Record Linkage Techniques -- 1997

Proceedings of an International Workshop and Exposition

March 20-21, 1997
Arlington, VA

Federal Committee on Statistical Methodology
Office of Management and Budget
Washington, DC 1997

Record Linkage Techniques -- 1997

Proceedings of an International Workshop and Exposition

Compiled and Edited by
Wendy Alvey and Bettye Jamerson

A Collaborative Effort by:

Bureau of the Census
Ernst and Young, LLP
Federal Committee on Statistical Methodology (OMB)
Internal Revenue Service
National Agricultural Statistics Service
National Cancer Institute
National Center for Health Statistics
National Research Council
National Science Foundation
Statistics Canada
Washington Statistical Society

Program Committee

FRITZ SCHEUREN (*Co-Chair*), Ernst and Young, LLP
WILLIAM WINKLER (*Co-Chair*), Bureau of the Census
MAXINE ANDERSON-BROWN (*Workshop Coordinator*), Bureau of the Census
WENDY ALVEY, Internal Revenue Service
JOHN ARMSTRONG, Statistics Canada
CHARLES DAY, National Agricultural Statistics Service
JOHN TUCKER, National Research Council
LAURA ZAYATZ, Bureau of the Census

Workshop Support

SHELBY FOLGER, Bureau of the Census
ANNA HOLAUS, Bureau of the Census
MICHAEL L. LUCAS, Bureau of the Census
CAROL MCDANIEL, Bureau of the Census
RUTH O'BRIEN, National Research Council
BARBARA WRIGHT, National Research Council

Acknowledgments

The Program Committee would like to thank the Bureau of the Census; Ernst and Young, LLP; the Office of Management and Budget's Federal Committee on Statistical Methdology; the Internal Revenue Service; the National Agricultural Statistics Service; the National Cancer Institute; the National Center for Health Statistics; the National Research Council'Theoretical Statistics; the National Science Foundation; Statistics Canada; and the Washington Statistical Society for supporting this Workshop project.

In addition, the Co-Chairs would like to thank John Tucker and Ruth O'Brien, of the National Research Council, for their support; Carol McDaniel, at Census, for overseeing the arrangements for the Workshop; Wendy Alvey, at the Internal Revenue Service, and Bettye Jamerson, of Jamie Productions, for compiling this compendium; and Nancy Kirkendall, at the Office of Management and Budget, and the Federal Committee on Statistical Methodology for printing this volume.

This publication will soon be available electronically through the Office of Management and Budget Web site. It can be reached through *http://www.fedstats.gov/* ; click on *Policy*; click on *Statistical Policy Working Papers*; and go down to *Other Multi-Agency Papers*.

Foreword

Introduction

On March 20-21 of this year, a two-day International Record Linkage Workshop and Exposition was held in Arlington, Virginia. Over 200 hundred people were in attendance; and, because the facilities were limited, we had to turn away yet another 200 interested individuals.

The Workshop had two main goals: first, we wanted an occasion to celebrate **Howard Newcombe's** pioneering practical work on computerized record linkage, which began in the 1950's -- see, e.g., Newcombe, Kennedy, Axford, and James (1959), Automatic Linkage of Vital Records, which appeared in *Science* -- and the theoretical underpinnings of his work, which were formalized in the 1960's by **Ivan Fellegi** and **Alan Sunter** in their classic 1969 paper on A Theory for Record Linkage, published in the *Journal of the American Statistical Association*. Second, we wanted a way to broadly update the methodological and technological developments in record linkage research and their applications since the Workshop on Exact Matching Methodologies, which was held in Washington, DC, in March 1985. The Proceedings from that earlier conference -- *Record Linkage Techniques -- 1985* -- have been widely cited; but much new work has been done since then.

Contents

The current volume is fortunate to present recent papers by two of the pioneers in record linkage research -- Fellegi and Newcombe -- as well as the work of many others who are exploring various aspects of exact matching techniques. Some of the new areas of related research involve increased privacy concerns due to record linkage; the growing interest in record linkage as a means for more efficient use of scarce statistical resources; the heightened importance of linkage technology, for such policy areas as health care reform; issues in physical security of data; and the measurement of nondisclosure and reidentification risks in public-use microdata files.

The format for this volume essentially follows that of the 1997 Workshop agenda -- with a section added to highlight key contributions made to the literature since the 1985 *Proceedings*. In those few cases where a paper was not available, the Conference abstract is provided. The report concludes with an Appendix, listing attendees who participated in the March 1997 Workshop and accompanying software expositions.

Copy Preparation and Reviews

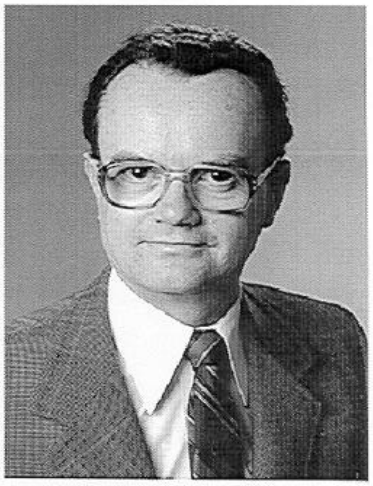
The contents of the papers included here are the responsibility of the authors – any opinions, findings, conclusions, or recommendations expressed in these material are those of the authors and do not necessarily reflect the views of the sponsors. With the exception of selected previously published papers, which were simply reproduced (with permission) as is, all of the papers in this volume underwent only a limited editorial review. Since this did not constitute a formal referee process, authors were also encouraged to obtain their own

technical review. Corrections and changes were either made by the authors, themselves, or cleared through them by the editors. Final layout of the papers was done by the editors, Wendy Alvey and Bettye Jamerson. Minor changes of a cosmetic nature were considered the prerogative of the editors.

Obtaining *Record Linkage Techniques -- 1985*

Tribute

Howard Newcombe



Ivan Fellegi

This workshop is dedicated to three innovative pioneers in the field of record linkage — to Howard Newcombe, for his seminal work in the practice of record linkage, begun over 40 years ago; and to Ivan Fellegi and Alan Sunter, for their joint theoretical work 30 years ago. Together they established, in large measure, what we still do today. The future will see us continuing to build on their work.

Table of Contents

Chapter 1

Record Linkage and Public Policy – A Dynamic Evolution

Ivan P. Fellegi, Statistics Canada

Abstract

Record linkage, as a major domain of substantive and technical interest, came about in the 1960s at the confluence of four closely inter-related developments:

- *First, the post-war evolution of the welfare state and taxation system resulted in the development of large files about individuals and business.*
- *Second, new computer technology facilitated the maintenance of these files, the practically unlimited integration of additional information, and the extraction of hitherto unimaginably complex information from them.*
- *Third, the very large expansion of the role of government resulted in an unprecedented increase in the demand for detailed information which, it was thought, could at least partially be satisfied by information derived from the administrative files which came about precisely because of this increase in the role of government.*
- *But there was a fourth factor present in many countries, one with perhaps the largest impact on subsequent developments: a high level of public concern that the other three developments represented a major threat to individual privacy and that this threat had to be dealt with.*

The paper traces the dynamic interaction of these factors over time, and their impact on the evolution of record linkage practice in three different domains of application: government statistics, public administration, and the private sector.

Introduction

It is a great honour and pleasure to be here today to think out loud about record linkage. I will say a few words about its evolution, its current status, and I will share with you some reflections about the future.

Let me start with the simplest possible definition of record linkage. There is a single record as well as a file of records and all records relate to some entities: persons, businesses, addresses, etc. Record linkage is the operation that, using the identifying information contained in the single record, seeks another record in the file referring to the same entity. If one accepts this definition, it is clear that people have been linking records ever since files existed: the filing clerk, for example, spent his or her entire working day looking for the "right" file to retrieve, or to insert new material in. The "right" file, of course, was the one that corresponded to the identification that was sought, where "identification" could be anything that uniquely described the "right" file.

Of course, this description of the traditional record linkage appears to be circuitous, but this did not matter to anyone: everyone knew what needed to be done, so the lack of definition had no operational consequence. The human mind could recognize the identification of a record in a file -- whether or not the descriptors contained some errors.

Four Critical Factors Shaping the Evolution of Record Linkage

Record linkage, as a major domain of substantive and technical interest, came about in the 1960s at the confluence of four closely inter-related developments:

- First, the post-war evolution of the welfare state and taxation system resulted in the development of large files about individuals and businesses (**opportunity**).

- Second, new computer technology facilitated the maintenance of these files, the practically unlimited integration of additional information, and the extraction of hitherto unimaginably complex information from them (**means**).

- Third, the very large expansion of the role of government resulted in an unprecedented increase in the demand for detailed information which, it was thought, could at least partially be satisfied by information derived from the administrative files which came about precisely as a consequence of the increase in the role of government (**need**).

- But there was a fourth factor present in many countries, one with perhaps the largest impact on subsequent developments: a high level of public concern that the other three developments represented a major threat to individual privacy and that this threat had to be dealt with (**constraint**).

As a result of the fourth factor, there was a very real commitment in these countries, whether formally taken or only implicitly accepted, that the creation of population registers must be avoided, indeed that even a uniform system of identifying persons would be unacceptable. In effect, files would be set up as and when needed, but personal information would not be integrated in a comprehensive manner. If all the relevant information had been kept together in a single large register, there would clearly have been little motivation to carry out the complex task of bringing together information from large and distinct files that were not designed for the purpose. Record linkage, it is important to remember, was therefore of particular interest in those countries which had a long history of striking a balance in favour of the individual in the tension between individual rights versus the needs of the state. In effect, record linkage came about to accomplish a task that was rendered difficult precisely because there was a social consensus that it **should** be difficult.

Much of the paper is devoted to an exploration of the dynamics among these four factors, and how they played themselves out in different domains of application: the statistical domain, other government applications, and the private sector. For simplicity and focus, I will mostly restrict my comments to files involving personal information.

A Historical Digression

While record linkage flourished because of government's administrative need, it is important to remember that the pioneering work of Howard Newcombe (Newcombe and Kennedy, 1962) was motivated by interest in genetic and biomedical research. Indeed, to this day a majority of record linkage applications carried out in Statistics Canada are health related.

However, my work with Alan Sunter (Fellegi and Sunter, 1969) was not motivated by health research issues. Rather, it was explicitly oriented to the problem of merging the information content of large administrative files in order to create a statistically useful source of new information. Our contribution can be summarized as follows:

- Newcombe recognized that linkage is a statistical problem: in the presence of errors of identifying information to decide which record pair of potential comparisons should be regarded as linked. Our first contribution involved formalizing this intuitive recognition and rigorously describing the space of record pairs consisting of all possible comparisons;

- Second, we provided a calculus for comparing the evidence contained in different record pairs about the likelihood that they refer to the same underlying unit;

- Third, we defined a linkage rule as a partitioning of a comparison space into the subset that we called "linked," i.e., record pairs about which the inference is that they indeed refer to the same underlying unit, a second subset for which the inference is that the record pairs refer to different underlying units, and a complementary third set where the inference cannot be made without further evidence;

- Fourth, as a formalization of the statistical character of the linkage rule, we identified the characteristic Type I and Type II errors associated with a given linkage rule: the proportion of record pairs that are falsely linked and the proportion that are incorrectly unlinked;

- Fifth, we showed that if the space of record pair comparisons is ordered according to our metric, this will result in a linkage rule that is optimal for any pre-specified Type I and Type II error levels;

- Our final contribution was to provide a framework that, in retrospect, turned out to be fruitful both for the design of operationally efficient record linkage systems and for the identification of useful areas for further research. Perhaps this was our most important contribution: facilitating the outstanding research that followed.

The very fact that there is a successful symposium here today is testimony to the productivity of that research. There has certainly been a spectacular evolution of methodology and techniques, signalling a continuing, perhaps even increasing interest in the topic. However, the basic tension among the four critical factors mentioned above was never fully resolved, even though the dynamics were quite different in the different domains of application. Let me turn to a brief overview of these application domains.

The Statistical Domain -- A Model?

Since most of us here are statisticians, I will start with statistical applications. The defining characteristic of this domain is that the output does not relate to identifiable individuals -- i.e., that statistical confidentiality is preserved. This very important distinction ought to result in a different public attitude to linkage by statisticians for statistical purposes. But I am not sure that it does -- for two related reasons. First, the process of linkage of personal records is intrinsically privacy intrusive, in the sense that information is brought together about a person without his or her knowledge and control. From that point of view it is largely irrelevant that the results can only affect particular individuals in an indirect manner. The second reason is that not everyone trusts us completely to maintain statistical confidentiality.

Statistical confidentiality protection in Canada, at least within government, is certainly tight -- both legally and de facto. As you know, unlike the United States, almost all government statistical activity is carried out within a single agency and is covered by a uniform and strong statistics act. In spite of that, we have taken what we think is a very cautious, though hopefully balanced attitude to record linkage. We have developed explicit

policies and strong mechanisms to make them effective.

Statistics Canada will undertake record linkage activities only if all the following conditions are satisfied:

- the purpose of the linkage activity is statistical/research;
- the products of the activity will be released only in accordance with the confidentiality provisions of the Statistics Act;
- the benefits to be derived from the linkage are substantial and clearly serve the public interest;
- record linkage is either the only option to acquire the needed information or, given the cost or burden implications of alternative approaches, it is the only feasible option;
- the record linkage activity will not be used for purposes that can be detrimental to the individuals involved;
- the record linkage activity is judged not to jeopardize the future conduct of Statistics Canada's programs; and finally
- the linkage satisfies a prescribed review and approval process.

Let me underline some features of this policy. Beyond the more or less obvious fact that we will not carry out linkage except for statistical purposes, and that we will protect confidentiality, the main feature of the policy is to seek a balance. We recognize that linkage is intrinsically intrusive of privacy, so we will only consider undertaking it where the public benefit is sufficiently important to tip the balance of decision. But even when this is the case, we want evidence that alternative methods to acquire equivalent information are either impossible or prohibitively costly. Another requirement is that the objective of the project should not be detrimental to the individuals concerned. This last point bears emphasis: we are not talking about individual jeopardy -- that is ruled out by our strict confidentiality protection -- but of possible harm to the group of people whose records are involved. Since typically we cannot contact them to obtain their informed consent, we want to make sure that we will not link their information if, as a group, they would not have given us informed consent had we been able to seek it.

However noble, no set of principles is likely to have operational impacts without a set of procedures designed to give them effect. In our case this involves a cascading set of approvals. Every manager who wishes to sponsor a linkage application has to submit a narrative describing the purpose, the expected public benefits, whether there is a possibility of harm to the individuals concerned, and whether there are feasible alternative approaches. In addition, the manager also has to describe the proposed methodology, any features that might enhance privacy or confidentiality protection, and has to propose a tight schedule for the destruction of linked identifiers. This information is assessed, in the first place, by a standing committee composed of several of Statistics Canada directors that is chaired by one of my direct assistants. Their assessment and recommendation is reviewed by the agency’s top level management group that I chair. If we decide that the public good indeed outweighs the privacy concern and that the objectives cannot reasonably be achieved without linkage, we next consider whether the project needs ministerial approval and/or external “stakeholder” review. Generally, ministerial approval is sought unless a previous approval has clearly established a precedent.

Least problematic are cases where both files to be linked were collected for a statistical purpose. Examples are routine linkages of successive rounds of panel surveys for purposes of editing, linkage of successive waves of longitudinal surveys, or the linkage of the census and a post-censal survey to assess the completeness of the census count.

Health applications provide another class of well established precedents. These typically involve a file, provided by an external organization, containing records of persons known to have been exposed to a health risk. These exposure files are linked either to a machine readable cancer register or to our mortality file. The purpose is to assess whether the exposed persons had a higher rate of some specified cancer, or whether they had a disproportionate number of deaths due to a suspected cause. If the proposal involves more than a scientific fishing expedition, i.e., if it is designed to explore a reasonably well-founded scientific hypothesis, then the linkage is normally approved by the senior management of Statistics Canada, since the precedent for ministerial approval is well established in these cases.

In other cases where there is no applicable precedent, the public benefit is considered carefully by Statistics Canada. If in our judgement the benefit is not sufficient to proceed, the request is rejected and that is the end of the matter. However, if we feel that there is considerable public benefit to be derived from the linkage and there is no alternative approach that is practical, then we make a positive recommendation to our minister.

Statistics Canada is a very autonomous organization, operating at arm's length from the political process. This is just about the only programmatic issue on which we seek ministerial guidance. Why do we do so here? Because privacy and information are both public goods and there is no methodology for a **professional** assessment of the right balance between them. Establishing the balance between competing public goods, however, is very much a function of elected politicians in democratic societies.

Our review process might involve an extra step in those rare cases where the potential public good is judged to be very high, but where the privacy issue is particularly sensitive. An example will illustrate. In Canada a substantial proportion of the social assistance to the poor is administered by provinces and there is no federal record of such disbursements. Conversely, the provinces have no access to the records of federal social assistance programs, e.g., unemployment insurance. A few years ago we received a very serious research proposal to study the combined effect on incomes of federal and provincial social assistance, as well as taxes. The objective was to assess the combined impact of all these programs: e.g. are they properly focused on the poor, or are there unintended disincentives to work. This was clearly a program of major potential public benefit, indeed of major potential benefit to the poor. But, equally clearly, it would have been prohibitively expensive to secure their informed consent. As the next best thing, we convened a seminar with the Privacy Commissioner, with interested researchers, and with representatives of advocacy groups for the poor, calling on the latter as proxy spokespersons. The linkage was endorsed by them and it subsequently received ministerial approval.

Our approach to privacy protection merited the following salutation from the Privacy Commissioner in his annual report to Parliament in 1990:

> "Worthy of special mention in an end-of-term report is the tangible commitment to privacy demonstrated by the Chief Statistician of Canada. It is especially noteworthy because many of the Privacy Act requirements do not apply to statistical records.
>
> The Chief Statistician took the initiative, as he has on other privacy matters, and sought the Privacy Commissioner's view on whether the public interest justified conducting this record linkage.
>
> The Privacy Commissioner agreed that the proposed pilot project had potential for contributing significantly to the public interest; most important, he considered it impossible to accomplish the goal without intruding on personal privacy...

> Some may consider the Privacy Act remiss in not subjecting personal information used for statistical purposes to the same requirements imposed on personal information used for administrative purposes. No one should doubt, however, that the privacy concerns about statistical data are being addressed in practice.
>
> The Chief Statistician of Canada is to be thanked for that."

I have quoted from this report at some length because it makes a number of important points. First, even a professional privacy advocate like the Commissioner agrees explicitly that the need to protect privacy must be weighed against the need for information. Second, because it makes clear that while statistical records are not legally subject to most requirements of the Privacy Act, it is strategically important for statisticians to be very prudent about record linkage. And last but not least, the quote (and our continuing experience since 1990 as well) proves that a well balanced policy, **together with concrete administrative practice to give it effect**, can effectively mediate between the two competing public goods of privacy and need to know.

The Dog That Does Not Bark: Public Administration and Public Unconcern

Large government data banks about persons have traditionally been regarded as threatening because of the visible power of the state: to make compulsory the provision of information that it needs, to make decisions on the basis of information in its possession, and finally to enforce the decisions made on the basis of the information held. One might expect, therefore, that the balance of forces affecting record linkage would follow Newton's third law of dynamics: the stronger the combined pro-linkage forces of opportunity, need and means, the stronger would become the constraining counterforce of concerns about privacy. Yet, during much of the last thirty years there has been a curious disjunction between, on the one hand, the level of public anxiety about record linkage and data banks, and on the other hand the level of actual government activity in these fields.

During the 1960s, 70s and much of the 80s, two main factors restrained the spread of government record linkage applications. First, the potential pay-off from this kind of linkage -- reduction of financial errors and outright fraud -- was not high on the agenda of governments. Consequently the widespread, though mostly latent, public hostility toward linkage effectively restrained governments, particularly in the absence of an overriding financial objective. And second, the level of technology available at the time kept the cost of linkage reasonably high. In addition, some measures of transparency introduced in most developed countries have also been helpful in alleviating concerns. In Canada, for example, every citizen has access to a register which describes the content and purposes of all government held personal data banks. Should they wish, they may obtain, free of charge, a copy of the information held about them. They also have a legal right to have their non-statistical records corrected or updated if they are in error.

The combination of these factors has effectively blunted the public's sense of concern about government data banks and record linkage. Yet during the last several years a significant change occurred in the balance of the forces at work. Deficit reduction rose to the top of governments' agenda, increasing the importance attached to controlling tax evasion and welfare fraud. At the same time, cuts in public services in the name of deficit reduction have caused substantial and widely reported hardships -- compared to which privacy fears seemed to have assumed a diminished importance. And, of course, the cost of linkage shrank rapidly. So, precisely at a time when record linkage applications by government are growing rapidly, there is hardly any public debate, let alone open concern, about the practice -- except for the one-day news triggered annually by the release of yet another report by the Privacy Commissioner.

So long as the public is properly informed but chooses to be unconcerned, we should all be pleased about the equilibrium that may have been reached. But I am concerned. First of all because the apparent equilibrium is not based on informed debates. Even more important, I believe the status quo to be fragile: a single egregious

error or accident might suffice to put a spotlight on the extent of linkage going on in government. In that case the incident might well balloon out of control if elementary questions cannot be answered about the weight given to privacy issues in approving each application, about procedural checks and balances, about accountability, and about the point in the bureaucratic and political hierarchy at which final approval was given.

I believe there is a great need for much increased transparency here. **We need to develop explicit and publicly debatable policies about both the criteria and processes involved in approving record linkage for administrative purposes** -- perhaps along the lines of the process used by Statistics Canada, but of course suitably modified to fit the different domains of application. This need not go as far as it has in some European countries where approval depends on an appointed privacy commissioner. Privacy Commissioners have an important role as advocates, i.e., as the public guardians of one side of the issue. But approval should entail a proper consideration of the conflict between the two competing public goods: privacy and enforcement. As such the approval should ultimately involve the political level. The process of political consideration can and must be supported by a bureaucratic process which reviews options, assesses benefits, and recommends approaches that reduce the risks to privacy.

Out of Control -- Linkage in the Private Sector

My concerns about the public sector are dwarfed by the discomfort I have about linkage in the private sector. Not that I know much about it -- and I suspect the same applies to most of you. But this is precisely the sign of a potentially very serious problem: the unrecognized and undiscussed threat of privately held data banks and large scale record linkage.

On the surface, and in comparison with the public sector, the private sector appears to be innocuous for two broad and interconnected reasons. First, one may think that its possible decisions about us affect us less profoundly, and hence the issue of control over personal information is less acute. And second, that the private sector has no legally enforceable sanctions to back up its data collection and its decisions about people. But none of the arguments about lack of sanctions or lack of impact stand up to scrutiny. On the one hand, the unregulated private sector rarely has the need to back up its information based decisions with sanctions against individuals -- it simply stops dealing with them. On the other hand, the ultimate threat of denying a benefit is probably sufficient to make the collection of information compulsory to all intents and purposes. After all, try to obtain a credit card, register a warranty, or seek insurance without providing the requested information. Of course, it can be argued that having a credit card or health insurance are not necessities -- but would you like to try living without them?

While some segments in the private sector clearly have de facto compulsory powers of data collection, others completely bypass the issue of informed consent by buying personal information. The impact of decisions on our lives made on the basis of indirectly obtained information can range from the inconvenient to the profound. Let me give you some examples.

You may or may not receive certain kinds of advertising depending on the information held about you by the distributor. You may miss some information that you might like to have or, conversely, you might be annoyed by the unnecessary “junk mail.”

The information held about you might affect your credit rating without you even being aware, let alone having the power to insist on the correction of erroneous information.

You may not receive insurance you might like to receive.

Your adversaries in a court case might gain undue advantage in preparing their case by accessing information held about you.

The list of examples could go on. I hope to have convinced you that there is, indeed, a serious privacy risk. How come that, as a society, we have allowed this situation to evolve? In effect, the four factors affecting record linkage have evolved differently in this domain compared to others.

- **Opportunity.--**The availability of cheap and powerful hardware and software facilitated the widespread use of information and communication technology. Therefore a capacity was created and made widely accessible for building up large files about clients as well as for making use of similar files created by others;

- **Means.--**Fragmented and dispersed lists of persons, even if widely available, used to be regarded as representing a negligible risk. But the cheap availability of computing, together with the powerful methodology-based software has altered the picture. Indeed, we in Statistics Canada were able to construct, as part of our 1991 census preparation, an excellent address list using client lists from the tax authority, telephone companies, hydro companies and municipal assessments. Even more worrisome is that the 3-4 largest credit card companies, among themselves, hold a list of consumers whose coverage of the adult population is probably close to universal. Furthermore, in addition to their excellent coverage, credit card company lists also hold a vast amount of information about us, including our income, expenditure patterns (and therefore our individual preferences and dislikes), our travel pattern, home repairs, etc.

- **Need.--**Advertising has evolved and it no longer relies solely on the mass media. Increasingly, companies prefer pinpoint approaches, using a variety of mailing or telephone lists. These lists are customized with great sophistication, using the information contained in them, to delineate the population group to be targeted. The advertising utility and value of the input files is directly determined by the amount of relevant personal information contained in them.

Thus three of our four critical factors have interacted positively, resulting in the creation of a mass market for large electronic files about consumers. In turn, this mass market created a new industry of service providers. These are information brokers specializing in consumer files, their updating, the upgrading of information held about people (i.e., record linkage), and the marketing of both the files themselves and services based on them. Although I have no objective evidence, there is little doubt that commerce in personal information has become a big business. And it is entirely unregulated. Which leads to the fourth critical factor:

- **Constraint.--**As indicated above, there is practically none.

From the perspective of privacy there are two quite distinct problems with the current situation. The first is that we have, indeed, lost control over the use made by others of information held about ourselves. The second basic problem is that we can't even control the accuracy of this information.

So there is a problem. Is there a solution?

The knee-jerk reaction might be to address these problems through regulation. But it only takes a few seconds of reflection to realize that this traditional tool, by itself, would not be workable. Electronic communication has become so cheap that the physical location of files can be moved anywhere in the world without the least impact on access and use. So if they don't like one country's regulations, the information brokers can simply take their files to another.

Are we completely defenceless? I don't think so. If regulations cannot be enforced by government, perhaps they can be designed to lead to a degree of self-enforcement within the private sector, based on their own enlightened self-interest. I will conclude this talk by outlining a possible approach.

Ideally, one may wish to achieve two objectives: to improve people's control over the use of information

about themselves; and to improve their ability to control the accuracy of such information.

I am not particularly sanguine about restoring individuals' control of the use of information about themselves. But we might nevertheless be able to improve the current situation. The approach could be based on the observation that while the operations of companies might be moved from one country to another, the transaction whereby members of a population provide information about themselves is intrinsically a domestic one, hence potentially subject to regulation. We could prescribe, therefore, that when information is requested from people, certain information must also be provided to them. This could include a description of the information management practices of the requesting company, the control methods they use, and whether and under what conditions they provide access to personal information to other companies. Can the government effectively monitor the adherence by companies to such standards? Certainly not directly. But at this point competitive pressures might come to the fore. Those who adhere to their promised information management standards incur some costs. It is in their interest that their competitors' misdemeanours should not be a source of unfair competitive advantage to the latter. The availability of formal complaint mechanisms, maintained either by the government or by the industry itself, might provide a constructive outlet for the policing of each firm's practice by its own competitors.

This approach would not be a guarantee against undesirable secondary information use. But at least it might go some distance towards a form of informed consent. If there is a significant number of consumers for whom control of their information is a priority, competitive pressure might help encourage some firms to cater to such consumer preferences, even at some small additional cost. If, as a result, these firms end up with better information, they will gain market share, eventually, perhaps squeezing out their less accommodating competitors. But in order to encourage this form of competition, government must provide a productive framework through its initial regulation and the creation of a suitable complaint mechanism.

I am a little more optimistic regarding the possibility of improving the reliability of the information content of personal information banks held in the private sector. First, accuracy of information is surely in the interest of an overwhelming proportion of users of personal information. So government could establish a licensing system for recognized carriers of personal information. The data banks held by such carriers (whether or not they are physically inside or outside the country) would be listed in a register of personal data banks, such as exists in Canada in respect of government operated personal data banks. A requirement for the license would be the obligation to provide free access by people to the information held about themselves, as well as the obligation to implement all corrections on demand. Since compliance with these regulations would improve the accuracy of information held in such data banks, registered carriers would have a competitive advantage. However, adherence to the regulations would also drive up their costs. If the competitive advantage due to higher accuracy is not sufficient to offset their higher costs, it might be necessary to consider additional measures. One possibility would be penalties against the usage of personal information held in unregulated data banks (the penalties would, of course, have to be assessed against the information users rather than the providers since the latter might be outside the national territory). The price differential between the regulated and unregulated operators would provide an incentive for the former to "police" the latter.

The combination of approaches proposed here would not restore to individuals full control over information about themselves -- the ultimate objective of privacy advocates. But I am firmly convinced that this objective is no longer attainable. They would, however, restore a significant element of informed consent to the process of providing personal information. It would also go some way to improving the accuracy of privately held personal information

banks. As such, these measures would be a major improvement over the current absolute free-for-all -- a situation which, I believe, is intolerable. If we do not at least acknowledge that we have a serious and rapidly worsening problem, and if we do not take practical measures to deal with it, then in effect we connive in its continuation and exacerbation.

Conclusion

We seem to have come full circle. As a society we did not want comprehensive population registers, largely because we did not want a large scale and routine merging of information contained in government files. But we did not want to rule out *some* merging for *some* well justified purposes. So, as a matter of conscious public policy, we made linkage very difficult. However, we allowed the development of record linkage methodology for use in exceptional circumstances. The applications were indeed important, often requiring a high level of accuracy, so we refined the methodology, and also made it vastly more efficient. Combined with rapidly diminishing computing costs, this efficient methodology rendered linkage into a truly ubiquitous tool: indeed, at this symposium there are several versions of the methodology on display, competing on efficiency and ease of use. The activity that was designed to be difficult has become quite easy and inexpensive.

As a society we have been concerned with the power of the state and the risk of that power being abused. So we constrained the ability of the state to use our personal information without our consent. It is perhaps a coincidence, but certainly not a contradiction, that as the relative power of the state is declining, we are, *de facto* and without much public discussion, allowing it more extensive latitude to link our personal information without our explicit agreement. It is, however, a paradox that as the relative balance of power is shifting to the private sector, we are allowing it to build up extensive personal data banks, without regulation or even assurances about the accuracy of its contents. The power that we used to be anxious to deny to the state, which is operating under the guidance of our elected representatives, we are allowing the private sector to acquire -- indeed we seem to be doing so with a shrug of the shoulders.

In a democratic society it is of paramount importance that major public issues be decided based on well informed public debate. This paper is intended as a modest contribution to ensuring that our **consent** is, indeed, based on **understanding**.

References

Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, 64, 1183-1210.

Newcombe, H. B. and Kennedy, J. M. (1962). Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information, *Communications of the A.C.M.*, 5, 563.

Chapter 2

OX-LINK: The Oxford Medical Record Linkage System

Leicester E. Gill, University of Oxford

Abstract

This paper describes the major features of the Oxford record linkage system (OX-LINK), with its use of the Oxford name compression algorithm (ONCA), the calculation of the names weights, the use of orthogonal matrices to determine the threshold acceptance weights, and the use of combinational and heuristic algebraic algorithms to select the potential links between pairs of records.

The system was developed using the collection of linkable abstracts that comprise the Oxford Record Linkage Study (ORLS), which covers 10 million records for 5 million people and spans 1963 to date. The linked dataset is used for the preparation of health services statistics, and for epidemiological and health services research. The policy of the Oxford unit is to comprehensively link all the records rather than prepare links on an ad-hoc basis.

The OX-LINK system has been further developed and refined for internally cross matching the whole of the National Health Service Central Register (NHSCR) against itself (57.9 million records), and to detect and remove duplicate pairs; as a first step towards the issue of a new NHS number to everyone in England and Wales. A recent development is the matching of general practice (primary care) records with hospital and vital records to prepare a file for analyzing referral, prescribing and outcome measures.

Other uses of the system include ad hoc linkages for specific cohorts, academic support for the development of test programs and data for efficiently and accurately tracing people within the NHSCR, and developing methodologies for preparing registers containing a high proportion of ethnic names.

Medical Record Linkage

The term record linkage, first used by H. L. Dunn (1946; Gill and Baldwin, 1987), expresses the concept of collating health-care records into a cumulative personal file, starting with birth and ending with death. Dunn also emphasised the use of linked files to establish the accuracy or otherwise of the recorded data. Newcombe (Newcombe et al., 1959; and Newcombe, 1967, 1987, and 1988) undertook the pioneering work on medical record linkage in Canada in the 1950's and thereafter, Acheson (1967, 1968) established the first record linkage system in England in 1962.

When the requirement is to link records at different times and in different places, in principle it would be possible to link such records using a unique personal identification number. In practice, a unique number has not generally been available on records in the UK of interest in medicine and therefore other methods such as the use of surnames, forenames and dates of birth, have been necessary to identify different records relating to

the same individual. In this paper, I will confine my discussion to the linkage of records for different events which relate to the same person.

Matching and Linking

The fundamental requirement for correct matching is that there should be a means of uniquely identifying the person on every document to be linked. Matching may be *all-or-none*, or it may be *probabilistic*, i.e., based on a computed calculation of the probability that the records relate to the same person, as described below. In probability matching, a threshold of likelihood is set (which can be varied in different circumstances) above which a pair of records is accepted as a match, relating to the same person, and below which the match is rejected.

The main requirement for all-or-none matching is a unique identifier for the person which is fixed, easily recorded, verifiable, and available on every relevant record. Few, if any, identifiers meet all these specifications. However, systems of numbers or other ciphers can be generated which meet these criteria within an individual health care setting (e.g., within a hospital or district) or, in principle, more widely (e.g., the National Health Service number). In the past, the National Health Service number in England and Wales had serious limitations as a matching variable, and it was not widely used on health-care records. With the allocation of the new ten digit number throughout the NHS all this is being changed (Secretaries of State, 1989; National Health Service and Department of Health, 1990), and it will be incorporated in all health-care records from 1997.

Numbering systems, though simple in concept, are prone to errors of recording, transcription and keying. It is therefore essential to consider methods for reducing errors in their use. One such method is to incorporate a checking device such as the use of *check-digits* (Wild, 1968; Hamming, 1986; Gallian, 1989; Baldwin and Gill, 1982; and Holmes, 1975). In circumstances where unique numbers or ciphers are not universally used, obvious candidates for use as matching variables are the person's names, date of birth, sex and perhaps other supplementary variables such as the address or postcode and place of birth. These, considered individually, are partial identifiers and matching depends on their use in combination.

Unique Personal Identifiers

Personal identification, administrative and clinical data are gradually accumulated during a patient's spell in a hospital and finalized into a single record. This type of linkage is conducted as normal practice in hospital information systems, especially in those hospitals having Patient Administration Systems (PAS) and District Information Systems (DIS) which use a centrally allocated check-digited District Number as the unique identifier (Goldacre, 1986).

Identifying numbers are often made up, in part, from stable features of a person's identification set, for example, sex, date of birth and place of birth, and so can be reconstructed in full or part, even if the number is lost or forgotten. In the United Kingdom (UK), the new 10-digit NHS number is an arbitrarily allocated integer, almost impossible to commit to memory, and cannot be reconstructed from the person's personal identifiers.

Difficulties arise, however, where the health event record does not include a unique identifier. In such cases, matching and linking depends on achieving the closest approach to unique identification by using several identifying variables each of which is only a partial identifier but which, in combination, provide a match which is sufficiently accurate for the intended uses of the linked data.

Personal Identifying Variables

The personal identifying variables that are normally used for person matching can be considered in five quite separate groups.

- **Group 1**.--Represents the persons proper names and with the exception of present surname when women adopt their husbands surname on marriage, rarely changes during a person's lifetime: birth surname; present surname; first forename or first initial; second forename or second initial; and, other forenames.

- **Group 2**.--Consists of the non-name personal characteristics that are fixed at birth and very rarely changes during the person's lifetime: gender (Sex at birth); date of birth; place of birth (address where parents living when person was born); NHS number (allocated at birth registration, both old and new formats); date of death; and ethnicity.

- **Group 3**.--Consists of socio-demographic variables that can change many times during the course of the person's lifetime: street address; post code; general practitioner; marital status; social class; number(s) allocated by a health district or special health-care register; number(s) allocated by a hospital or trust; number(s) allocated by a general practitioner's computing system; and, any other special hospital allocated numbers.

- **Group 4**.--Consists of other variables that could be used for the compilation of special registers: clinical specialty; diagnosis; cancer site; drug idiosyncrasy or therapy; occupation; date of death; and other dates (for example, LMP, etc.).

- **Group 5**.--Consists of variables that could be used for family record linkage: other surnames; mother's birth surname; father's surname; marital status; number of births; birth order; birth weight; date of marriage; and number of marriages.

File Ordering and Blocking

Matching and linkage in established datasets usually involves comparing each new record with a master file containing existing records. Files are ordered or *blocked* in particular ways to increase the efficiency of searching. In similar fashion to looking up a name in a telephone directory the matching algorithm must be able to generate the "*see also*" equivalent to this surname for variations in spelling (e.g., Stuart and Stewart, Mc, Mk, and Mac). Searching can be continued, if necessary, under the alternative surname.

Algorithmics that emulate the "see also" method are used for computer matching in record linkage. In this way, for example, Stuarts and Stewarts are collated into the same block. A match is determined by the amount of agreement and disagreement between the identifiers on the "incoming" record and those on the master file. The computer calculates the statistical probability that the person on the master file is the same as the person on the record with which it is compared.

File Blocking

The reliability and efficiency of matching is very dependent on the way in which the initial grouping or the "file-blocking" step is carried out. It is important to generate blocks of the right size. The balance between the number and size of blocks is particularly important when large files are being matched. The selection of variables to be used for file blocking is, therefore, critical and will be discussed before considering the comparison and decision-making stages of probability matching.

Any variable that is present on each and every record on the dataset to be matched could be used to divide or block the file, so enhancing the search and reducing the number of unproductive comparisons. Nevertheless,

if there is a risk that the items chosen are wrongly recorded -- which would result in the records being assigned to the wrong file block, then potential matches will be missed. Items that are likely to change their value from one record to another for the same person, such as home address, are not suitable for file blocking. The items used for file blocking must be universally available, reliably recorded and permanent. In practice, it is almost always necessary to use surnames, combined with one or two other ubiquitous items, such as sex and year of birth, to subdivide the file into blocks that are manageable in size and stable. Considerable attention has been given to the ways in which surnames are captured and algorithmic methods to reduce, or eliminate, the effects of variations in spelling and reporting, and which "compress" names into fixed-length codes.

Phonemic Name Compression

In record linkage, name compression codes are used for grouping together variants of surnames for the purposes of blocking and searching, so that effective match comparisons can be made using both the full name and other identifying data, despite misspelled or misreported names.

The first major advance in name compression was achieved by applying the principles of phonetics to group together classes of similar-sounding groups of letters, and thus similar-sounding names. The best known of these codes was devised in the 1920's by Odell and Russell (Knuth, 1973) and is known as the Soundex code. Other name compression algorithms are described by Dolby (1970) and elsewhere.

Soundex Code and the Oxford Name Compression Algorithm (ONCA)

The Soundex code has been widely used in medical record systems despite its disadvantages. Although the algorithm copes well with Anglo-Saxon and European names, it fails to bring together some common variants of names, such as Thomson/Thompson, Horton/Hawton, Goff/Gough, etc., and it does not perform well where the names are short, as is the case for the very common names, have a high percentage of vowels, or are of Oriental origin.

It is used principally, for the transformation of groups of consonants within names, to specific combinations of both vowels and consonants (Dolby, 1970). Among several algorithms of this type, that devised by the New York State Information and Intelligence System (NYSIIS) has been particularly successful, and has been used in a modified form by Statistics Canada and in the USA for an extensive series of record linkage studies (Lynch and Arends, 1977). A recent development in the Unit of Health-Care Epidemiology (UHCE) (Gill and Baldwin, 1987; Gill et al., 1993), referred to as the Oxford Name Compression Algorithm (ONCA), uses an anglicised version of the NYSIIS method of compression as the initial or pre-processing stage, and the transformed and partially compressed name is then Soundexed in the usual way. This two-stage technique has been used successfully for blocking the files of the ORLS, and overcomes most of the unsatisfactory features of pure Soundexing while retaining a convenient four-character fixed-length format.

The blocks produced using ONCA alone vary in size, from quite small and manageable for the less common surnames, to very large and uneconomic for the more common surnames. Further subdivision of the ONCA blocks on the file can be effected using sex, forename initial and date of birth either singly or in combination.

ORLS File Blocking Keys and Matching Variables

The file blocking keys used for the ORLS are generated in the following fashion:

- The primary key is generated using the ONCA of the present surname.

- The secondary key is generated from the initial letter of the first forename. Where this forename is a nickname or a known contraction of the "formal" forename, then the initial of the "formal" forename is used. For example, if the recorded forename was BILL, the "formal" forename would be William, and the initial used would be W. A further record is set up on the master file where a second forename or initial is present; the key is derived from this second initial.

- Where the birth surname is not the same as the present surname, as in the case of married women, a further record is set up on the master file under the ONCA code of birth surname and again subdivided by the initial. (This process is termed *exploding* the file.)

- Further keys based on the date of birth and other blocking variables are also generated.

In addition to the sorting header, four other variables are added to each record before sorting and matching is undertaken:

- **Accession Number**.--A unique number allocated from a pool of such numbers, and is absolutely unique to this record. The number is never changed and is used for identification of this record for correction and amendment. The number is check digited to modulus 97.

- **Person or System Number**.--A unique number allocated from a pool of such numbers. The number can be changed or replaced if this record matches with another record. The number is check digited to modulus 97.

- **Coding Editions**.--Indicators that record the various editions of the coding frames used in this record, for example the version of the ICD (International Classification of Diseases) or the surgical procedure codes. These indicators ensure that the correct coding edition is always recorded on each and every record and reliance is not placed on a vague range of dates.

- **Input and Output Stream Number**.--This variable is used for identifying a particular dataset during a matching run, and enables a number of matches to be undertaken independently at the same pass down the master file.

Generating Extra Records Where a Number of Name Variants Are Present

To ensure that the data record can match with the blocks containing all possible variants of the names information, multiple records are generated on the master file containing combinations of present and birth surnames, and forenames. To illustrate the generation of extra records where the identifying set for a person contains many variants of the names, consider the following example:

birth surname:	SMITH
present surname (married surname):	HALL
first forename:	LIZ (contraction of Elizabeth)
second forename:	PEGGY (contraction of Margaret)
year of birth:	1952 (old enough to be married).

Eight records would be generated on the master file and each record indexed under the various combinations of ONCA and initial, as follows:

Indexed under the present surname HALL: i.e., ONCA H400:

H400L	for Liz
H400E	for Elizabeth (formal version of Liz)

H400P	for Peggy
H400M	for Margaret (formal version of Peggy);
Indexed under the birth surname SMITH: i.e., ONCA S530:	
S530L	for Liz
S530E	for Elizabeth (formal version of Liz)
S530P	for Peggy
S530M	for Margaret (formal version of Peggy).

Mrs. Hall would have her master file record included under each of the above eight ONCA/initial blocks. A data record containing any combination of the above names would generate an ONCA/initial code similar to any one of the eight above, and would have a high expectation of matching to any of the variants during the matching phase.

To reduce the number of unproductive comparisons, a data record will only be matched with *an other record in the same block* provided that the year of birth on both records are within 16 years of each other. This constraint has been applied, firstly, to reduce the number of unproductive matches, and secondly to confine matching to persons born within the same generation, and in this way eliminate father/son, mother/daughter matches. Further constraints could be built into the matching software for example, matching only within the same sex, logically checking that the dates on the two records are in a particular sequence or range, or that the diagnoses on the two records are in a specified range, as required in the preparation of a cancer registry file.

Matching Methods

There are two methods of matching data records with a master file.

- The **two file method** is used to match a data record from a data file with a block on the master file, and in this way compare the data record with every record in the master file block.

- The **one file/single pass** method is used to combine the data file block and the master file block into one block, and to match each record with every other in the block in a triangular fashion, i.e., first with the rest, followed by second with the rest etc. In this way every record can be matched with every other record. Use of a stream number on each record enables selective matching to be undertaken, for example data records can be matched with the master file and with each other, but the master file records are not matched with themselves.

Match Weights

Considerable work has been undertaken to develop methods of calculating the probability that pairs of records, containing arrays of partial identifiers which may be subject to error or variation in recording do, or do not, relate to the same person. Decisions can then be made about the level of probability to accept. The issues are those of reducing false negatives (Type I errors) and false positives (Type II errors) in matching (Winkler, 1995; Scheuren and Winkler, 1996; and Belin and Rubin, 1995). A false negative error, or "missed match," occurs when records which relate to the same person are not drawn together (perhaps because of minor variations in spelling or a minor error in recorded dates of birth). Matches may also be missed if the two records fall into different blocks. This may happen if, for example, a surname is misspelled and the phonemic compression algorithm puts them into two different blocks.

Methods for probability matching depend on making comparisons between each of several items of identifying information. Computer-based calculations are then made which are based on the *discriminating power* of each item. For example, a comparison between two different records containing the same surname has greater discriminating power if the surnames are rare than if they are common. Higher scores are given for

agreement between identifiers (such as particular surnames) which are uncommon than for those which are common. The extent to which an identifier is uncommon or common can be determined empirically from its distribution in the population studied. Numerical values can then be calculated routinely in the process of matching for the amount of agreement or disagreement between the various identifying items on the records. In this way a composite score or *match weight* can be calculated for each pair of records, indicating the probability that they relate to the same person. In essence, these weights simulate the subjective judgement of a clerk. A detailed discussion of match weights and probability matching can be found in publications by Newcombe (Newcombe et al., 1959; and Newcombe, 1967, 1987, and 1988), and by Gill and Baldwin (1987) (See also Acheson, 1968.)

Calculating the Weights for the Names Items

Name identifiers are weighted in a different fashion to the non-name identifiers, because there are many more variations for correctly spelled names. Analysis of the NHS central register for England and Wales shows that there are:

57,963,992	records
1,071,603	surnames
15,143,043	surname/forename pairs.

The low frequency names were mainly non Anglo-Saxon names, hyphenated names and misspelled names. In general the misspellings were due to embedded vowel changes or to miss keying. A more detailed examination of the register showed that 954 different surnames covered about 50% of the population, with the following frequency distribution:

10% population	24 different surnames
20% population	84 different surnames
30% population	213 different surnames
40% population	460 different surnames
50% population	954 different surnames
60% population	1,908 different surnames
70% population	3,912 different surnames
80% population	10,214 different surnames
90% population	100,000 different surnames
100% population	1,071,603 different surnames.

Many spelling variations were detected for the common forenames. Using data from the NHSCR register, various forename directories and other sources of forenames, a formal forename lexicon was prepared that contained the well known contractions and nicknames. The problem in preparing the lexicon was whether to include forenames that had minor spelling errors, for example JOHN and JON. This lexicon is being used in the matching algorithm, to convert nicknames and contractions, for example LIZ, to the formal forename ELIZABETH, and both names are used as part of the search strategy.

Calculation of Weights for Surnames

The binit weight calculated from the frequency of the first letter in the surname (26 different values) was found to be too crude for matching files containing over 1 million records. The weights for Smith, Snaith, Sneath, Smoothey, Samuda, and Szabo would all have been set to some low value calculated from the frequency of Smith in the population, and ignoring the frequency of the much rarer Szabo. Using the frequencies of all of the 1 million or more different surnames on the master match file is too cumbersome, time consuming to keep up-to-date, and operationally difficult to store during the match run. The list would also have contained all of the one-off surnames generated by bad transcription and bad spelling. A compromise solution was de-

vised by calculating the weights based on the frequency of the ONCA block (8,000 values), with a cut-off value of 1 in a 1,000 in order to prevent the very rare and one-off names from carrying very high weights. Although this approach does not get round the problem of the very different names that can be found in the same ONCA block (Block S530: contains Smith, Smithies, Smoothey, Snaith, Sneath, Samuda, Szabo, etc.) it does provide a higher level of discrimination and, in part, accommodate the erroneous names.

The theoretical weight based on the frequency of the surname in the studied population is modified according to the algorithm devised by Knuth-Morris-Pratt (Stephen, 1994; Gonnet and Baeza-Yates, 1991; and Baeza-Yates, 1989), and takes into account the length of the shortest of the two names being compared, the difference in length of the two names, the number of letters agreeing and the number of letters disagreeing. Where the two names are absolutely identical, the weight is set to +2N, but falls down to a lower bound of -2N where the amount of disagreement is quite large.

If the birth surname and present surname are swapped with each other, exploding the file as described previously enables the system to find and access the block containing the records for the appropriate surnames. The weights for the present and birth surname pairs are calculated, then the present surname/birth surname and birth surname/present surname pairs are also calculated. The highest of the two values is used in the subsequent calculations for the derivation of the match weight.

In cases where the marital status of the person is single, i.e., never married, or the sex is male, or the age is less that 16 years, it is normal practice in the UK for the present surname to be the same as the birth surname, and for this reason only the weight for the present surname is calculated and used for the determination of a match.

Forenames

The weights derived for the forenames are based on the frequency of the initial letter of the forename in the population. Since the distribution of male and female forenames are different, there are two sets of different weights, one for males and a second for females. Since the forenames can be recorded in any order, the weights for the two forenames are calculated and the highest value used for the match. Where there are wide variations in the spelling of the forenames, the Daitch-Motokoff version of Soundex ("Ask Glenda") is being evaluated for weighting the forenames in a fashion similar to that used for the surnames.

Calculating the Weights for the Non-Names Items

The weights for date of birth, sex, place of birth and NHS number are calculated using the frequency of the item on the ORLS and on the NHSCR file. The weight for the year of birth comparison has been extended to allow for known errors, for example, only a small deduction is made where the two years of birth differ by 1 or 10 years, but the weight is substantially reduced where the year of births differ by say, 7 years.

The weight for the street address is based on the first 8 characters of the full street address, where these characters signify a house number (31, High Street), or house name (High Trees), or indeed a public house name (THE RED LION). Terms like "Flat" or "Apartment" are ignored and other parts of the address are then used for the comparison. The postcode is treated and weighted as a single field although the inward and outward parts of the code can be weighted and used separately.

The range of binit weights used for the ORLS is shown in Table 1.

When the matching item is present on both the records, a weight is calculated expressing the amount of agreement or disagreement between the item on the data record and the corresponding item on the master file record.

Table 1. -- The Range of Binit Weights Used for Matching

Identifying Item		Score in Binits[1]		
		Exact Match	Partial Match	No Match
Surnames:	Birth	+2S	+2S to -2S	-2S
	Present[2]	+2S	+2S to -2S	-2S
	Mother's birth	+2S	+2S to -2S	-2S
	(where: common surname S = 6, rare surname S = 17)			
Forenames[3]		+2F	+2F to -2F	-2F
	(where: common forename F = 3, rare forename F = 12)			
NHS number		+7	NP[4]	0
Place of birth (code)		+4	+2	-4
Street address[5]		+7	NP	0
Post Code		+4	NP	0
GP (code)		+4	+2	0
Sex [6]		+1	NP	-10
Date of birth		+14	+13 -> -22	-23
Hospital and Hospital unit number		+7	NP	-9

[1]Where an item has been recorded as not known, the field has been left blank, or filled with an error flag, the match weight will be set to 0, except for special values described in the following notes.

[2]Where the surname is not known or has been entered as blank, the record can not be matched in the usual way, but is added to the file to enable true counts of all the events to be made.

[3]Forename entries, such as boy, girl, baby, infant, twin, or not known, are weighted as -10.

[4]Where the weight is shown as NP (not permissible), this partially known value cannot be weighted in the normal fashion and is treated as a NO MATCH.

[5]No fixed abode is scored 0.

[6]Where sex is not known, blank, or in error, it is scored -10. (All records input to the match are checked against forename/sex indices and the sex is corrected where it is missing or in error.)

It is possible for the calculated weight to become negative where there is extreme disagreement between the item on the data record and the corresponding item on the master file. In matching street address, postcode and general practitioner the score cannot go negative, although it can assume zero, because the individual may have changed their home address or their family doctor since they were last entered into the system, this is really a change in family circumstances and not errors in the data and so a negative weight is not justified.

Threshold Weighting

The procedure for deciding whether two records belong to the same person, was first developed by Newcombe, Kennedy, Axford, and James (1959), and rigorously examined by Copas and Hilton (1990), Belin and Rubin (1995), and Winkler (1995). The decision is based on the total binit weight, derived by summing algebraically the individual binit weights calculated from the comparisons of each identifying item on the master file and data file. The algebraic sum represents a measure of the probability that two records match. By comparing

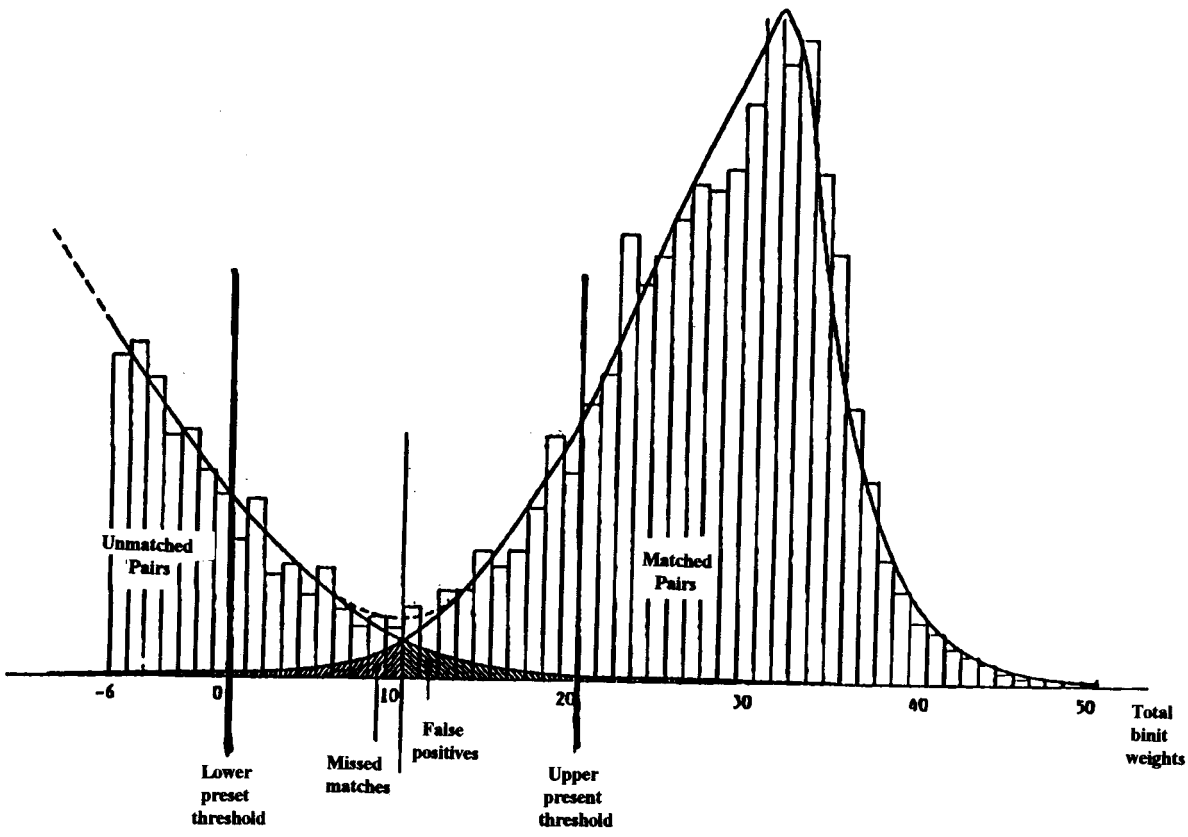
the total weight against a set of values determined empirically, it is possible to determine whether the two records being compared refer to the same person.

Two types of error can occur in record matching. The first, *false negative* matching or Type I error, is the more common and is a failure to collate records which refer to the same person and should have the same system number instead the person is assigned two or more person/system numbers and their records are not collated together. The second, *false positive* or Type II error, is less common but potentially more serious in allocating the same system number to two or more persons, where their records are wrongly collated together. The frequency of both types of error is a sound measure of the reliability of the record matching procedure.

In preparing earlier versions of the ORLS linked files, a range of binit weights was chosen and used to select records for clerical scrutiny. This range was delimited by the upper and lower pre-set thresholds, see Figure 1. The *false positive* and *false negatives* are very sensitive to the threshold cut-off weight: too low gives a very low *false positive* rate and a high *false negative* rate; too high gives a high and unacceptable *false positive* rate with a low *false negative* rate. The values selected for the threshold cut-off are, of course, arbitrary, but must be chosen with care, having considered the following objectives:

- The minimisation of *false positives*, at the risk of increased missed matches;
- The minimisation of missed matches, at the risk of increased false positives; and
- The minimisation of the sum of *false positives* and *missed matches*.

Figure 1. — Frequency Distribution of the Binit Weights for Pairs of Records



The simple approach for the determination of a match based on the algebraic sum of the binit weights, ignores the fact that the weight calculated for names is based on the degree of commonness of the name, and is passed on from other members of the family, whereas the weight for the non-names items are based on distributions of those items in the population, all values of which are equally probable.

An unusual set of rare names information would generate high weights which would completely swamp any weights calculated for the non-names items in the algebraic total, and conversely, a common name would be swamped by a perfect and identical set of non-names identifiers. This would make it difficult for the computer algorithm to differentiate between similarly-named members of the population without resort to clerical assistance.

In the determination the match threshold, a number of approaches have been developed, the earliest being the two stage primary and secondary match used in building the early ORLS files, through a graphical approach developed in Canada for the date of birth, to the smoothed two dimensional grid approach developed by the UHCE and used for all its more recent matching and linking (Gill, et al., 1995; Vitter and Wen-Chin, 1987).

Algebraic Summation of the Individual Match Weights

In recent years we have, therefore, developed an approach in which a two dimensional orthogonal matrix is prepared, analogous to a spreadsheet, with the names scores forming one axis and the non-names scores the other axis. In the development of the method, sample runs are undertaken; pairs of records in cells in the matrix are checked clerically to determine whether they do or do not match; and the probability of matching is derived for each cell in the sample. These probabilities are stored in the cells of an orthogonal matrix designated by the coordinates (names score, non-names score). The empirical probabilities entered into the matrix are further interpolated and smoothed across the axes using linear regression methods.

Match runs using similar data types would access the matrix and extract the probability score from the cell designated by the coordinates. The array of probabilities can be amended after experience with further runs, although minor tinkering is discouraged. Precise scores and probabilities may vary according to the population and record pairs studied. A number of matrices have therefore been prepared for the different types of event pairs being matched, for example, hospital to hospital records, hospital to death records, birth to hospital records, hospital and District Health Authority (DHA) records, cancer registry and hospital records, and so on.

Over 200,000 matches were clerically scrutinized and the results recorded in the two axes of a orthogonal matrix, with the algebraic sum of the weights for the names items being X coordinate ("X" axis), and the algebraic sum of the fixed and variable statistics items plotted on the Y coordinate ("Y" axis). In each cell of the orthogonal matrix the results of the matches were recorded, with each cell holding the total number of matches, the number of good matches and the number of non-matches. A sample portion of the matrix is shown in Figure 2.

A graphical representation of the matrix is shown in Figure 3, where each cell contains the empirical decision about the likelihood of a match between a record pair. The good matches are shown as "Y," the non-matches as "N" and the doubtful matches that require clerical intervention as "Q." This graph is the positive quadrant where both the names and non-names weights are greater that zero. In the microcomputer implementation of the software, this graph is held as a text file and can be edited using word-processing software.

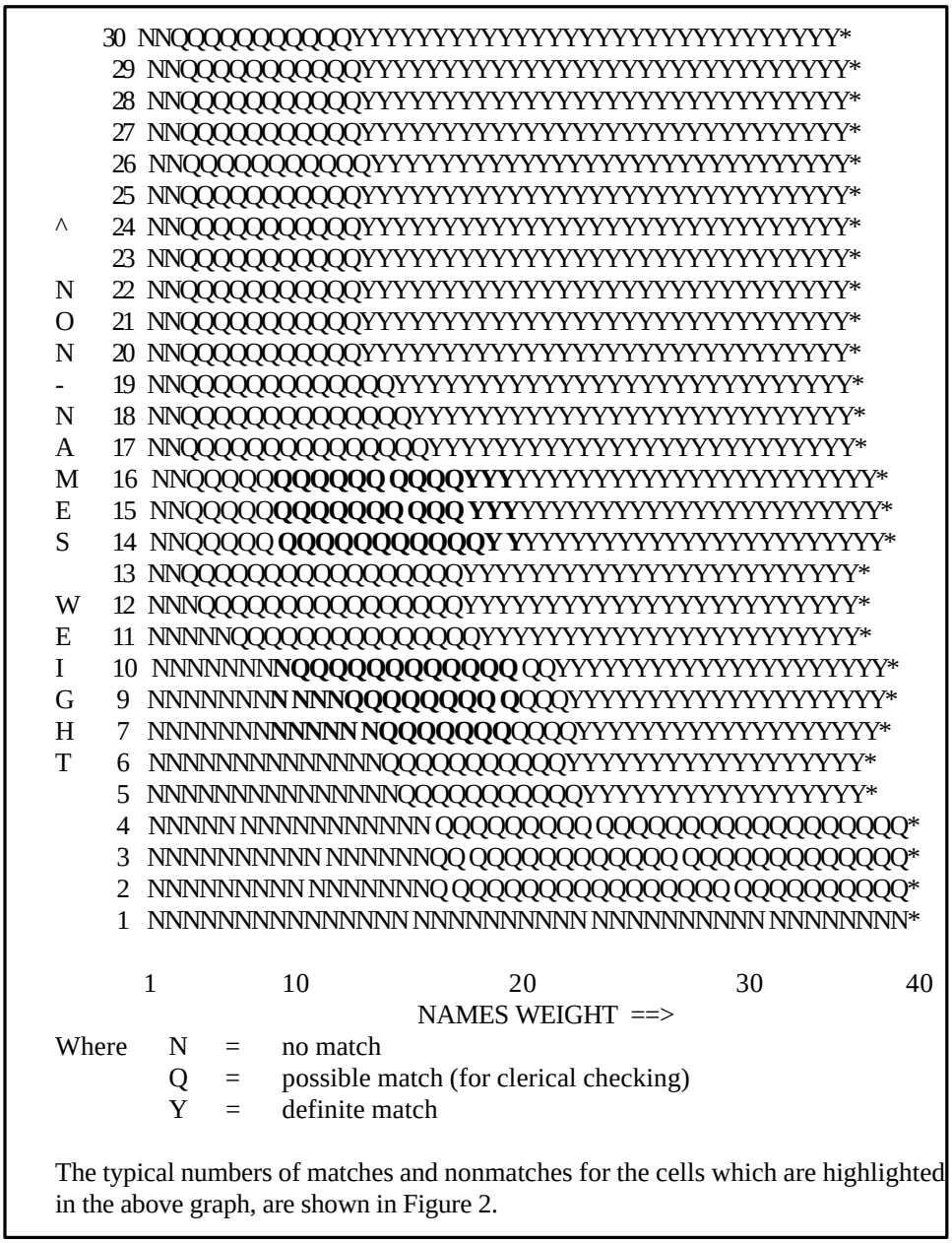
Figure 2. — Sample Portion of the Threshold Acceptance Matrix Showing the Number of Matches and Nonmatches, by Binit Weight for Names and Non-names Identifiers

WT=16	Percentage	37	41	45	58	83	77	91	98	99	99		
	Matches	198	177	231	255	319	277	413	298				
	Nonmatches	537	255	298	145	65	83	41	4				
WT=15	Percentage	41	38	42	56	61	75	87	98	99	99		
	Matches	190	223	211	316	410	329	218	523	322			
	Nonmatches	273	364	293	245	265	109	33	11	4			
WT=14	Percentage	18	25	21	19	31	56	77	93	89	97	99	
	Matches	113	87	90	110	190	198	660	422	161	377		
	Nonmatches	514	261	330	460	412	162	197	34	19	11		
WT=10	Percentage	4	7	8	8	14	11	22	26				
	Matches	17	35	28	34	50	50	69	75				
	Nonmatches	341	404	284	382	277	295	235	203				
WT=9	Percentage	2	4	4	4	8	12	13	15				
	Matches	18	42	28	47	64	90	90	87				
	Nonmatches	737	966	637	952	706	644	588	474				
WT=8	Percentage	2	7	7	9	12	16	20	22				
	Matches		95	70	118	113	140	147	170				
	Nonmatches		1,234	812	1,106	785	728	583	588				
WT=7	Percentage	0	1	1	1	2	2	3	4				
	Matches	5	45	43	55	58	57	68	93				
	Nonmatches	2,721	3,919	2,733	3,576	2,458	2,542	1,952	1,848				

Record pairs with weights that fall in the upper right part of the matrix and shown in Figure 3 as "Y" are considered to be “good” matches and only a 1% random sample is printed out for clerical scrutiny. Record pairs with weights that fall between the upper and lower thresholds and shown in the figure as "Q" are considered to be “query” matches and all the record pairs are printed out for clerical scrutiny and the results keyed back into the computing system. Record pairs with weights falling below the lower threshold and shown on the map as "N" are considered to belong to two different people and a 1% random sample is taken of record pairs that fall adjacent to N-Q boundary.

At the end of each computer run, the results of the clerical scrutiny are pooled with all the existing matching results and new matrices are prepared. The requirement is to reduce the “Q” zone to the minimum consistent with the constraints of minimum false positives and false negatives. Clerical intervention is invariably the most costly and rate determining stage.

Figure 3. -- A Sample Portion of the Matrix Used for Matching Hospital Records with Hospital Records



Where N = no match
Q = possible match (for clerical checking)
Y = definite match

The typical numbers of matches and nonmatches for the cells which are highlighted in the above graph, are shown in Figure 2.

Separate matrices have been modelled for the different types of record pairs entering the system, for example:

hospital discharge / hospital discharge
hospital discharge / death record
birth record / hospital discharge
hospital discharge / primary care/FHSA record
hospital discharge / Cancer registry.

Further matrices have also been prepared that record the number of match items used in matching a record pair, for example, number of surnames, forenames and numbers of other matching variables. Since the number of matrices can become quite large, intelligent systems and neural net techniques are being developed for the interpretation of the N dimensional matrices and the determination of the match threshold (Kasabov, 1996; Bishop, 1995).

Special procedures have been developed for the correct matching of similarly-named same sex twins. Where the match weights fall within the clerical scrutiny area, the clerks are able to identify the two records involved and take the appropriate action.

The marked records are printed out for clerical scrutiny and the match amended where necessary. This situation also arises where older people are recorded in the information system under a given set of forenames but, on a subsequent hospital admission or when they die, a different set of forenames are reported by the patient or by the next of kin.

Linking

The output from the matching run, is a text file that contains details about each pair of records that were matched together. A sample portion of this file is shown in Figure 4, the layout of which is:

Details of data record	Person/system number Accession number Record type
Details of main file record	Person/system number Accession number Record type
Details about the match run	Output stream (good match or query match) Names weight Non-names weight Cross-reference to the clerical printout Matching probability/decision (either Y or N).

The number of records written to the output file for any one person can be very large, and is approximately the number of records on data file multiplied by the number of records on the master file. Using combinational and heuristic algebraic methods these records are reduced to a small number for each potential match pair, ideally one (Hu, 1982; Cameron, 1994; Lothaire, 1997; and Pidd, 1996).

Figure 4.—A Sample of the Typical Output from the Match Run

Example of OX-LINK System Number Output

389447756	860895558	GS	229800034	352-68394	GN	2	50	26	(GH1/ 500001)	Y	O
379194856	858751858	GS	233513082	369890337	GN	2	29	24	(GH1/ 500002)	Y	O
379194856	858751858	GS	233513082	911759078	TU	2	29	15	(GH1/ 500003)	Y	O
379194856	858751858	GS	233513082	911759078	TU	2	29	15	(GH1/ 500004)	Y	O
437096752	781384114	GS	323947927	524582350	BL	2	31	19	(GH1/ 500005)	Y	O
357816810	726892961	GS	249173530	472792138	GN	2	31	23	(GH1/ 500006)	Y	O
357816810	726892961	GS	249173530	343537893	GN	2	31	21	(GH1/ 500007)	Y	O
357816810	726892961	GS	249173530	406349427	GN	2	31	23	(GH1/ 500008)	Y	O
540814037	883641514	GS	210500551	448983383	GM	2	50	19	(GH1/ 500009)	Y	O
110463907	559719951	GN	408578989	738005030	GS	2	50	30	(GH1/ 500010)	Y	O
110463907	262969219	GH	408578989	738005030	GS	2	50	30	(GH1/ 500011)	Y	O
110463907	63685552	GH	408578989	738005030	GS	2	50	26	(GH1/ 500012)	Y	O
133714360	188729480	GH	414567239	748873845	GS	2	50	25	(GH1/ 500013)	Y	O
133714360	205039688	GH	414567239	748873845	GS	2	50	23	(GH1/ 500014)	Y	O

The rules for undertaking this reduction are:

- Ideally, all records for the same person will have the same person/system number.

- The records for a person who has only one set of identification details will be of the following type, where each record only carries one person/system number (A):

  A = A = A = A, etc. (= signifies matches with).

- Where a single woman gets married within the span of the file, records will be recorded under maiden name, person/system number (A) and also under her married name (B). Links will be effected between (A) and (B) and all the records will be converted to person/system number (A). The person/system number (B) will be lost to the system. Future matches will link to either her single or married records, both of which will carry the person/system number (A):

  A = A = B = B = A = B, etc.

  A being links under her maiden name
  B being links under her married name.

- Where there are records for a women recorded under her maiden name (A), and records that contain details of both her maiden and married name (B) and just her married name (C), these chains are will be made up of three types of links,

  A = A = B = B = C = B = C, etc.

  Successive matches will convert all the records to person/system number (A). If the linked file contains records type (A) and (C) only, linkage cannot be effected between (A) and (C) until records of type (B) are captured and linked into the system.

- Where the person has had many changes of name and marital status, the number of different types of links will increase. Over the 30 year span of the file, links up to 5 deep have been found.

Each record entering the system is given a new purely arbitrary person/system number from a pool of such numbers. Where the record on the data file matches with a record on the master file, the person/system number stored on the master file record is copied over the person/system number on the data record, over-writes it, and the original person/system number on the data record number is then lost from the system and cannot be re-issued.

Where two sets of records for the same person, but having two different person/system numbers are brought together during a subsequent matching run; all the records are given the lowest person/system number and any other person/system numbers are destroyed.

Results

When the matching, linking and clerical stages are completed, the file of linked records will contain two types of error. Firstly, the records that have matched together but do not belong to the same person, these are known as *false positives*. Secondly, records belonging to the same person that have not been brought together, i.e., reside on the file under two or more different person identifiers, these are known as "*false negatives or missed matches*."

The *false positive* rate was estimated using two different methods. Firstly, all the records for a random sample of 5,000 people having two or more records were extracted from the ORLS file and printed out for clerical scrutiny. Secondly, all the record pairs that matched together with high match weights but where the forenames differed, were printed out for clerical scrutiny.

The "*false negative or missed match*" rate was estimated, by extracting a subset of people who had continuing treatment, such as repeated admissions for diabetics, nephritics, etc., and for those patients who had died in hospital, where the linked file should contain both the hospital discharge record and the death record.

The latest results from the ORLS file and the Welsh and Oxfordshire Cancer registry files are very encouraging, with the *false positive* rate being below 0.25 percent of all people on the file, and the *missed match* rate varying between 1.0 percent and 3.0 percent according to the type of sample investigated. Recent works on matching 369,000 records from a health district with 71 million *exploded* records from NHS Central Register has given a *false positive* rate of between 0.2 and 0.3%; the higher figure is produced from records which have very common Anglo-Saxon or Asian names.

The worst *false negative* rate was found where hospital discharges were matched with the corresponding death record. The identifying information on the hospital discharge was drawn from the hospital master index supplemented by information supplied by the patient or immediate family. The identifying information on the death record is usually provided by the next of kin from memory and old documents.

The completed ORLS file is serial file that is indexed using the person/system number, and contains the partial identifiers, administrative and socio-demographic variables and clinical items. This file used for a wide range of epidemiological and health services research studies. For ease of manipulation and other operational reasons, subsets of the file are prepared for specific studies, usually by selecting specified records or record types, or by selecting on geographical area or span of years or on clinical specialty.

Acknowledgments

The Unit of Health Care Epidemiology and the work on medical record linkage is funded by the Research and Development Directorate of the Anglia and Oxford Regional Health Authority. The Office of Population Censuses and Surveys (now the Office of National Statistics) for permission to publish the frequencies of the surnames from the NHS central register.

References

Acheson E.D. (1967). *Medical Record Linkage*, Oxford: Oxford University Press.

Acheson E.D. (ed) (1968). Record Linkage in Medicine, *Proceedings of the International Symposium,Oxford, July 1967*, London: ES Livingstone Limited.

"Ask Glenda," Soundex History and Methods, World Wide Web: *http://roxy.sfo.com/~genealogysf/glenda.html* .

Baeza-Yates, R.A. (1989). Improved String Searching, *Software Practice and Experience*, 19, 257-271.

Baldwin, J.A. and Gill, L.E. (1982). The District Number: A Comparative Test of Some Record Matching Methods, *Community Medicine*, 4, 265-275.

Belin, T.R. and Rubin, D.B. (1995). A Method for Calibrating False-Match Rates in Record Linkage, *Journal of the American Statistical Association*, 90, 694-707.

Bishop, C.M. (1995). Three Layer Networks, in*: Neural Networks for Pattern Recognition*, United Kingdom: Oxford University Press, 128-129.

Cameron, P.J. (1994). Graphs, Trees and Forests, in: *Combinatorics*. United Kingdom: Cambridge University Press, 159-186.

Copas, J.R. and Hilton, F.J. (1990). Record Linkage: Statistical Models for Matching Computer Records, *Journal of the Royal Statistical Association*, Series A, 153, 287-320.

Dolby, J.L. (1970). An Algorithm for Variable Length Proper-Name Compression, *Journal of Library Automation*, 3/4, 257.

Dunn, H.L. (1946). Record Linkage, *American Journal of Public Health* , 36, 1412-1416.

Gallian, J.A. (1989). Check Digit Methods, *International Journal of Applied Engineering Education*, 5, 503-505.

Gill, L.E. and Baldwin, J.A. (1987). Methods and Technology of Record Linkage: Some Practical Considerations, in: *Textbook of Medical Record Linkage* (Baldwin,J.A., Acheson, E.D., and Graham, W.J., eds). Oxford: Oxford University Press, 39-54.

Gill, L.E.; Goldacre, M.J.; Simmons, H.M.; Bettley, G.A.; and Griffith, M. (1993). Computerised Linkage of Medical Records: Methodological Guidelines, *Journal of Epidemiology and Community Health*, 47, 316-319.

Goldacre, M.J. (1986). The Oxford Record Linkage Study: Current Position and Future Prospects, *Proceedings of the Workshop on Computerised Record Linkage in Health Research* (Howe, G.R. and Spasoff, R.A., eds). Toronto: University of Toronto Press, 106-129.

Gonnet, G.H. and Baeza-Yates, R. (1991). Boyer-Moore Text Searching, *Handbook of Algorithms and Data Structure*, 2nd ed, United States: Addison-Wesley Publishing Co Inc, 256-259

Hamming, R.W. (1986). *Coding and Information Theory*, 2nd ed., Englewood Cliffs, NJ: Prentice Hall.

Holmes, W.N. (1975). Identification Number Design, *The Computer Journal*, 14, 102-107.

Hu, T.C. (1982). Heuristic Algorithms, in: *Combinatorial Algorithms*., United States: Addison-Wesley Publishing Co. Inc, 202-239.

Kasabov, N.K. (1996). Kohonen Self-Organising Topological Maps, in*: Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering*, Cambridge, MA, USA: MIT Press, 293-298.

Knuth,D.E. (1973). Sorting and Searching, in: *The Art of Computer Programming*, 3, United States: Addison-Wesley Publishing Co. Inc., 391.

Lothaire, M. (1997). Words and Trees, in: *Combinatorics on Words*, United Kingdom: Cambridge University Press, 213-227.

Lynch, B.T. and Arends, W.L. (1977). *Selection of a Surname Encoding Procedure for the Statistical Reporting Service Record Linkage System*, Washington, DC: United States Department of Agriculture.

National Health Service and Department of Health (1990*). Working for Patients: Framework for Implementing Systems:The Next Steps*, London: HMSO.

Newcombe, H.B. (1967). The Design of Efficiency Systems for Linking Records into Individual and Family Histories, *American Journal of Human Genetics* , 19, 335-339.

Newcombe, H.B. (1987). Record Linking: The Design of Efficiency Systems for Linking Records into Individual and Family Histories, in*: Textbook of Medical Record Linkage* (Baldwin, J.A.; Acheson, E.D.; and Graham, W.J., eds), Oxford: Oxford University Press, 39-54.

Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press.

Newcombe, H.B.; Kennedy, J.M.; Axford, S.J.; and James, A.P. (1959). Automatic Linkage of Vital Records, *Science*, 130, 3381, 954-959.

Pidd, M. (1996). Heuristic Approaches, *Tools for Thinking, Modelling in Management Science*, England: John Wiley and Sons, 281-310.

Scheuren, F. and Winkler, W.E. (1996). Recursive Merging and Analysis of Administrative Lists and Data, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Secretaries of State for Health, Wales, Northern Ireland and Scotland (1989). *Working for Patients*., London: HMSO, CM 555.

Stephen, G.A. (1994). Knuth-Morris-Pratt Algorithm, in: *String Searching Algorithms*, Singapore: World Scientific Publishing Co. Pte. Ltd, 6-25.

Vitter, J.S and Wen-Chin,C. (1987). The Probability Model, *Design and Analysis of Coalesced Hashing*, United Kingdom: Oxford University Press, 22-31.

Wild, W.G. (1968). The Theory of Modulus *N* Check Digit Systems, *The Computer Bulletin*, 12, 308-311.

Winkler,W.E. (1995). Matching and Record Linkage, *Business Survey Methods* (Cox, Binder, Chinnappa, Christianson, Culledge, and Kott, eds.), New York: John Wiley and Sons, Inc., 355-384.

Chapter 2

Complex Linkages Made Easy

John R. H. Charlton, Office for National Statistics, UK
Judith D. Charlton, JDC Applications

Abstract

Once valid key fields have been set up, relational database techniques enable complex linkages that facilitate a number of statistical analyses. Using one particular example, a classification of types of linkages is developed and illustrated. The naive user of such data would not necessarily know how to use a relational database to perform the linkages, but may only know the sort of questions they want to ask. To make data (anonymous to protect the confidentiality of patients and doctors) generally accessible, a user-friendly front-end has been written using the above concepts, which provides flat-file datasets (tabular or list) in response to answers from a series of questions. These datasets can be exported in a variety of standard formats. The software will be demonstrated, using a sample of the data.

Introduction

Most of the papers at this workshop are concerned with establishing whether or not different records in the database match. This paper starts from the point where this matching has already been established. It will thus be assumed that the data have already been cleaned, duplicates eliminated, and keys constructed through which linkages can be made. Procedures for matching records when linkage is not certain have been discussed for example by Newcombe *et al.* (1959, 1988), Fellegi and Sunter (1969), and Winkler (1994). We also assume database software that can:

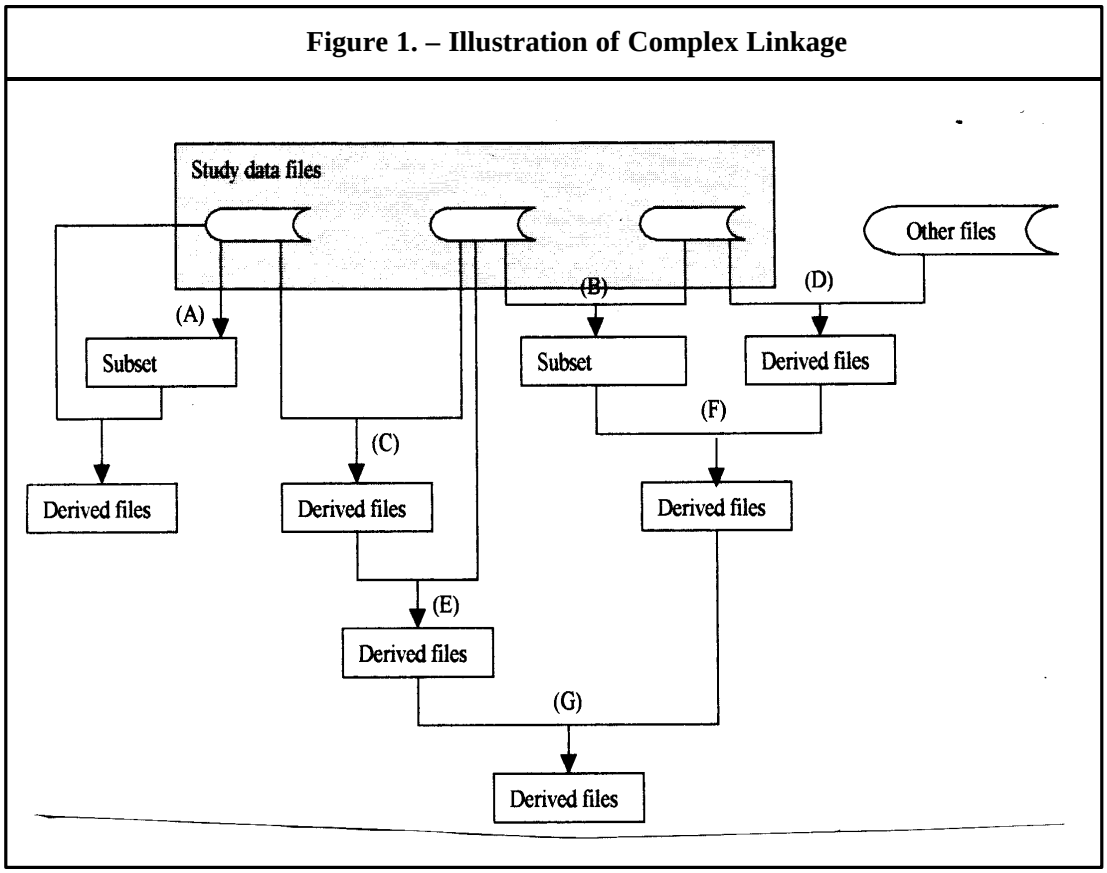
- select fields from a file of records;
- extract from a file either;
  - all records
  - distinct records which satisfy specified criteria; and
- link files using appropriate key and foreign fields.

The purpose of this paper is firstly to illustrate the huge potential of using relational databases for data linkage for statistical analyses. In the process a classification of linkages will be developed, using a particular database to illustrate the points. Some results will be presented by way of example. We will show how the complex linkages required for statistical analyses can be decomposed into a sequence of simple database queries and linkages. Finally a user-friendly program that has been written for extracting a number of different types of dataset for analyses will be described. The advantages and disadvantages of such approaches will be discussed.

Relational databases are ideal for storing statistical data, since they retain the original linkages in the data, and hence the full data structure. They also facilitate linking in new data from other sources, and are economical in data storage requirements. However, most statistical analyses require simple rectangular files, and complex database queries may be required to obtain these.

The linkages required to obtain the flat files for statistical analysis vary from the relatively simple to the extremely complex (Figure 1). Subsets of data files may be found (A), possibly by linkage with another file (B). Derived files may be created by linking files (or their subsets) within the study data files (C), or to files outside the study data (D). The derived files may be further linked to files in or outside the dataset (E), subsets (F), or other derived files (G), to obtain further derived files, and this process may continue at length.

Figure 1. – Illustration of Complex Linkage

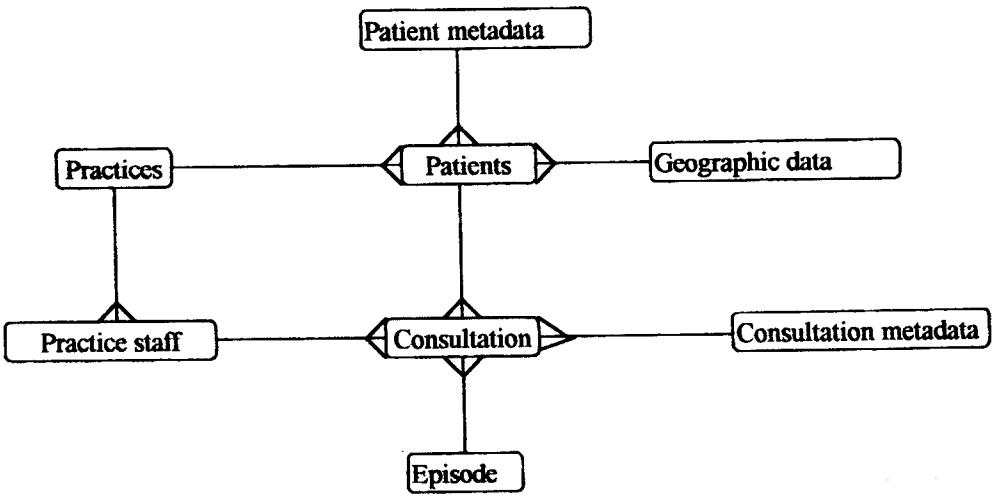


The Example Database

In a major survey in England and Wales (MSGP4) some 300 general medical practitioners (GPs) in 60 practices collected data from half a million patients, relating to every face to face contact with them over the course of a year (McCormick *et al.*, 1995). In the UK nearly the entire population is registered with a GP, and only visit a doctor in the practice in which they are registered, except in an emergency, when they may attend an accident and emergency department of a hospital or another GP as a temporary patient. For all patients in MSGP4 there was information on age, sex, and postcode. In addition socio-economic data were successfully collected by interview for 83 per cent of the patients on these doctors registers. There was a core of common questions, but there were also questions specific to children, adults, and married/cohabiting women. Information was also collected about the practices (but not individual GPs). Geographic information related to postcodes was also available. The structure of the data is illustrated in

Figure 2 (simplified). MSGP4 was the fourth survey of morbidity in general practice. In previous MSGP surveys output consisted only as a series of tables produced by COBOL programs, and MSGP4 was the fist survey for which relational databases were used to provide flexible outputs.

Figure 2.—Example Database (Simplified) – Patient Consultations of General Medical Practitioners



Some Definitions

- **Read Code**.--A code used in England and Wales by general practice staff to identify uniquely a medical term. This coding was used in the MSGP project because it is familiar to general practice staff, but it is not internationally recognised and the codes have a structure that does not facilitate verification.

- **ICD Code**.--International classification of disease. Groups of Read codes can be mapped onto ICD9 codes. For example Read code F3810= "acute serious otitis media," maps to ICD A381.0 = "acute nonsuppurative otitis media"). Such mappings form part of the consultation metadata (see below).

- **Consultation**.--A "consultation" refers to a particular diagnosis by a particular member of staff on a particular date at a particular location, resulting from a face to face meeting between a patient and doctor/ nurse. A "diagnosis" is identified by a single Read code.

- **"Patients Consulting"**.--Some registered patients did not consult a doctor or other staff member during the study. "Patients consulting" is therefore a subset of the practice list of all registered patients. Consultations must be carefully distinguished from "patients consulting." A combination of patient number, date and place of consultation and diagnosis uniquely define each record in the consultation file. Patient numbers are not unique because a patient may consult more than once, nor are combinations of patient number and diagnosis unique. On the other hand, a "patient consulting" file will contain at most one record for each patient consulting for a particular diagnosis (or group of diagnoses), no matter how many times that patient has consulted a member of the practice staff. "Consultations" are more relevant when work-load is being studied, but if

prevalence is the issue then "patients consulting," i.e., how many patients consulted for the illness, is more useful.

- **Patient Years at Risk**.--The population involved in the MSGP project did not remain constant throughout the study. Patients entered and left practices as a result of moving house or for other reasons, and births and deaths also contributed to a changing population. The "patient years at risk" derived variable was created to take account of this. The patient file contains a "days in" variable, which gives the number of days the patient was registered with the practice (range 1-366 days for the study). "Patient years at risk" is "days in" divided by 366, since 1992 was a leap year.

Database Structure

To facilitate future analyses some non-changing data were combined at the outset. For example some consultation metadata were added to the consultation dataset, such as International Classification of Disease (ICD) codes and indicators of disease seriousness. The resultant simplified data structure is thus:

Practice: Practice number; information about practice (confidential)
Primary Key: Practice number

A practice is a group of doctors, nurses, and other staff working together. Although patients register with a particular doctor, their records are kept by the practice and the patient may be regarded as belonging to a practice. Data on practice and practice staff are particularly confidential, and not considered in this paper. Individual practice staff consulted are identified in the consultation file by a code.

Patients: Patient number; age; sex; post code; socio-economic data
Primary key: Patient number
Foreign key: Postcode references geographic data

These data were stored as four separate files relating to: all patients; adult patients; children; married cohabiting women, because different information was collected for each subgroup.

Consultation: Patient number; Practice number; ID of who consulted; date of contact; diagnosis; place of consultation; whether referred to hospital; other consultation information
Primary key: Patient number, doctor ID, date of contact, diagnosis
Foreign keys: Practice number references practice; Patient number references patients; Staff ID references staff (e.g., doctor/nurse).

Episode: For each consultation the doctor/nurse identified whether this was the "first ever," a "new," or "ongoing" consultation for that problem. An "episode" consists of a series of consultations for the same problem (e.g., Read code).

Geographically-referenced data: Post codes, ED, latitude/longitude, census ward, local authority, small area census data, locality classifications such as rural/ urban, prosperous/inner city, etc.

These data were not collected by the survey, but come from other sources, linked by postcode or higher level geography.

Patient metadata: These describe the codes used in the socio-economic survey (e.g., ethnic group, occupation groups, social class, housing tenure, whether a smoker, etc.)

Consultation metadata: The ReadICD file links Read codes with the corresponding ICD codes. In addition a lookup table links 150 common diseases, immunisations and accidents to their ICD codes. Each diagnosis is classified as serious, intermediate or minor.

Derived files: The MSGP database contains information on individual patients and consultations. To make comparisons between groups of patients, and to standardise the data (e.g., for age differences), it is necessary to generate files of derived data, using database queries and linkages as described below. In some derived files duplicate records need to be eliminated. For example, we may wish to count **patients** consulting for a particular reason rather than consultations, and hence wish to produce at most one record per patient in a "patients consulting" derived file -- see "Some Definitions above).

Types of Linkage (with Examples)

In this section we classify a variety of linkage types that are possible into three main types, illustrating the linkages with examples based on the MSGP4 study.

Simple Linkage

- Straightforward data extracts (lists) combining several sources.—

  Example: Making a list of patients with asthma including age, sex and social class for each patient.

- Observed frequencies.—

  Example: Linking the "all patients" file, and the "consultations" file to count the number of consultations by the age, sex and social class of the patient, or cross-classifying home-visits and hospital referrals with socio-economic characteristics.

- Conditional data, where the availability of data items depends on the value of another variable.—

  Example: In MSGP4 some data are available only for adults, or children, or married/cohabiting women. Smoking status was only obtained from adult patients, so tabulating "home visits" by "smoking status" by "age," and "sex" involves linking the "all patients" (to find age and sex), "adult patients" (to find smoking status) and "consultations" (to find home visits) files. Linking the "adult" file to the "all patients" file excludes records for children.

- Linking files with "foreign" files.— Useful information can often be obtained by linking data in two or more different datasets, where the data files share common codes. For example data referenced by postcode, census ED or ward, or local authority are available from many different sources as described above.

  Example: The MSGP4 study included the postcode of residence for each patient, facilitating studies of neighbourhood effects. The crow-fly distance from the patient's home to the practice was calculated by linking patient and practice postcodes to a grid co-ordinates file and using Pythagoras's theorem. The distance was stored permanently on the patient file for future use.
- Linking to lookup tables (user-defined and pre-defined).—

  Examples: The information in the MSGP database is mostly held in coded form, with the keys to the codes held in a number of lookup tables linked to the main database. Most of these are quite small and simple (e.g., ethnic group, housing tenure, etc.) but some variables are linked to large

tables of standard codes (e.g., occupational codes, country of birth). . In some cases the coded information is quite detailed and it is desirable to group the data into broader categories, e.g., group diagnostic codes into broad diagnostic groups such as ischaemic heart disease ICD 410-414. For some diseases a group of not necessarily contiguous codes are needed to define a medical condition. A lookup file of these codes can be created to extract the codes of interest from the main data, using a lookup table that could be user-defined. Missing value codes could also be grouped, ages grouped into broad age groups, social classes combined, etc.

Auto-Linkage Within a File (Associations Within a File)

- Different records for the same "individual."— Records for the same individual can be linked together to analyse patterns or sums of events, or associations between events of different kinds. In general a file is linked to a subset of itself to find records relating to individuals of interest.

  Example: Diabetes is a chronic disease with major complications. It is of interest to examine, for those patients who consulted for diabetes, what other diseases they consulted for. Consultations for diabetes can be found from their ICD code (250). Extracting just the patient identification numbers from this dataset, and eliminating duplicates, results in a list of patients who consulted for diabetes at least once during the year. This subset of the consultation file can be linked with the original consultation file to produce a derived file containing the consultation history of all diabetic patients in the study, which can be used for further analysis. Note that in this example only the consultation file (and derived subsets) has been used.

- Different records for same households/ other groups.—

  Example: Information on households was not collected as part of MSGP4. However "synthetic" households can be constructed, using postcode and socio-economic data, where the members of the same "household" must, by definition, share the same socio-economic characteristics and it would be rare for two distinct households to have exactly the same characteristics. These "households" can be used to discover how the behaviour of one "household" member may affect another. For example, we can examine the relationship between smoking by adults, and asthma in children. Clearly in this example some sort of check needs to be made on how accurately "households" can be assembled from the information available and the algorithm used.

- Temporal relationships.— Files containing "event" data can be analysed by studying temporal patterns relating to the same individual.
  Example:
  - The relationship between exposure to pollution or infection and asthma can be studied in terms of both immediate and delayed effects. Consultations for an individual can be linked together and sorted by date, showing temporal relationships.

  - The duration of clinical events can sometimes be determined by the sequence of consultations. In MSGP4 each consultation for a particular medical condition was labelled "first ever," "new," or "ongoing" and the date of each consultation recorded. Survival analysis techniques cater for these types of data.

Complex Linkages

Linkages that are combinations of the two types of linkage previously described could be termed "complex linkages." These can always be broken down into a sequence of simpler linkages. A number of examples of complex linkages are given, in order of complexity.

- Finding subsets through linkage.—

  Example: In the MSGP4 data this is particularly useful in the study of chronic conditions such as diabetes and heart disease. Linking the file of patients consulting for diabetes discussed in section 3.2 with the patient dataset results in a subset of the patient file, containing only socio-economic details of diabetic patients.

- Linking a derived file to a lookup table and other files.—

  Example: Diabetes is particularly associated with diseases of the eye (retinopathy), kidney, nervous system and cardiovascular systems. It is of interest to analyse the relationship between diabetes and such diseases, which are likely to be related to diabetes. In this slightly more complex situation it is necessary to create a lookup table containing the diseases of interest and their ICD codes and link this to the "consultations by diabetic patients" file to create a further subset of the consultation file containing consultations for diabetes and its complications. It is likely that this file as well as the simpler one described above would be linked to the patient file to include age and sex and other patient characteristics before analysis using conventional statistical packages.

- Linking a derived file with another derived file.—

  Example:
  - Rates for groups of individuals.— Rates are found by linking a derived file of numerators with a derived file of denominators. The numerators are usually found by linking the patient and consultation files, for example, age, sex, social class or ethnic group linked to diagnosis, referral or home visits. Denominators can be derived from the patient file (patient years at risk) or the consultation file (consultations or patients consulting) for the various categories age, sex, etc.

  - Standardised ratios.— This is the ratio of the number of events (e.g., consultations or deaths) observed in a sub-group to the number that would be expected if the sub-group had the same age-sex-specific rates as a standard population (e.g., the whole sample), multiplied by 100. Examples of sub-groups are different ethnic groups or geographical areas. The calculation of standard population rates involves linking the whole population observed frequencies to whole population patient years at risk. Each of these is a derived file, and the result is a new derived file. Calculating expected numbers involves linking standard population rates to the sub-groups' "years at risk" file. This produces two new derived files, "Observed" and "Expected." Age-standardised patient consulting ratios are obtained by linking these two derived files together, using outer joins to ensure no loss of "expected" records where there are no observed in some age-sex categories.

- Establishing population rates for a series of nested definitions.—

  Example: Individuals at particular risk from influenza are offered vaccination. In order to estimate how changes in the recommendations might affect the numbers eligible for vaccination, population rates for those living in their own homes were estimated for each of several options. People aged 65 and over living in communal establishments are automatically eligible for vaccination, and hence were selected out and treated separately. The options tested were to include patients with:

  A - any chronic respiratory disease, chronic heart disease, endocrine disease, or immune-suppression;

B - as A but also including hereditary degenerative diseases;
C - as B but also including thyroid disease;
D - as C but also including essential hypertension.

The MSGP dataset was used to estimate the proportion of the population in need of vaccination against influenza according to each option. The problem was to find all those patients who had consulted for any of the diseases on the list, taking care not to count any patient more than once. This involved creating a lookup table defining the disease groups mentioned in options A-D, linking this to the consultation dataset, eliminating duplicates and linking this to the patient dataset (to obtain age-group and sex), and then doing a series of queries to obtain appropriate numerator data files. A denominator data file was separately obtained from the patient dataset to obtain patient years at risk, by age-group and sex. The numerator and denominator files were then joined to obtain rates. These rates were then applied to census tables to obtain the estimated numbers of patients eligible for vaccination under assumptions A-D.

- Record matching for case-control studies.— These are special studies of association-extracting "cases" and "controls" from the same database.

  Example: what socio-economic factors are associated with increased risk of Crohn's disease? All patients who consulted for ICD555 (regional non-infective enteritis) during the MSGP4 study were selected and referred back to their GP to confirm that they were genuine cases of Crohn's disease. Patients who were not confirmed as having Crohn's disease were then excluded. This resulted in 294 cases. Controls were selected from patients who did have the disease – those who matched cases for practice, sex and month and year of birth. In each of two practices there were two cases who were of the same sex and the same month and year of birth. In each of these practices the controls were divided randomly between these cases as equally as possible. There were 23 cases for whom no controls could be found using these criteria. In 20 of these cases it was possible to find controls who matched on practice and sex and whose date of birth was within two months of the case's date of birth. The remaining three cases were excluded from the analysis. This procedure resulted in 291 cases and 1682 controls.

User-Friendly Linkage Software

The MSGP4 practice software was originally written so that participating practices could gain access to the data collected from their own practice. The software was designed to be used easily by people with no knowledge of database technology and because the software runs directly under DOS or Windows, no specialised database software is needed. The structure of the MSGP database is transparent to the user who can refer to entities (e.g., diseases or occupation) by name rather than codes.

Later, a modified version of the software was developed to enable researchers to use the complete dataset (60 practices).

Although it may be possible for some of these linkages to be performed as a single query it is generally best to do a series of simple linkages for two reasons. Firstly, database software creates large temporary files of cross products, which is time consuming and may lead to memory problems. Secondly, queries involving complex linkages are often difficult to formulate and may easily turn out to be incorrect. The order in which the linkages are performed is also important for efficiency. In general, only the smallest possible files should be linked together. For example, rather than linking the patient and consultations files together, then finding the diseases and patient characteristics of interest, it is better to find the relevant subsets of the two files first, then link them together.

The software performs the required linkages and then analyses the data in two stages. The first part of the program performs the sequence of linkages and queries needed to find subsets required for the second stage, and the second part performs the analyses and displays the output. The data flow through the program is shown in Figure 3.

It can be seen from the diagram that any of the three input files may be linked to themselves or to either of the others in any combination to form subsets of the data, or the entire dataset can be used.

Finding Subsets

- The program enables the user to find any combination of characteristics required, simply by choosing the characteristic from menus. The program finds subsets of individual files, as well as linking files in the dataset to each other and to lookup tables, and finding subsets of one file according to data in another. For example the program can produce a list of young women with asthma who live in local authority accommodation, or of patients with a particular combination of diagnoses. It is also possible to examine the data for a particular group of people (for example, one ethnic group), or for a particular geographical area.

- Dealing with missing values.--When the data for MSGP4 was collected it was not possible to collect socio-economic data for all patients. The user is given the option to exclude missing values, or to restrict the data to missing values only should they want to find out more about those patients for whom certain information is missing. For example, an analysis of the frequency of cigarette smoking in each age/sex group in the practice might include only those patients for whom smoking information is available.

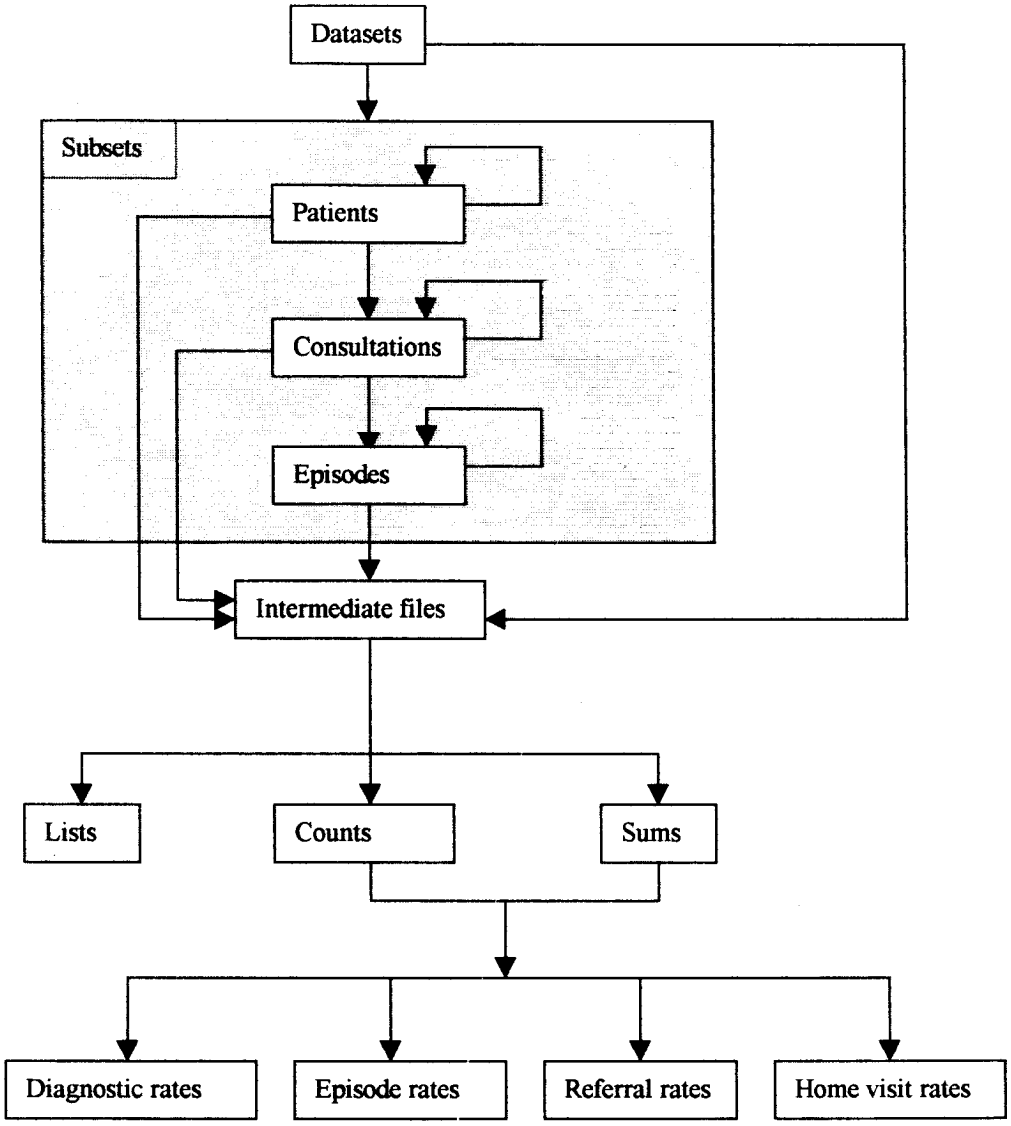
The Output

The output from the program is of three types, any of which may be exported by the program in a variety of formats (e.g., WK1, DBF, TXT, DB) for further statistical analyses.

- Lists output consists of one record for each patients, consultation or episode of interest, with files linked together as appropriate. Each record contains a patient number together with any other information that the user has requested. These flat files can be used for further analysis using spreadsheet or statistical software.

- Frequency output consists of counts of the numbers of patients, consultations or episodes in each of the categories defined by the fields selected by the user.
- Rate output enables a variety of rate with different types of numerators and denominators to be calculated. Any of the following rates may be chosen: Diagnostic rates for a specified diagnostic group (patients consulting; consultations; episodes); referral rates; and home visit rates. Rates are generally calculated for standard age and sex groups but other appropriate patient and consultations characteristics may be included in the analysis. Denominators can be consultations, patients consulting or patient years at risk.

Figure 3.—Data-flow Diagram for MSGPX Data Extractor Program


Datasets
Subsets
Patients
Consultations
Episodes
Intermediate files
Lists
Counts
Sums
Diagnostic rates
Episode rates
Referral rates
Home visit rates

Discussion and Conclusions

We have demonstrated through the use of one example database the potential that relational databases offer for storing statistical data. These are also the natural way to capture the data, since they reflect real data relationships, and are economical in storage requirements. They also facilitate linking in new data from other sources. However most statistical analyses require simple rectangular files, and complex database queries may be required to obtain these. We have shown that such complex linkages can be decomposed into a sequence of simple linkages, and user-friendly software can be developed to make such complex data readily available to users who may not understand the data structure or relational databases fully. The major advantage of such software is that the naïve user can be more confident in the results than if they were to extract the data themselves. They can also describe their problem in terms closer to natural language.

Although such programs enable the user with no knowledge of database technology to perform all the linkages shown above, they do have their limitations. Choosing options from several dialogue boxes is simple but certainly much slower than performing queries directly using SQL, Paradox or other database technology. Since the most efficient way to perform a complex query depends on the exact nature of the query, the program will not always perform queries in the most efficient order. The user is also restricted to the queries and tables defined by the program, and as more options are added the program must of necessity become more unwieldy and possibly less efficient.

User friendly software remains, however, the useful for the casual user who may not be familiar with the structures of a database, and essential for the user who does not have access to or knowledge of database technology.

References

Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, 64: 1183-1210.

McCormick, A.; Fleming, D.; and Charlton, J. (1995). *Morbidity Statistics from General Practice*, Fourth National Study 1991-92, Series MB5, no 3, London: HMSO.

Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press.

Newcombe, H. B.; Kennedy, J. M.; Axford, A. P.; and James, A. P. (1959). Automatic Linkage of Vital Records, *Science*, 130: 954-959.

Winkler, W. E. (1994). Advanced Methods of Record Linkage, *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 467-472.

Chapter 2

Tips and Techniques for Linking Multiple Data Systems: The Illinois Department of Human Services Consolidation Project

John Van Voorhis, David Koepke, and David Yu
University of Chicago

Abstract

This project involves the linkage of individuals across more than 20 state-run programs including TANF (AFDC), Medicaid, JOBS, Child Protection, Child Welfare Services, Alcohol and Substance Abuse programs, WIC, and mental health services. The count before linking is over 7.5 million records of individuals. Unduplicating the datasets leaves 5.9 million records. And the final linked dataset contains records for 4.1 million individuals. This study will provide the basic population counts for the State of Illinois's planning for the consolidation of these programs into a new Department of Human Services.

In the context of linking multiple systems, we have done a number of different things to make using AutoMatch easier. Some features of the process relate to standardized file and directory layouts, automatically generating match scripts, "data improvement" algorithms, and false match detection.

The first two issues, files and directories and scripts, are primarily technical, while the second two issues have more general substantive content in addition to the technical matter.

Properly laying out the tools for a matching project is a critical part of its success. Having a standard form for variable standardization, unduplication and matching provides a firm and stable foundation for linking many files together. Creating additional automation tools for working within such standards is also well worth the time it takes to make them.

With multiple sources of data it is possible to improve the data fields for individuals who are linked across multiple datasets. We will discuss both how we extract the information needed for such improvements and how we use it to improve the master list of individuals. One particular example of these improvements involves resolving the false linking of family members.

Chapter 3

Linking Administrative Records Over Time: Lessons from the Panels of Tax Returns

John L. Czajka, Mathematica Policy Research, Inc.

Abstract

In 1985 and again in 1987, the Statistics of Income (SOI) Division of the Internal Revenue Service initiated panel studies of taxpayers. Taxpayer identification numbers (TINs) reported on a sample of tax returns from the 1985 and 1987 filing years were used to identify panel members and search for their returns in subsequent years. The 1987 panel also included efforts to capture dependents, based on the TINs reported on Aparents@ and dependents' returns. This paper describes and assesses the strategy used to identify panel members and then capture and link their returns. While the availability of a unique identifier greatly simplifies data capture and record linkage and, as in this case, may determine whether or not a record linkage project is operationally feasible, imperfections in the identifiers generate a range of problems. Issues addressed in this paper include elements of operational performance, validation, and measuring the completeness of matching or data capture. Recommendations for improving the success of such efforts are presented, and implications for linkage across administrative records systems are discussed.

Introduction

How often, when confronted with a task requiring the linkage of records with imperfectly listed names and addresses, recorded in nonstandard formats, do we long for a unique identifier? This paper addresses some of the problems that analysts may face when they perform exact matches using a unique identifier. The paper deals, specifically, with records that have been linked by an exact match on social security number (SSN). The question it poses is, when is an exact match not an exact match? The paper is more about "unlinkage" than linkage per se. The linkages created by exact matches on SSNs represent the starting point. The work that ensues involves breaking some of these linkages as well as creating additional ones. The findings reported here may be relevant to any effort to link administrative records by SSN, whether longitudinally or cross-sectionally.

Overview of the Statistics of Income (SOI) Panel Studies

Over the years, the SOI Division of the Internal Revenue Service (IRS) has conducted a number of panel studies of individual (1040) tax returns. These studies employ a common methodology, for the most part. A base year panel sample is selected from the annual SOI cross-sectional sample, which provides a large and readily available sampling frame for such studies. Panel members are identified by their SSNs, as reported on their base year sample returns. The IRS searches for and captures all returns that list panel SSNs as filers in subsequent years. The returns captured by this procedure are then linked longitudinally. In reality, what are linked over time are persons, and these person linkages imply linkages between tax returns. In the two most recent panel studies, described below, the SSNs were edited, after this initial linkage, to correct errors and fill in missing values. After the editing was completed, the linkages were

re-established. As a result of this process, some of the original links were eliminated while others were added.

The 1985-based Sales of Capital Assets (SOCA) Panel began with about 13,000 base year returns. All filers on these returns were initially designated as panel members. Joint returns, which can be filed only by married couples, have two filers. Returns with other filing statuses have one filer. A SOCA Panel file covering the years 1985 to 1991 has been completed.

The 1987-based Family Panel began with about 90,000 base year returns. Not only filers but also their dependents (as claimed on base year returns) were defined as panel members. Returns filed by separately filing spouses, whether panel members or not, are to be captured and linked to the returns filed by their panel spouses. Returns filed by the dependents who are claimed in any year after the base year, whether they are original panel members or not, are to be captured and linked as well -- but only for the years in which they are claimed. Work to implement and review the SOI edits and prepare a panel file is only beginning; further editing will take place over the next few months.

Problems Created by Incorrect SSNs

Incorrect SSNs create a number of problems affecting not only record linkage and data capture but subsequent analysis of the data. In describing these problems, it is helpful to distinguish between incorrect SSNs on base year returns, which by definition include only panel returns, and incorrect SSNs on out-year returns, which include both panel and nonpanel returns.

Incorrect SSNs reported on base year returns have two types of consequences. Both stem from the fact that base year panel SSNs provide the means for identifying and capturing out-year panel returns. First, incorrect base year SSNs produce pseudo-attrition. Individuals whose SSNs were listed incorrectly in the base year will drop out of the panel when they file with correct SSNs. If these individuals are married to other panel members, they will remain in the database, but unless their base year SSNs are corrected their later data will not be associated with their earlier data. These missed linkages lead to incorrect weight assignments, which have a downward bias. A second consequence of incorrect base year SSNs is that the IRS will look for and may link the out-year returns of the wrong individuals to the base year records of panel members. The editing of SSNs is intended to eliminate both kinds of linkage errors.

Incorrect SSNs on out-year returns, as was stated, may involve both panel and nonpanel returns. If a panel member's SSN is misreported on an out-year return, after having been reported correctly in the base year, the out-year SSN will not be identified as panel, which may prevent the panel member's return from being captured at all. This is true if the panel member whose SSN is incorrect is the only panel member to appear on the return. While many panel returns continue to be selected for the annual cross-sectional sample in the years immediately following the base year, such that a panel return may still be captured despite the absence of a panel SSN, the incorrect SSN will prevent the panel member's being linked to the earlier returns. If a *nonpanel* return incorrectly includes a panel SSN, this error will result in, first, the return's being captured for the panel and, second, the wrong individual's data being linked to the panel member's base year record.

The bias that may be introduced by incorrect SSNs is distributed unevenly. Certain types of returns appear to be more prone to erroneous SSNs than others. Clearly, error rates are higher among lower income returns than among higher income returns. They may be higher as well among joint returns filed by couples who have a better than average chance of divorce in the next few years, although this observation is more speculative.

The dollar costs of incorrect SSNs cannot be overlooked either. In addition to the editing costs, there is

a cost to collecting and processing excess returns.

Identifying Incorrect SSNs

The SSN lacks a check digit. The SSN was established long before it became commonplace to include in identification numbers an extra digit or set of digits that can be used in an arithmetic operation to verify that the digits of the number "add up" right. As a result, there is no quick test to establish that a reported SSN was recorded incorrectly. Instead, it is necessary to make use of a number of other techniques to validate and correct the SSNs that are reported on tax returns or other administrative records.

Range checks are an important tool in screening out incorrect SSNs early in processing. Range checks of SSNs build on what is known and knowable about the distribution of numbers that have been issued by the Social Security Administration (SSA). A very limited range check can be based on the fact that the first three digits of the nine-digit number must fall into either of the ranges 001-626 or 700-728. SSNs with lead digits that fall outside these ranges must be incorrect. (The IRS uses an additional range to assign taxpayer identification numbers to persons who cannot obtain SSNs; these numbers are valid for IRS purposes but cannot be linked to other data.) More elaborate tests may utilize the fact that the 4th and 5th digits of the SSN have been assigned in a set sequence, historically. For each set of first three digits, SSA can report what 4th and 5th digits have been assigned to date or through a specific date. Most of the nine-digit numbers that have never been issued -- and, therefore, are incorrect -- can be identified in this manner. In addition, the SSNs that were assigned to persons who have since died can be obtained from SSA. Brief records for most SSA decedents can be accessed via the Internet.

The IRS maintains a validation file, using data obtained from SSA, to verify not only that particular numbers have ever been issued, but that they were issued to the persons who report them. The validation file contains up to 10 "name controls" for each SSN, where a name control consists of the first four characters of an individual's surname. If an individual changes his or her name numerous times and registers these changes with SSA, the different name controls will be present on the validation file, sorted from the latest to the earliest. The name control is a relic of period of much more limited computing capacity and less powerful software. The inability of name controls to differentiate among members of the same family, for example, restricts their utility for the editing of tax panel data, since misreporting among family members is a common type of error.

SSA maintains much more extensive data for its own validation purposes as well as other uses. Essentially all of the information collected on applications for new or replacement social security cards is retained electronically. The SSA will also perform validation exercises for other agencies. This was not an option for the IRS data, which could not be shared with SSA, but it may be a viable path for other users to take. In performing its validation and other matching exercises, SSA relies heavily on exact matches on multiple characteristics. SSA utilizes partial matches as well but without the framework of a probabilistic matching algorithm. As a result, SSA's validation tends to be conservative, erring on the side of making too few matches rather than making false matches.

In editing the SSNs reported on tax panel records, the IRS staff employed a number of evaluation strategies. These are discussed below.

The SOI Editing Strategy

The editing strategy employed by SOI staff for the two panel databases included several key elements. The first was the use of automated procedures to flag probable errors. The second was the reliance on manual or clerical review to evaluate the cases that were flagged as containing probable errors. Automated validation tests were not always definitive in identifying false matches, so expert review was

often necessary. Furthermore, there was no attempt to automate the identification of the appropriate corrections. The clerical review was responsible, then, for determining if an SSN was indeed incorrect, identifying the correct SSN or an appropriate substitute, and then implementing the needed corrections. The third element of the editing strategy was to correct the base year panel SSNs to the fullest extent possible. This is an important task because the corrected SSNs identify panel members in future years. The fourth element was to eliminate cross-sectional "violations" in the out-years -- that is, instances where particular SSNs appeared as filers multiple times in the same tax year, or where the SSNs listed as dependents matched to filers who were not the dependents being claimed. The last element of the editing strategy was to use automated procedures to apply SSN corrections to other years, where errors might exist but may not have been flagged. These corrections are directed at situations where a taxpayer continues to report an incorrect SSN for a filer, a separately filing spouse, or a dependent, year after year or at least for multiple years. These misreported SSNs may not always be flagged as probable errors. Furthermore, it is highly inefficient to rely on independent identification and correction of these errors.

Limitations of the Editing Strategy

The overall strategy has two notable limitations. First, the sheer number of cases that could be flagged as probable errors in a panel database containing nearly a million records, as the Family Panel file does, is very imposing. The obvious response is to limit clerical review to cases whose probabilities of error are judged to be very high. The SOI Division designed a number of validation tests. Certain tests were considered to be fatal; all violations had to be corrected. For other tests, multiple failures or specific combinations of failures were necessary in order to trigger a review. If a test is associated with a low probability of error, clearly it is inefficient to review all cases. But if there is no other test that in combination with this one can identify true errors with a high enough probability to warrant review, then errors will be missed. Below we discuss some of the problems associated with identifying incorrect secondary SSNs.

Another limitation is that cross-sectional error detection strategies have been favored over longitudinal strategies. This can be attributed to two things. First, some of the desired linkages are cross-sectional in nature, and cross-sectional tests have a direct impact on the quality of these matches. Second, it is difficult to define longitudinal tests that identify cases with high probabilities of error. The kinds of longitudinal conditions that suggest errors in SSNs involve breaks in continuity -- for example, changes in the SSN of a spouse or in some aspect of filing behavior. While incorrect SSNs will produce such breaks, most of the occurrences are attributable to genuine change.

Validating SSNs Against IRS/SSA Records

In editing the SOCA and Family Panel files, SOI staff used an IRS validation file that contained fields obtained, ultimately, from SSA. These fields were the SSN, up to 10 name controls, and the date of birth. Identifying variables that were present on the panel records included:

- SSNs (primary, secondary, and dependent);
- Return name control (derived from surname of first-listed filer);
- City and state;
- Full name line -- starting in 1988; and
- Name of separately filing spouse -- starting in 1988.

That the SOI Division did not begin to obtain full names until 1988 proved to be unfortunate for both panels. Having full names for the base year would have allowed panel members to be identified by both name and SSN. Some of the problems of validation that grew out of the limited identifying information that was present for the base year returns in both panels are discussed below.

Use of the Return Name Control

Until full names became available, the only identifying information about a filer was the return-level name control, which is derived from the surname of the primary filer, which may differ from that of the secondary filer and one or more dependents. Testing for exact agreement between the return name control and any of the name controls on the validation file for the primary SSN, the secondary SSN, and any dependent SSNs could be automated easily and reliably. Exact agreement was interpreted as validating the SSN. For primary SSNs, the application of this test dispensed with well over 99 percent of the sample cases. In a clerical review of cases failing this test in the base year of the SOCA Panel, more than half were judged to be true matches. The test failures occurred in these cases because of the misspelling of a name control on either file or because the order of the SSNs on the return did not correspond to the order of the names. That is, a couple may have filed as John Smith and Mary Wesson but listed Mary's SSN in the primary position. In this case the return name control of SMIT would not have matched the name control, WESS, associated with the primary SSN in the validation file. For secondary SSNs, the application of the return-level name control test dispensed with over 90 percent of the sample cases in the base year of the SOCA Panel. Still, the remainder were too many to review. Moreover, clerical review of the cases with name control mismatches could not be expected to resolve all of these cases. A secondary filer with a different surname than the primary filer would fail the test. Without a full name line, it was not possible to establish the secondary filer's surname or even that it differed from the primary filer's surname.

Use of Full Name Lines

Full name lines were not available to validate base year SSNs for either panel. From the standpoint of correctly establishing base year names, the one year lag for the Family Panel was not as bad as the three year lag for the SOCA Panel. Still, given that many erroneous SSNs are incorrect for only one year, the problem presented by changes in SSNs for secondary filers is a significant one.

The single most useful piece of information that a full name line provides is a surname for the secondary filer, from which a name control can be constructed. Basing validation tests for secondary SSNs on a secondary name control will yield substantially fewer false failures than tests that use the return level name control. With this improved targeting, clerical review of all violations becomes not only feasible but desirable.

Because the format of the name line is not exactly standard, there will be errors in constructing name controls for the secondary filer. Many of these errors, however, may occur in situations where the secondary filer has the same surname as the primary filer. For example, John and Mary Smith might list their names as John Smith and Mary. While an overly simple algorithm might yield MARY as the secondary name control, which would be incorrect and would produce a test failure, this need not undermine the validation procedures. Any strategy for using secondary name controls generated in this manner should include testing the secondary SSN against both the return name control and the secondary name control. In this example, the incorrect secondary name control would be irrelevant, as Mary Smith's SSN would be validated successfully against the return name control.

Strategies When Name Lines Were Not Available

For the SOCA Panel, name lines did not become available until year four. Birth dates provided important alternative information with which to evaluate the secondary SSNs. The birth date of the primary filer implies a probability distribution of secondary filer birth years. An improbable birth year for the secondary SSN may be grounds for determining that the SSN is incorrect when it also fails a name control test based on the return name control. Birth dates proved to be particularly helpful in choosing between two

alternative secondary SSNs when the reviewer had reason to believe that they referred to the same individual.

Name lines for later years may be valid substitutes for name lines in the base year when the SSNs in question do not change. But what if the secondary SSN does change? In particular, what if the base year secondary SSN failed a validation test based on the return name control and then changed the next year? Was this a true change in spouse or was it simply the correction of an SSN? Unless the two SSNs were so similar as to leave no doubt that one of the two SSNs was in error, the editors had to consider whether the change in SSN coincided with any pronounced change in circumstances, as reflected in the data reported on the two tax returns. Did the couple move, or did the earnings change markedly? These cases reduced to judgment calls on the part of the editors. In the SOCA Panel editing, such calls appear to have favored the determination that the filer changed, not just the SSN.

Multiple Occurrences within Filing Year

Incorrect panel SSNs may belong to other filers. If a panel member continues to use an incorrect SSN after the base year, and this SSN belongs to another filer, multiple occurrences of the SSN in question may be observed within a filing period. Such occurrences provide unambiguous evidence of the need for a correction. If the panel member does not continue to use the SSN, however, the false matches of out-year returns back to the incorrectly reported base year SSN become less easy to detect.

Findings

Table 1 summarizes our findings with respect to the frequency of erroneous SSNs in the population of tax returns filed for 1985, based on the editing of the base year data for the SOCA Panel. Of the SSNs that were determined to be incorrect, 42 percent belonged to other persons who filed during the next six years. Thus, 58 percent of the incorrect SSNs had to be identified without the compelling evidence provided by other filers using those SSNs correctly.

Table 1. -- Percentage of 1985 SSNs Determined to be Incorrect

Type of SSN	Percent incorrect
Primary SSN	0.57%
Secondary SSN	1.97

Source: SOI Division SOCA Panel.

Table 2 summarizes the findings for the 1987 filing year, based on the first year of the 1987 Family Panel. These findings include dependent SSNs, which taxpayers were required to report for the first time in that year. It is striking, first of all, how closely the estimated error rates for primary and secondary SSNs match those of the much smaller SOCA Panel. Second, the error rate for all dependent SSNs is just over twice the error rate for secondary SSNs. This is lower than pessimistic predictions would have suggested, but it could also be an understatement of the true error rate. Most dependents do not file tax returns, and so the evidence on which to base the error determinations may not be as solid as the evidence for primary and secondary filers. The other surprising feature is how the error rate for dependent SSNs takes off after the fourth listed dependent, rising to 24 percent for dependents listed in the 7th through 10th positions. It remains to be determined whether this high error rate is a phenomenon of higher order dependents or, more broadly, of all dependents on returns that report seven or more dependents. The number of sample cases involving more than five dependents is quite small, however, so the precision of these estimates for higher

order dependents is relatively low.

Table 2. -- Percentage of 1987 SSNs Determined to be Incorrect

Type of SSN	Percent Incorrect
Primary SSN	0.49%
Secondary SSN	1.65
All dependent SSNs	3.39
1st dependent SSN	3.36
2nd dependent SSN	3.04
3rd dependent SSN	3.63
4th dependent SSN	3.56
5th dependent SSN	7.78
6th dependent SSN	13.59
7th-10th dependent SSNs	24.31

Source: SOI Division Family Panel

Conclusions and Recommendations

The quality of SSNs reported on IRS records in 1985 and 1987 appears to be quite good. For primary SSNs the error rate is exceedingly low, which can be attributed in large part to the quality checks that primary SSNs must pass before the IRS will "post" their returns to its master file. Secondary SSNs have more than three times the error rate of primary SSNs, but the error rate is still low. Moreover, the IRS has increased its validation efforts with respect to secondary SSNs, so their quality should improve over time. Dependent SSNs had twice the error rate of secondary SSNs in 1987, but 1987 was the first year that dependent SSNs were required to be reported. These error rates are likely to decline as taxpayers become accustomed to the new requirements and as the cumulative effect of IRS validation efforts grows. In offering a preliminary assessment of the impact of SSN errors on data quality, I would say that, as of now, there is no evidence from the SOCA Panel that matches lost or incorrectly made due to bad SSNs will seriously compromise analytical uses of the data.

With respect to SOI editing procedures, I would make the following broad recommendations. First, the SOI Division needs to increase the amount of automation in the validation procedures and reduce the amount of unproductive clerical review time. Much of the clerical review time, currently, is spent on cases that are judged, ultimately, to be correct. The strategy that I discuss below for constructing and using secondary name controls will directly address this recommendation. In addition, the application of record linkage technology to the name control validation tests could significantly reduce the potential clerical review by allowing SSNs to pass validation when a name control contains a simple error. What I have in mind is modifying the tests so that they can take account of partial matches. Second, validation and editing must be carried out in a more timely manner. Data capture relies on an exact match to a list of panel SSNs. Unless corrected SSNs are added to the list as soon as possible, returns that could otherwise be captured will be lost.

Finally, I want to encourage the SOI Division to develop secondary name controls from the name lines

that became available in 1988 and use these name lines to edit the secondary SSNs in the Family Panel. Secondary name controls derived by even a simple algorithm from the full name line could substantially reduce the subset of cases that are flagged as possibly containing incorrect secondary SSNs. Reviewing all of the secondary SSNs that fail name control tests based on both the return name control and the secondary name control should then be feasible. Doing so will very likely prove to be an efficient way to identify virtually all cases with erroneous secondary SSNs.

Acknowledgments

I would like to thank the SOI Division for its support of this work. I would particularly like to acknowledge Michael Weber for his efforts in designing and overseeing the editing of both panel files, and Peter Sailer for encouraging attention to data quality. Finally, I would like to thank my colleague Larry Radbill for building the data files and generating the output on which the findings presented here are based.

Chapter 3

Record Linkage of Census and Routinely Collected Vital Events Data in the ONS Longitudinal Study

Lin Hattersley, Office for National Statistics, U.K.

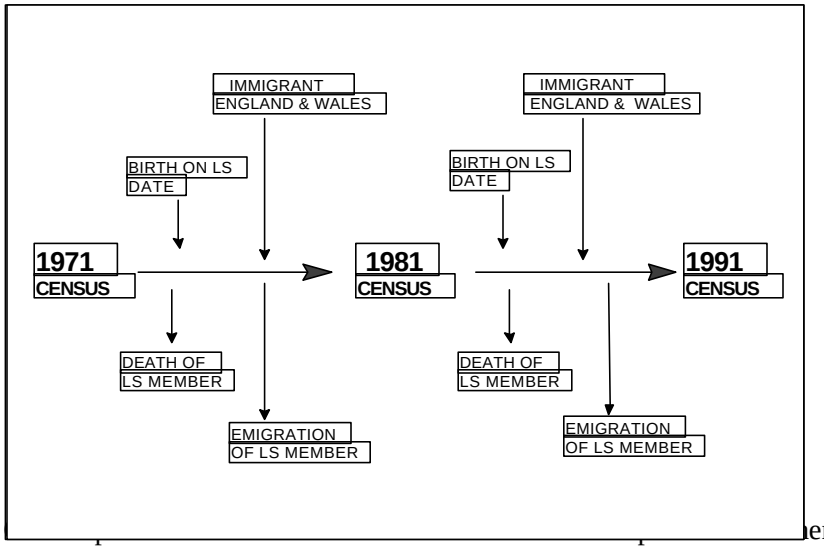
Abstract

Both manual and computerized methods of record linkage are used in the Office for National Statistics' Longitudinal Study (LS) -- a representative one percent sample of the population of England and Wales, containing census and vital events data. Legal restrictions mean that individual name and address data cannot be carried on either census or vital events computer files. Linkage of records has to be achieved by the use of the National Health Central Register (NHSCR) database, where names and addresses are carried together with information on date of birth and medical registration. Once an individual has been identified as a bona-fide LS member and flagged at the NHSCR, data carried on their census record or vital events record(s) can be extracted from the appropriate census file and vital event(s) file and added to the LS database. At no time are the two computer systems linked. This paper will describe the record linkage process and touch on some of the key confidentiality concerns.

What Is the ONS Longitudinal Study?

The ONS Longitudinal Study (LS) is a representative 1 percent sample of the population of England and Wales containing linked census and vital events data. The study was begun in 1974 with a sample drawn from the population enumerated at the 1971 Census using four possible dates of birth in any year as the sampling criterion. Subsequent samples have been drawn and linked from the 1981 and 1991 Censuses using the LS dates of birth. Population change is reflected by the addition of new sample members born on LS dates and the recording of exits via death or emigration. The structure of the population in the LS is shown below.

Figure 1. -- The Structure of the ONS Longitudinal Study

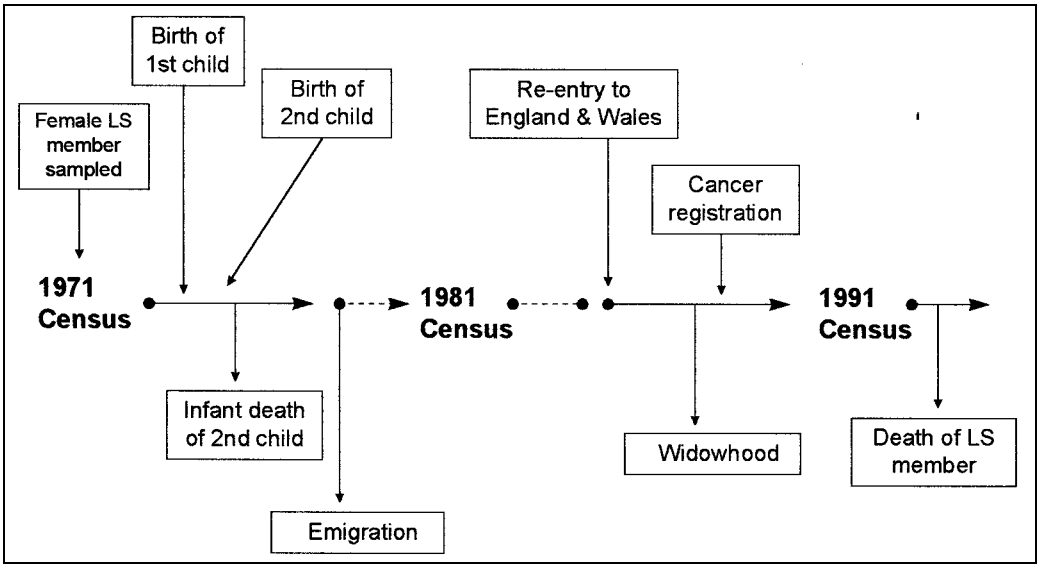


At each of the [illegible] er (non-LS) members

in the same household are included in the database. However, it should be noted that linkages of routinely collected events data are only performed for the LS members. The household an LS member resides in at one Census may well be different from the household they are part of in the next, and other (non-LS) household members may therefore change over time.

Routinely collected data on the mortality, fertility, cancer registrations, infant mortality of children born to LS sample mothers, widow(er)hoods and migration of LS members are linked into the sample using the National Health Service Central Register to perform the link (Figure 2). Marriages and divorces cannot be linked to the sample in Britain as the marriage certificate includes age, not date of birth.

Figure 2. -- Event Linkage --The Event History of an LS Member



Creation of the Sample and Methods of Linkage

Linkage methods vary depending on the source of data, but all linkages are made using the National Health Service Central Register (NHSCR). NHSCR performs the vital registration function for England and Wales and is part of the Office for National Statistics. The register was begun in 1939 using the data from the full census of the population carried out on the outbreak of the Second World War. Each enumerated individual was given an identification number which was used to allocate food rationing cards. This number became the National Health Service (NHS) number in 1948 when the NHS was created. Subsequently NHS numbers were issued at birth, or if the person was an immigrant, an NHS number was allocated when they first signed on with a General Practitioner (GP). The NHS number is thus the only identification number that is almost universally held among the population of England and Wales.

NHSCR was computerized in 1991 and prior to that date all records were kept in hand written registers containing one line per person in NHS number order. Events such as births, deaths, cancer registrations, enlistment into the armed forces, entries into long-stay psychiatric hospitals, re-entries to the NHS, embarkation's and internal migration were noted in the registers together with any ciphers denoting membership of medical research studies. In 1991 an electronic register was created.

The Creation of the Original LS Sample

When the LS was begun in 1974 an index card was created for each potential sample member who was

born on an LS date and enumerated in the 1971 Census. A unique 8 digit number was assigned to each LS member and printed on each card together with information that could identify the relevant census forms (such as ward, form number and enumeration district, sex, date of birth, marital status, person number and a usual residence indicator). The relevant census forms were then selected and from these name, usual address and enumeration address were written onto the cards. The cards were then sorted alphabetically and sent to NHSCR where they were matched against the registers. NHS numbers were added to the cards if the person was registered, and the register entries were flagged as LS. These cards were then used to create an LS alphabetical index held at NHSCR.

The essential element in the linkage of events to LS members is the possession of an NHS number and their presence as a member of the NHS register. Those LS members who do not possess NHS numbers are known as *"not traced"* and although Census data can be linked to them, vital event notifications, which are used by NHSCR in maintaining the registers, cannot. By the end of 1976 all but 3.2 percent of the 1971 sample LS members were traced in the register.

Different mechanisms are employed for census record linkage and event record linkage but both are covered by Acts of Parliament which restrict the use of certain data and at present prevent an electronic link being performed between the computer systems of NHSCR and the rest of ONS. Census data is covered by the Census Act which prevents the use of any data that can be used to identify an individual. As a result all completed census forms are stored for one hundred years before public release. Data from the schedules are held in electronic form but exclude names and addresses by law. However, dates of birth of all persons enumerated on each census form are included in the data. This inclusion of date of birth allows the identification of *potential* LS members at any census and together with the data which identifies each form uniquely, allows ONS to extract the forms and provide NHSCR with the names and addresses which can be used to match with their records.

The Linkage of Census Data to the LS

ONS have performed two LS-Census links to date, the first linking the 1971 and 1981 LS Census samples together, the second linking the 1981 and 1991 Census samples. Both LS-Census links were done in the same manner, although the computerisation at NHSCR in 1991 helped to speed the process of the second link.

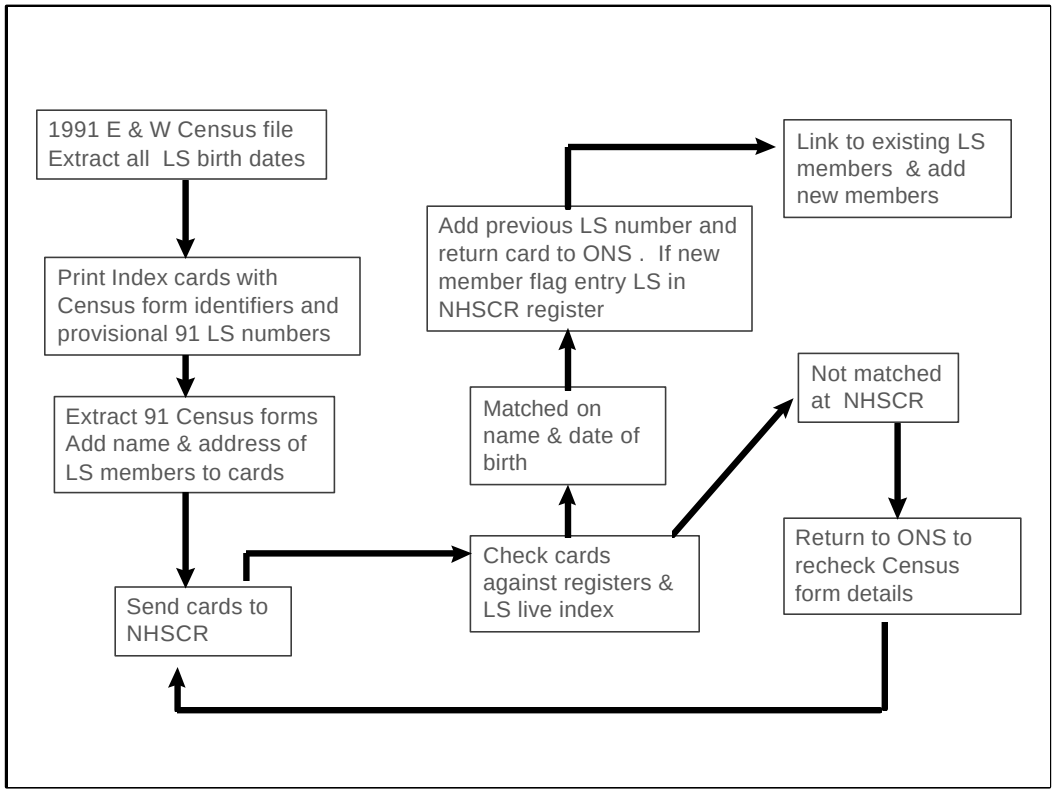
After the 1981 Census, index cards and listings of potential LS members were created from the Census GB Households file by extracting data for each household which contained any person with an LS date of birth. Each potential LS member was allocated a unique 1981 LS serial number which was printed on both the cards and the listings. As in 1971, when the LS was created, the information printed on the index cards was used to locate the Census forms and the name and address were transcribed from the forms. These cards were then sent to NHSCR for matching against the LS alphabetical index. If the LS member already existed in the index (that is had been enumerated in 1971 or had been born or immigrated after the 1971 Census) the 1971 LS number was added to the 1981 card which was then returned to OPCS (now ONS) for processing. If the cards were not matched with any entry in the LS index then a search of the NHS registers was made and if a match was found then the Central Register was flagged LS81 and that person entered the LS as a new member. Further searches against the electoral registers, birth indexes, marriage indexes and the Family Practitioner Committee's GP patient registers were also made for unmatched cards in 1981. The cards were also checked for "traced" or "not traced" status and were then returned to OPCS for processing as one of five types. These five types were:

- matched to an existing "traced" LS member;
- new "traced" 1981 entrant;

- new "not traced" 1981 entrant;
- matched to an existing "not traced" LS member (these could be "traced" or "not traced" in 1981); or
- matched to a 1971 LS member but a double enumeration.

The LS numbers on the returned cards were validated and the resulting file was run against the 1981 LS Households file in order to add the 1971 LS number or intercensal entry LS number to the records. The final process in the link was the creation of separate LS personal and household files for 1981. After this was completed the cards were returned to NHSCR for addition to the LS index.

Figure 3. -- The Linkage Process



The 1981-1991 LS - Census link was completed in 1995 (Figure 3). Although NHSCR had been computerized in 1991 a manual linkage process was used to fulfill the confidentiality rules. As in 1981 index cards and listings were produced giving census schedule identifiers and 1991 LS serial numbers. The Census forms were extracted and the names and addresses were added for NHSCR identification purposes. Once the cards had been completed and checked they were sent to NHSCR for matching and tracing against the registers and LS indexes. The matching and tracing process was easier and faster than in 1981 as the cards were initially matched and traced against the NHSCR database entries rather than manually against the two rooms full of index cards which formed the LS alphabetical index. Only if no previous LS number existed or there was no entry on the NHSCR database were the cards checked against the clerical registers and indexes. Any cards not matched were returned to OPCS for re-checking against the census forms to identify transcription errors. The NHS number and any pre-1991 LS numbers were added to the cards before their return to OPCS for processing.

How Good Was the Linkage Between Censuses?

The two LS-Census links so far performed have been extremely successful, with at least 90 percent of traced LS member's records being linked together. It should be noted that 97 percent of 1971 LS members, 99 percent of 1981 LS members and 98 percent of 1991 LS members were traced at NHSCR at the time of linkage (Table 1).

Table 1. -- Forward Linkage Rates for the 1971-1981 LS-Census Link and the 1981-1991 LS-Census Link

Forward Linkage Rates					
	1971 Census Sample *	71-81 Linked Sample	1981 Census Sample **	81-91 Linked Sample	1991 Census Sample ***
	N = 512,881		N = 530,248		N = 534,647
Died prior to next census	58,911		58,931		
Embarked prior to next census	5,625		4,399		
Eligible to be in next census	448,345		466,918		
Recorded in next census		408,451		420,267	
Forward linkage rate		91%		90%	

*Traced at NHSCR prior to the 1981 Census (97%).
**Traced at NHSCR prior to the 1991 Census (99%).
***Traced at NHSCR at the 1991 Census (98%).

However, even allowing that LS-Census forward linkage rates were extremely good there were still approximately 10 percent linkage failures at each census. This problem of linkage failure was investigated using the NHSCR records to examine 1 percent samples of linkage failures as part of each of the LS-Census Link exercises.

Table 2. -- Reasons for Failure to Link

	Number Believed to Still be in Sample But Not Found at Census

Reasons for Failure to Link	1971-81 Link N = 39,616	1981-91 Link N = 46,652	All LS Members Who Failed to Link by the 1991 Census* N = 92,580
Date of birth discrepancy between Census & NHSCR	37%	21%	18%
Cancelled NHS registration -- whereabouts not known	6%	9%	16%
Missed event (emigration, death, enlistment)	14%	5%	4%
Not known	10%	5%	18%
Currently registered at NHSCR but not enumerated	38%	61%	44%

*Includes LS members lost to link in 1981 and still not linked in 1991 and LS members linked in 1981 but not in 1991. Excludes LS members who were lost to link in 1981 but were linked in 1991.

The total number of LS members lost to link between 1971 and 1991 was 92,580 (Table 2). Date of birth discrepancies were a major cause of failure to link providing at least 37 percent of failures in 1971. Those in the "Not known" category may well also have included sample members who had given dates of birth other than LS dates on their Census forms. The rise noted in "Cancelled NHS registrations," which tend to occur if a person has not been seen by their GP for over two years, suggests that many of the persons in this category may have in fact emigrated but not reported it.

Vital Events Linkage

While the LS-Census links only take place once every ten years, vital events linkage occurs annually for most events and six monthly for some. There are two methods of identifying vital events occurring to LS members – firstly, through routine notification of events to NHSCR, where the LS member is identified by the presence of an LS flag in the register; and secondly, through the annual vital events statistics files compiled by ONS. Some types of event, deaths and cancer registrations are identified using both methods as a cross checking device (Table 3).

Table 3. -- Vital Events and the Methods of Linkage

Event Type Currently Collected	Linked Through Routine Notification	Linked Through Stated Date of Birth
New births into sample		**X**

Births (live & still) to LS mothers		**X**
Infant deaths of LS mothers children		**X**
Widowerhoods		**X**
Deaths of LS members	**X**	**X**
Cancer registrations	**X**	**X**
Immigrants into sample	**X**	
Emigrations	**X**	
Enlistment into armed forces	**X**	
Re-entries from emigration and enlistment	**X**	

How the Linkage Process Works

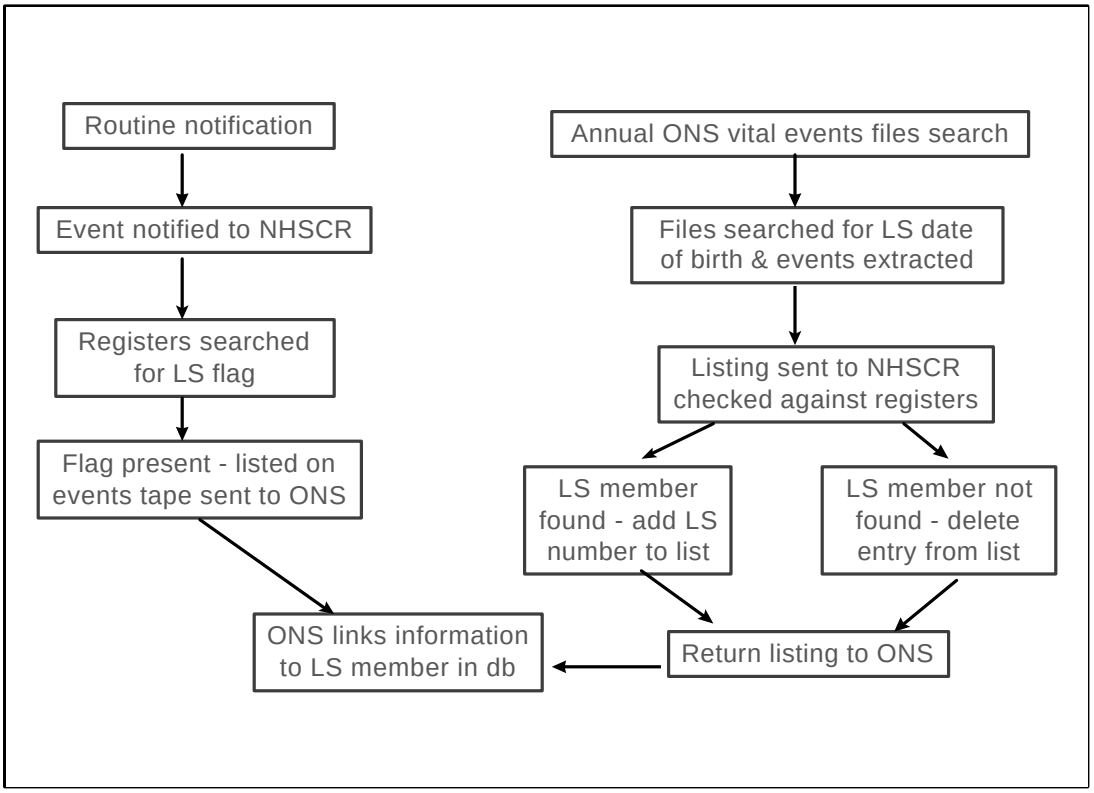
The identification of immigrants into the LS sample, emigrations out of England and Wales by sample members, enlistments into the armed forces and re-entries from emigration or enlistment are all made through the routine notification of these events to NHSCR (Figure 4). When NHSCR updates their database the LS flag is noted for all existing sample members and the details including the LS number are entered onto a tape that is sent twice yearly to ONS to update the LS database. Included in the tapes are details of date of emigration, enlistment or re-entry together with the relevant LS numbers, and for immigrants date of birth and entry details. ONS returns a listing to NHSCR containing the new LS numbers allocated to immigrants joining the sample and this is used by NHSCR to flag their database.

New births into the sample, births to sample mothers, infant deaths of LS members children and widow(er)hoods are all identified using date of birth searches of the annual vital events statistics files. The process involves extracting a subset of data from the statistics files using the LS birth dates as the selection criteria. In the case of new births an LS number is allocated and a listing is sent to NHSCR containing LS number, Date of birth and NHS number. The entry is checked and the LS number is added to the register entry and the new LS member is flagged.

Births to LS mothers are also extracted from the annual England and Wales births file, but the criterion used here is the date of birth of the mother which must be an LS date. A listing including registration details is sent to NHSCR where it is used to extract the relevant birth drafts to identify the name of the mother. The mother's name is then used to find the LS number which is added to the listing which is then returned to ONS for processing.

Infant death details are extracted from the annual deaths file and the mothers date of birth is then matched with the data on the LS births to sample mothers file. Any queries are sent to NHSCR for resolution using the registers. Widow(er)hoods are also linked using the annual deaths file. An LS date of birth search for the surviving spouse is used to extract the data and a listing giving the date of death and registration details is sent to NHSCR. NHSCR have access to the ONS deaths system and use this to identify the names of the deceased and their surviving spouse. The register is then searched for the surviving spouse’s name and the LS number extracted and added to the listing.

Figure 4. -- The Linkage Process



How Good Is the Linkage of Events?

The quality of event linkage is extremely good for new births into the sample and deaths occurring to sample members. Virtually 100 percent of these events are linked (Table 4). The rate of linkage for other events is high, with the exception of migration events. Unlike events directly associated with births and deaths which have to be registered within set times by law, migration events do not have to be compulsorily registered. Immigrants can only be linked to the sample when they register with a GP and this may be long after the date of immigration. The date of birth for immigrants is that taken from their NHS registration details and may not be accurate. Certainly, between 1971 and 1981, 62 percent more immigrants were linked to the LS than were expected based on the England and Wales immigration figures. Emigrations of LS members out of England and Wales are only captured if an LS member returns their medical card to their Family Health Service Authority on leaving the country or if the Department of Social Security informs NHSCR when a pensioner or a mother with children is no longer resident. As a result not only are emigrations undercounted but they are often notified to NHSCR many years after the event.

Table 4. -- How Good is the Linkage of Events?

Event	Percentage Linked Between 1971 And 1981 Census	Percentage Linked Between 1981 And 1991 Census
New births into sample	101%	100%

Immigrants into sample	162%	106%
Deaths of sample members	98%	109%
Emigrations of sample members	65%	36%
Births to sample mothers	92%	93%
Widow(er)hoods	77%	84%
Cancer registrations	98%	103%**
Infant mortality	86%*	91%

* Available from 1976
** Available until 1989

Confidentiality Issues

There are two sets of confidentiality issues involved with the maintenance and usage of LS data. First, how to link data without breaching the legal restrictions on the release of census and certain vital statistics data, and second how to ensure that confidentiality is maintained by researchers using the data for analysis.

The processes of data linkage would be accelerated if electronic linkage could be achieved between ONS and NHSCR. However, at present this would contravene all legal requirements including that of current UK data protection legislation. The LS is not a survey where an individual gives their consent for the use of personal data but a study where administrative data collected for other purposes is used to provide a rich source of socio-demographic and mortality data about the England and Wales population over time. Given the restrictions imposed by this situation, the maintenance of the study must not only be done in such a manner as to comply with the legal instruments but must also be publicly seen to do so.

The restrictions on the methods used for linkage of the data also apply to the release of data for analysis by outside researchers. Any data which could conceivably identify an individual such as the LS dates of birth and LS number are used only within the database to achieve linkage between data files. Extraction of data is done within ONS itself and data is only released to researchers in aggregated form which will not permit the identification of an individual.

Conclusion

The LS is a complex linkage study which, by using the only universal identifier held by members of the population of England and Wales (the NHS number), has provided extremely high quality linked data on a 1 percent sample of that population for over 20 years.

The linkage methods used are partially computerised but because of legal restrictions much of the linkage is still labour intensive and reliant on the skills of ONS and NHSCR staff. Automatic linkage would be the ideal, but until it is legally feasible to electronically link the LS system to all other ONS systems (including the Census database) and to NHSCR, this is unlikely to be achieved.

Chapter 3

Use of Probabilistic Linkage for an Analysis of the Effectiveness of Safety Belts and Helmets

Dennis Utter, National Highway Traffic Safety Administration

Abstract

This presentation will describe the use of linked data by the National Highway Traffic Safety Administration to generate population-based crash and injury state data that include the medical and financial outcome for specific crash, vehicle, and behavior characteristics. The linked data were used by NHTSA for a Report to Congress as mandated by the Intermodal Surface Transportation Efficiency Act (ISTEA) of 1991. Benefits were to be measured in terms of about their impact on mortality, morbidity, severity, and costs.

Hawaii, Maine, Missouri, New York, Pennsylvania, Utah, and Wisconsin, states with the most complete statewide crash and injury data, were funded by NHTSA to implement Crash Outcome Data Evaluation Systems (CODES). The states linked crash to hospital and EMS and/or emergency department data using their most recent data year available at the time, ranging from 1990-1992. Implementation of a uniform research model by the seven CODES states was successful because of the linked data. The presentation will discuss how the linked data were used to standardize non-uniform data and expand existing data for analysis.

Introduction

Motor vehicle traffic crashes continue to be a significant problem in the United States. Each year there are more than 6 million crashes investigated by police agencies. In these crashes 3.5 million people are injured, 450,000 of them severely, and nearly 42,000 are killed. Crashes produce a staggering economic toll, too. Nearly $151 billion are lost due to medical costs, property damage, legal costs, productivity losses, and other factors. Clearly, reducing the number of crashes and their severity is a necessity.

The National Highway Traffic Safety Administration (NHTSA) was created to reduce the number of deaths, the severity of injuries, and other damage resulting from motor vehicle traffic crashes. It does so through a variety of programs aimed at making vehicles safer, therefore mitigating the results of crashes, and by getting vehicle drivers and occupants to do things that would either prevent crashes or mitigate their outcomes. Evaluation of these programs requires a significant amount of data. Data linkage provides NHTSA, and the traffic safety community at large, with a source of population-based crash and injury state data that include the medical and financial outcome for specific crash, vehicle, and behavior characteristics.

Data files created from police reported motor vehicle crash data alone do not include medical outcome information for everyone involved in a motor vehicle crash. Thus, linking data became necessary when NHTSA was required by the Intermodal Surface Transportation Efficiency Act (ISTEA) of 1991 to report to the Congress about the benefits of safety belts and motorcycle helmets. Benefits were to be defined in terms of mortality, morbidity, severity, and costs. Statewide crash data files were determined by NHTSA to be the only source of population-based information about the successes (those who use the countermeasure and receive no or a less serious injury), the failures (those who do use the countermeasure and receive an injury), those not affected (those who do not use the countermeasure and receive no injury) and those who were not injured as seriously as they might have been because of the safety device.

CODES

Hawaii, Maine, Missouri, New York, Pennsylvania, Utah, and Wisconsin, states with the most complete statewide crash and injury data, were funded by NHTSA to implement Crash Outcome Data Evaluation Systems (CODES). The states linked crash to hospital and EMS and/or emergency department data using their most recent data year available at the time, ranging from 1990-1992. The study population was chosen from police reported data because of the importance of the safety belt and helmet utilization data contained in the crash file. The safety belt analysis included 879,670 drivers of passenger cars, light trucks and vans and the motorcycle analysis consisted of 10,353 riders of motorcycles. This presentation will describe how linked data made it possible for NHTSA to conduct a medical and financial outcome study of the benefits of safety belt and motorcycle helmets using routinely collected, population-based, person-specific state data.

Use of Linked Data to Standardized Non-Uniform Data for Analysis

Outcome Analysis Using "As Reported" Data

Measuring outcome is complicated when using "as reported" utilization data. Using this type of data, the CODES results indicated that although each state was different, all safety-belt odds ratios from all states agreed that safety belts are highly effective at all analysis levels at less than the .001 significance level. The non-adjusted effectiveness rates indicated that safety belts were 89% effective for preventing mortality and 52% effective for preventing any injury. The downward shift in severity was demonstrated by the decreasing effectiveness rates ranging from 89% for victims who die to 75% for those who die or are inpatients and to 54% for those who die, are inpatients, or are transported by EMS. But these results are inflated. When safety belt usage is mandated, human beings being human have a tendency to exaggerate their use of a safety belt, particularly when crash evidence or their injury type and severity are not likely to indicate otherwise. Over reporting of belt use moves large numbers of unbelted uninjured persons into the belted uninjured column thus inflating belt effectiveness. NHTSA repeated the research model to incorporate observed safety belt utilization rates into the analysis. Adjustments were made based on the assumption that 35 percent of the belted who were uninjured or slightly injured may have misreported their belt use at the time of the crash. These adjustments obtained the more realistic effectiveness rates of 60 percent for preventing mortality and 20-45 percent for preventing morbidity. In the future, as state injury data systems are improved to include safety utilization and external cause of injury information, linkage will make it possible to use the injury data to confirm utilization of the safety device.

Definition of the Occurrence of an Injury and Injury Severity

Although the study population was defined from the crash report, the linked data were used to define the occurrence of an injury and the various levels of injury severity. This standardization was necessary to compensate for inconsistent implementation of the police reported KABCO severity scale by the different states. For example New York classified one-third of the occupants involved in crashes as suffering "possible" injuries compared to about 10 percent in the other CODES states. For CODES, injury and the severity levels were defined by combining "injury severity" on the crash report with "treatment given" on the injury records to create five levels -- died, inpatient, transported by EMS or treated in the ED, slightly injured or no injury. Police reported "possible" injuries were classified as non-injured unless the crash report linked to an injury or claims record. The severity levels were used to define the outcome measures (mortality, morbidity, injury severity, and cost) for the uniform research models for both the belt and helmet analyses as follows:

Mortality:	Died versus all other crash-involved victims.
Morbidity:	Any injured compared to those not injured.
Shift in Severity:	Separate effectiveness rates for each severity level were calculated and then compared to measure the downward shift in injury severity
Cost:	Defined as inpatient charges because non-inpatient charges were not comparable among the seven states.

Use of Linked Data to Expand Existing Data

Identifying Injuries Not Documented by the Police

Police are required to document only those crashes and injuries that occur on public roads and meet mandated reporting thresholds. In addition, some reportable injuries are not documented because of non-compliance with the requirements. CODES excluded cases not documented by the police because of the need for standardized safety device utilization information. But using only crash reports to document the injuries understates the total injuries. The CODES states used the linked crash and injury records to identify those injuries not documented by the police.

Identifying Financial Outcome

Data linkage provides highway safety with access to financial outcome information related to specific characteristics of the crash event. Lack of uniformity in the documentation of EMS and emergency department charges limited the CODES analysis to inpatient billed charges as indicated in the hospital data. These data were used to calculate average charges for inpatient drivers and all crash involved drivers. The analysis indicated that the average inpatient charge for unbelted drivers admitted to a hospital was 55% higher than for the belted, $13,937 compared to $9,004. If all drivers involved in police-reported crashes in the CODES states had been wearing a safety belt, costs would be reduced 41 percent (approximately $68 million in reduced inpatient charges or $47 million in actual costs). This type of information is powerful in the political arena and is unavailable to highway safety except through data linkage.

Identifying the Type of Injury

Linked data were crucial for the helmet analysis. By using only the level of severity NHTSA found that the effectiveness rates were low for helmets, 35% effective in preventing mortality, and only 9% effective in preventing morbidity. The downward shift in injury severity was much less than for safety belts. The linked data enabled NHTSA to redirect the analysis to brain injuries which the helmet is designed to prevent and found that helmets were 67 percent effective in preventing brain injury. That means 67% of the unhelmeted brain injured would not have been so injured if they had been helmeted. Looking at the costs for the brain injuries also justified focusing the analysis. Average inpatient charges for the brain injured were twice as high. Approximately $15,000 in inpatient charges would be saved during the first 12 months for every motorcycle rider who, by wearing the helmet, did not sustain a brain injury. Again, this type of information is more powerful than the overall effectiveness rate for helmets.

Barriers to Linkage of Crash and Injury Data

Probabilistic linkage requires computerized data. Unfortunately, not all states have crash and injury data that are statewide and computerized. Almost all of the states have computerized crash data statewide. Half of the states have developed state EMS data systems, but only a few have state emergency department data systems. A majority of the states have computerized state hospital discharge data systems. All of the states have computerized Medicaid and Medicare data systems, but few states have statewide computerized data files for private vehicle or health insurance claims data. Access to data for the less seriously injured victims, a group that includes many of the successes for highway safety, is difficult to obtain because the data may not be computerized. Or if computerized, they are computerized by provider or by insurance group and rarely statewide. Injury data are particularly useful to highway safety because they document what happens to all victims injured in motor vehicle crashes, regardless of whether the crash itself meets police reporting thresholds.

Benefits of Data Linkage

Data linkage provides documentation, generated from a state's own linked data, that is more credible among local decision makers who may be tempted to repeal the safety mandates, such as helmet legislation. And the data linkage process itself has the added benefit of making data owners and users more aware of the quality, or lack thereof, of the data being linked. The CODES states found that important identifiers that should have been computerized uniformly statewide were not; or if the identifiers were computerized, some of the attribute values were missing or inaccurate. All of the states became adept in discovering errors in the data and were motivated to revise their edits and logic checks. Thus annual linkage of the crash and injury state data provides the states, NHTSA, public health and injury control, with a permanent and routine source of outcome information about the consequences of motor vehicle crashes at the same time that the quality of state data are improved for their originally intended purposes.

Chapter 3

Multiple Causes of Death for the National Health Interview Survey

John Horm, National Center for Health Statistics

Abstract

The National Health Interview Survey (NHIS) is a nationally representative health survey of the United States population. The NHIS is a rich resource for national and subnational health information such as chronic and acute conditions, doctor visits, hospital stays and a wide variety of special health topics knowledge, attitudes, and behaviors each year. Basic socio-demographic information is routinely collected on each person in the NHIS. The NDI contains records for virtually 100 percent of persons who die in the United States. Respondents to the NHIS who are age 18 or over are now routinely linked with the National Death Index (NDI) to create a new resource of immense public health and epidemiologic potential. An automated probabilistic approach has been used to link the two data files from the date of interview through 1995 and classify the linked records as either true (deceased) or false (alive) matches. It is estimated that over 97 percent of deceased persons and 99 percent of living persons are correctly classified as to vital status. The linked NHIS-NDI files contain all of the survey information along with vital status, multiple causes of death and date of death if deceased.

Introduction

The National Health Interview Survey (NHIS) is a large in-person health survey of the United States population conducted annually by the National Center for Health Statistics (Dawson and Adams, 1987). Health and health-related information is collected on approximately 122,000 persons per year (42,000 households) among the civilian, non-institutionalized population (note that since matching with the NDI is done only for persons aged 18 and over, the sample size for this purpose is about 85,000 persons). The NHIS consists of a basic health and demographic questionnaire (BHD) with information on every person in the household. The BHD contains basic socio-demographic information, acute and chronic conditions, doctor visits, hospital stays, and related items. In addition to the BHD, one or more surveys on Current Health Topics (CHT) is also conducted each year. The CHT surveys are usually administered to one randomly selected sample person over the age of 18 in each family although there are some family-style CHT surveys. The sample-person CHT surveys yield information on about 42,000 persons per year. Recent CHT surveys include the following content areas: alcohol use; cancer epidemiology and control; child health; health insurance; adult immunization; Year 1990 health objectives; Year 2000 health objectives and others. All questionnaires and topic areas included from 1985 through 1989 have been published by Chyba and Washington (1993). Response rates for both components of the NHIS are high: 95 percent for the BHD and about 85 percent for the CHT's.

The NDI is a central computerized index with a standard set of identifying information on virtually every decedent in the United States since 1979 (Boyle and Decoufle, 1990) managed by the National Center for Health Statistics and can be used to enumerate and identify decedents in epidemiologic studies. The NDI produces matches between user records and death records based on a set of twelve criteria. The user must then develop a methodology to classify the potential matches returned by the NDI as either true or false matches.

The approach taken here to classify the NHIS-NDI potential matches is a modification of the probabilistic approaches developed by Fellegi and Sunter (1969) and refined by Rogot, Sorlie, and Johnson (1986).

Methods

The NDI contains records on all deaths occurring in the United States since 1979 and is fully documented in the National Death Index User's Manual (1990). The NDI has developed a set of 12 criteria under which matches between user records and NDI records are produced. These criteria are based on various combinations of Social Security Number, date of birth, first name, middle initial, and last name. The 12 matching criteria are:

- Social security number and first name;
- Social security number and last name;
- Social security number and father's surname;
- If female, Social security number, last name (user's record) and father's surname (NDI record);
- Month and year of birth and first and last name;
- Month and year of birth and father's surname;
- If female, month and year of birth, first name, last name (user's record) and father's surname (NDI record);
- Month and year of birth, first and middle initials, and last name;
- Month and $\pm$1 year of birth, first and middle initials, and last name;
- Month and $\pm$1 year of birth, first and last names;
- Month and day of birth, first and last names; and
- Month and day of birth, first and middle initials, and last name.

An NDI record is matched to a user record if any one of the above 12 criteria result in a match.

An indication of agreement between the user record and the NDI record is returned to the user for each of the seven items involved in the twelve matching criteria. In addition to the items involved in the matching criteria the NDI returns an indication of agreement/disagreement between the user record and the NDI record on five additional items: age at death; race; marital status; state of residence; and state of birth. Multiple NDI records may be matched to a single user record and a possibly large number of false positive matches may be returned by the NDI. Matches between NDI records and NHIS records are referred to as ***potential matches***.

The NHIS routinely collects all of the seven data items used by the NDI for matching as well as the five additional items used for assessing the quality of potential matches. The NHIS has essentially 100 percent complete reporting of these items except for social security number (SSN) and middle initial. Completeness of reporting of SSN and middle initial varies by year but is generally between 65 and 75 percent. Various studies have indicated that the NDI is capable of identifying over 90 percent of known deaths (Patterson and Bilgrad, 1986; Stampfer et al., 1984; Williams, Demitrack and Fries, 1992) with some studies finding that the proportion is in the upper 90's when a full set of identifiers is available (Calle and Terrell, 1993; Curb et al., 1985; Horm and Wright, 1993). Social Security Number is a key identifier in the matching process. When the SSN is not available the proportion of known deaths identified drops to about 90 percent.

Tepping (1968) developed a model for computerized matching of records from the perspective of the cost of making correct or incorrect decisions about potential matches. Fellegi and Sunter (1969) developed a theory-based approach for record linkage which incorporated the concept of weighting factors with the weight being positive if the factor agreed and negative if it disagreed With the magnitude of the weight being inversely proportional to the frequency of the factor in the population. This approach was refined by Rogot, Sorlie, and Johnson (1986) who used binit weights [$\text{Log}_2 (1/p_i)$] where p_i is the proportion of the population with the i^{th} characteristic. Newcombe, Fair, and Lalonde (1992) while not espousing a particular form for the weights did make a case for the necessity of weighting by something more than simple agreement/disagreement weights.

Weights

Weights for each of the eleven items used for assessing the quality of the potential matches were constructed based on the composition of the 1988-91 NHIS and 1986-91 U. S. deaths (SSN is handled separately).

A weight is the base 2 logarithm of the inverse of the probability of occurrence of the characteristic based on the above files. For example, since males constitute about 46.3 percent of the population aged 18 and over, the weight is $\log_2(1/.463) = 1.11$. Weights are constructed in a similar manner for race, last name, father's surname, birth month, day, and year, state of residence, and state of birth. Since middle initials are sex-specific, sex-specific weights were constructed for middle initial. Weights for marital status were constructed to be jointly age and sex specific. First name weights are both sex and birth year cohort (<1926, 1926-1935, 1936-1955, and >1955) specific because of secular trends in the assignment of first names.

Weights may be either positive or negative. If a particular item matches between the NHIS record and the NDI record, the weight is positive. If the item does not match, the weight is negative. Weights for items missing from the NHIS file, the NDI file, or both are assigned a weight of zero.

Last name weights have been modified for females. Since some females change their surnames upon marriage, divorce, remarriage, etc., matching on surname only may produce false non-matches. The NDI returns an indication of a match on the father's surname as well as last name which is used as auxiliary information for females. If last name does not match on the two records (the last name weight is negative), the last name weight is replaced with the father's surname weight if positive, otherwise the last name weight is retained. This approach provided the best classification performance for females.

Because all information provided to the NDI is proxy reported and information provided to the NHIS may be proxy reported, there is a considerably likelihood that one of the two files may contain a respondent's given first name while the other contains his/her commonly used nickname. We have constructed files of common nicknames which are used in the classification process if the first name on file does not provide a good match.

Frequency-based weighting schemes such as proposed by Fellegi and Sunter and Rogot, Sorlie, and Johnson are attractive since the rarer occurrences of a matching item is given more weight than more common occurrences. However, the user is still left with the problem of properly classifying matched records into at least minimal categories of true matches, false matches, and questionable matches. Recent work by Belin (1993) and Belin and Rubin (1993) suggests that the false-match rate is sensitive to the setting of cut-points.

Calibration Samples

Calibration samples need to have known vital status information such as date and location of death, and ideally, death certificate number on the sample subjects based on sources independent of the NDI. Two NCHS surveys meet this criteria.

The 14,407 persons who participated in the NHANES I examination survey (1971-75) were used as the first calibration sample. Active followup was conducted on this sample to ascertain the vital status of the participants and death certificates obtained for persons found to be deceased (Finucane et al., 1990). NHANES is a large nationally representative survey and is sufficiently similar to the NHIS to be used as a calibration sample for developing a methodology for classification of the NHIS-NDI matches.

The NHANES I followup sample was then matched to the NDI and randomly stratified into two samples, a developmental sample and a confirmation sample.

Any one calibration sample may have an inherent structural process which differs systematically from the

target sample. Even though the NHANES sample was randomly stratified into two samples, systematic differences between NHANES and the NHIS could exist in both parts. Thus a second calibration sample was used to counteract potential structural differences. The second calibration sample used was the Longitudinal Study on Aging (LSOA) (Kovar, Fitti, and Chyba, 1992), a subset of the 1984 NHIS. The data used from this sample were those participants aged 70 and over at the time of interview and followed through August, 1988. Vital status was obtained independent of the NDI by interviewer followback in both 1986 and 1988.

Classification of Potential Matches

Potential matches returned by the NDI must be classified into either true or false matches. This is done by assigning a score, the sum of the weights, to each match.

$$\begin{aligned}\text{Score} = {} & W_{\text{firstname X sex X birthcohort}} + W_{\text{middleinitial X sex}} + W_{\text{lastname}} \\ & + W_{\text{race}} + W_{\text{maritalstatus X sex X age}} + W_{\text{birthday}} \\ & + W_{\text{birthmonth}} + W_{\text{birthyear}} + W_{\text{stateofbirth}} + W_{\text{stateofresidence}} .\end{aligned}$$

The NHANES I developmental sample suggested that classification efficiency could be increased by grouping the potential matches into one of five mutually exclusive classes based on which items matched and the number of items matching. These classes are:

Class 1: Exact match on SSN, first, middle, and last names, sex, state of birth, birth month and birth year.

Class 2: Exact match on SSN but some of the other items from Class 1 do not match although certain cases were moved from Class 2 to Class 5 because of indications that the reported SSN belonged to the spouse.

Class 3: SSN unknown but eight or more of first name, middle initial, last name, birth day, birth month, birth year, sex, race, marital status, or state of birth match.

Class 4: Same as Class 3 but less than eight items match.

Class 5: SSN known but doesn't match. Some cases were moved from Class 5 to Class 3 because of indications that the reported SSN belonged to the spouse.

In this classification scheme all of Class 1 are considered to be true matches implying that the individuals are deceased while all of the Class 5 matches are considered false matches. Assignment of records falling into one of Classes 2, 3, or 4, as either true matches or false matches was made based on the score and cut-off points within class. Records with scores greater than the cut-off scores are considered true matches while records with scores lower than the cut-off scores are considered false matches.

The cut-off scores were determined from the NHANES I developmental sample using a logistic model. The logistic model was used within each of classes 2, 3, and 4 to determine cut-off scores in such a manner as to jointly maximize the ***number and proportion*** of records correctly classified while minimizing the ***number and proportion*** of records incorrectly classified. The cut-off scores were then applied to the NHANES I confirmation sample for refinement. Slight fine-tuning of the cut-off scores was required at this stage because of the relatively small sample sizes. Finally the weights and cut-off scores were applied to the LSOA sample for final confirmation. Further refinements to the cut-off scores were not made.

Results

The recommended cut-off scores are estimated to correctly classify over 97 percent of NHIS decedents and over 99 percent of living persons. It is known that the NDI misses about five percent of known decedents. An adjustment for this has not been included in these classification rates.

Subgroup Biases in Classification

The correct classification rate for females who were known to be deceased is about 2.5 percentage points poorer for females than males. This is due to linkage problems caused by changing surnames through marriages, divorces, and widowhood. Even though father's surname is being used to provide additional information there still remain problems of correctly reporting and recording surnames in both the survey and on the death certificates. Both males and females have the same correct classification rates for living persons.

Among non-whites there are multiple problems including lower reporting of social security numbers and incorrect spelling/recording of ethnic names. The correct classification rates for non-white decedents dropped to 86 percent while the classification rate for living persons remained high at over 99 percent. The classification rate for deceased non-white females was about three percent lower than that for non-white male decedents (84.7 percent and 87.8 percent, respectively). These biases are due to the relatively large proportions of non-white decedents in Class 4 because of incorrect matching information. Females and non-whites falling into Classes 1, 2, 3, or 5 have the same classification rates as white males.

Discussion

Application of the above outlined matching and classification methodology to 1986 through 1994 NHIS survey year respondents provides death follow-up from the date of interview through 1995. The linkage of these files yields approximately 900 deaths for each survey year for each year of follow-up. For example, there are 7,555 deaths among respondents to the 1987 survey with an average of 8½ years of follow-up. Although years can be combined to increase the sample sizes for data items included in the NHIS core (BHD items), this is not generally the case for supplements which change topic areas each year. NHIS supplements are usually administered to one randomly chosen person age 18 or over in each household. This results in an annual sample size for the NHIS of about 42,000 persons. The number of deaths among such supplement respondents would be approximately one-half the number of deaths listed above (e.g., about 450 deaths per survey year per year of follow-up).

The NHIS-NDI linked files (NHIS Multiple Cause of Death Files) can be used to estimate mortality rates (although caution must be given to biases), life expectancies, and relative risks or odds ratios of death for a wide variety of risk factors while controlling for the influence of covariates. For example, the impact of poverty or health insurance status on the risk of dying could be explored while simultaneously controlling for age, sex, race, acute or chronic conditions. Or, mortality rates according to industry or occupation could be developed or for central city residents relative to rural residents. Such analyses are possible because the NHIS carries its own denominators (number at risk).

References

Belin, T.R., and Rubin, D.B. (1995). A Method for Calibrating False-Match Rates in Record Linkage, *Journal of the American Statistical Association*, 90, 430, 694-707.

Belin, T.R. (1993). Evaluation of Sources of Variation in Record Linkage Through a Factorial Experiment, *Survey Methodology*, 19, 1, 13-29.

Boyle, C.A., and Decoufle, P. (1990). National Sources of Vital Status Information: Extent of Coverage and Possible Selectivity in Reporting, *American Journal of Epidemiology*, 131, 160-168.

Calle, E. E., and Terrell, D.D. (1993). Utility of the National Death Index for Ascertainment of Mortality among Cancer Prevention Study II Participants, *American Journal of Epidemiology,* Vol 137, 235-241.

Chyba, M.M., and Washington, L.R. (1993). Questionnaires from the National Health Interview Survey, 1985-89, National Center for Health Statistics, *Vital and Health Statistics*, 1(31), DHHS Publication No. (PHS) 93-1307, Public Health Service, Washington, D.C. U.S. Government Printing Office.

Curb, J.D.; Ford, C.E.; Pressel, S.; Palmer, M.; Babcock, C.; and Hawkins, C.M. (1985). Ascertainment of Vital Status Through the National Death Index and the Social Security Administration, *American Journal of Epidemiology*, 121, 754-766.

Dawson, D.A., and Adams, P.F. (1987). Current Estimates from the National Health Interview Survey, United States, 1986, National Center for Health Statistics, *Vital and Health Statistics*, Series 10, No. 164, DHHS Pub. No. (PHS) 87-1592, Public Health Service, Washington, D.C. U.S. Government Printing Office.

Fellegi, I.P., and Sunter, A.B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association,* 64, 1183-1210.

Finucance, F.F.; Freid, V.M.; Madans, J.H.; Cox, M.A.; Kleinman, J.C.; Rothwell, S.T.; Barbano, H.E.; and Feldman, J.J. (1990). Plan and Operation of the NHANES I Epidemiologic Followup Study, 1986, National Center for Health Statistics, *Vital and Health Statistics*, Series 1, No. 25, DHHS Pub. No. (PHS) 90-1307, Public Health Service, Washington, D. C. U.S. Government Printing Office.

Horm, J.W., and Wright, R.A. (1993). A New National Source of Health and Mortality Information in the United States, *Proceedings of the Social Statistics Section, American Statistical Association,* San Francisco.

Kovar, M.G.; Fitti, J.E.; and Chyba, M.M. (1992). The Longitudinal Study on Aging: 1984-90. National Center for Health Statistics, *Vital and Health Statistics*, 1(28).

National Center for Health Statistics (1990). *National Death Index User's Manual*, U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Health Statistics, DHHS Pub. No. (PHS) 90-1148.

Newcombe, H.B.; Fair, M.E.; and Lalonde, P. (1992). The Use of Names for Linking Personal Records. *Journal of the American Statistical Association,* 87, 1193-1208.

Patterson, B.H., and Bilgrad, R. (1986). Use of the National Death Index in Cancer Studies, *Journal of the National Cancer Institute*, 77, 877-881.

Rogot, E.; Sorlie, P.; and Johnson, N.J. (1986). Probabilistic Methods in Matching Census Samples to the National Death Index, *Journal of Chronic Diseases,* 39, 719-734.

Stampfer, M.J.; Willett, W.C.; Speizer, F.E.; Dysert, D.C.; Lipnick, R.; Rosner, B.; and Hennekins, C.H. (1984). Test of the National Death Index, *American Journal of Epidemiology,* 119, 837-839.

Tepping, B.J., (1968). A Model for Optimum Linkage of Records, *Journal of the American Statistical Association,* 63, 1321-1332.

Williams, B.C.; Demitrack, L.B.; and Fries, B.E. (1992). The Accuracy of the National Death Index When Personal Identifiers Other than Social Security Number Are Used, *American Journal of Public Health,*

82, 1145-1147.

Chapter 4

A Method for Calibrating False-Match Rates in Record Linkage*

Thomas R. Belin, UCLA and Donald B. Rubin, Harvard University

Specifying a record-linkage procedure requires both (1) a method for measuring closeness of agreement between records, typically a scalar weight, and (2) a rule for deciding when to classify records as matches or nonmatches based on the weights. Here we outline a general strategy for the second problem, that is, for accurately estimating false-match rates for each possible cutoff weight. The strategy uses a model where the distribution of observed weights are viewed as a mixture of weights for true matches and weights for false matches. An EM algorithm for fitting mixtures of transformed-normal distributions is used to find posterior modes; associated posterior variability is due to uncertainty about specific normalizing transformations as well as uncertainty in the parameters of the mixture model, the latter being calculated using the SEM algorithm. This mixture-model calibration method is shown to perform well in an applied setting with census data. Further, a simulation experiment reveals that, across a wide variety of settings not satisfying the model's assumptions, the procedure is slightly conservative on average in the sense of overstating false-match rates, and the one-sided confidence coverage (i.e., the proportion of times that these interval estimates cover or overstate the actual false-match rate) is very close to the nominal rate.

KEY WORDS: Box–Cox transformation; Candidate matched pairs; EM algorithm; Mixture model; SEM algorithm; Weights.

1. AN OVERVIEW OF RECORD LINKAGE AND THE PROBLEM OF CALIBRATING FALSE-MATCH RATES

1.1 General Description of Record Linkage

Record linkage (or computer matching, or exact matching) refers to the use of an algorithmic technique to identify records from different data bases that correspond to the same individual. Record-linkage techniques are used in a variety of settings; the current work was formulated and first applied in the context of record linkage between the census and a large-scale postenumeration survey (the PES), which comprises the first step of an extensive matching operation conducted to evaluate census coverage for subgroups of the population (Hogan 1992). The goal of this first step is to declare as many records as possible "matched" without an excessive rate of error, thereby avoiding the cost of the resulting manual processing for all records not declared "matched."

Specifying a record-linkage procedure requires both a method for measuring closeness of agreement between records and a rule using this measure for deciding when to classify records as matches. Much attention has been paid in the record-linkage literature to the problem of assigning "weights" to individual fields of information in a multivariate record and obtaining a "composite weight" that summarizes the closeness of agreement between two records (see, for example, Copas and Hilton 1990; Fellegi and Sunter 1969; Newcombe 1988; and Newcombe, Kennedy, Axford, and James 1959). Somewhat less attention has been paid to the problem of deciding when to classify records as matches, although various approaches have been offered by Tepping (1968), Fellegi and Sunter (1969), Rogot, Sorlie, and Johnson (1986), and Newcombe (1988). Our work focuses on the second problem by providing a predicted probability of match for two records, with associated standard error, as a function of the composite weight.

The context of our problem, computer matching of census records, is typical of record linkage. After data collection, preprocessing of data, and determination of weights, the next step is the assignment of candidate matched pairs where each pair of records consists of the best potential match for each other from the respective data bases (cf. "hits" in Rogot, Sorlie, and Johnson 1986; "pairs" in Winkler 1989; "assigned pairs" in Jaro 1989). According to specified rules, a scalar weight is assigned to each candidate pair, thereby ordering the pairs. The final step of the record linkage procedure is viewed as a decision problem where three actions are possible for each candidate matched pair: declare the two records matched, declare the records not matched, or send both records to be reviewed more closely (see, for example, Fellegi and Sunter 1969). In the motivating problem at the U.S. Census Bureau, a binary choice is made between the alternatives "declare matched" versus "send to followup," although the matching procedure attempts to draw distinctions within the latter group to make manual matching easier for follow-up clerks. In such a setting, a cutoff weight is needed above which records are declared matched; the false-match rate is then defined as the number of falsely matched pairs divided by the number of declared matched pairs. Particularly relevant for any such decision problem is an accurate method for assessing the probability that a candidate matched pair is a correct match as a function of its weight.

1.2 The Need for Better Methods of Classifying Records as Matches or Nonmatches

Belin (1989a, 1989b, 1990) studied various weighting procedures (including some suggested by theory, some used in practice, and some new simple ad hoc weighting schemes) in the census matching problem and reached three primary

* Thomas R. Belin is Assistant Professor, Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA, 90024. Donald B. Rubin is Professor, Department of Statistics, Harvard University, Cambridge, MA, 02138. The authors would like to thank William Winkler of the U.S. Census Bureau for a variety of helpful discussions. The authors also gratefully acknowledge the support of Joint Statistical Agreements 89-07, 90-23, and 91-08 between the Census Bureau and Harvard University, which helped make this research possible. Much of this work was done while the first author was working for the Record Linkage Staff of the Census Bureau; the views expressed are those of the authors and do not necessarily reflect those of the Census Bureau.

Journal of the American Statistical Association
June 1995, Vol. 90, No. 430, Theory and Methods

conclusions. First, different weighting procedures lead to comparable accuracy in computer matching. Second, as expected logically and from previous work (e.g., Newcombe 1988, p. 144), the false-match rate is very sensitive to the setting of a cutoff weight above which records will be declared matched. Third, and more surprising, current methods for estimating the false-match rate are extremely inaccurate, typically grossly optimistic.

To illustrate this third conclusion, Table 1 displays empirical findings from Belin (1990) with test-census data on the performance of the procedure of Fellegi and Sunter (1969), which relies on an assumption of independence of agreement across fields of information. That the Fellegi–Sunter procedure for estimating false-match rates does not work well (i.e., is poorly calibrated) may not be so surprising in this setting, because the census data being matched do not conform well to the model of mutual independence of agreement across the fields of information (see, for example, Kelley 1986 and Thibaudeau 1989). Other approaches to estimating false-match rates that rely on strong independence assumptions (e.g., Newcombe 1988) can be criticized on similar grounds.

Although the Fellegi–Sunter approach to setting cutoff weights was originally included in census/PES matching operations (Jaro 1989), in the recent past (including in the 1990 Census) the operational procedure for classifying record pairs as matches has been to have a human observer establish cutoff weights manually by "eyeballing" lists of pairs of records brought together as candidate matches. This manual approach is easily criticized, both because the error properties of the procedure are unknown and variable and because, when linkage is done in batches at different times or by different persons, inconsistent standards are apt to be applied across batches.

Another idea is to use external data to help solve this classification problem. For example, Rogot, Sorlie, and Johnson (1986) relied on extreme order statistics from pilot data to determine cutoffs between matches and nonmatches; but this technique can be criticized, because extreme order statistics may vary considerably from sample to sample, especially when sample sizes are not large. One other possibility, discussed by Tepping (1968), requires clerical review of samples from the output of a record-linkage procedure to provide feedback on error rates to refine the calibration procedure. Such feedback is obviously desirable, but in many applications, including the census/PES setting, it is impossible to provide it promptly enough.

Table 1. Performance of Fellegi–Sunter Cutoff Procedure on 1986 Los Angeles Test-Census Data

Acceptable false-match rate specified by user of matching program	Observed false-match rate among declared matched pairs
.05	.0627
.04	.0620
.03	.0620
.02	.0619
.01	.0497
10^{-3}	.0365
10^{-4}	.0224
10^{-5}	.0067
10^{-6}	.0067
10^{-7}	.0067

A more generally feasible strategy is to use the results of earlier record-linkage studies in which all candidate matched pairs have been carefully reviewed by clerks. This type of review is common practice in operations conducted by the Census Bureau. Each such training study provides a data set in which each candidate pair has its weight and an outcome, defined as true match or false match, and thus provides information for building a model to give probability of match as a function of weight.

1.3 A Proposed Solution to the Problem of Calibrating Error Rates

There are two distinct approaches to estimating the relationship between a dichotomous outcome, $Z_i = 1$ if match and $Z_i = 0$ if nonmatch, from a continuous predictor, the weight, W_i: the direct approach, typified by logistic regression, and the indirect approach, typified by discriminant analysis. In the direct approach, an iid model is of the form $f(Z_i \mid W_i, \nu) \times g(W_i \mid \zeta)$, where $g(W_i \mid \zeta)$, the marginal distribution of W_i, is left unspecified with ζ a priori independent of ν. In the indirect approach, the iid model is of the form $h(W_i \mid Z_i, \phi)[\lambda^{Z_i}(1-\lambda)^{(1-Z_i)}]$, where the first factor specifies, for example, a normal conditional distribution of W_i for $Z_i = 0$ and for $Z_i = 1$ with common variance but different means, and the second factor specifies the marginal probability of $Z_i = 1$, λ, which is a priori independent of ϕ. Under this approach, $P(Z_i \mid W_i)$ is found using Bayes's theorem from the other model specifications as a function of ϕ and λ. Many authors have discussed distinctions between the two approaches, including Halperin, Blackwelder, and Verter (1971), Mantel and Brown (1974), Efron (1975), and Dawid (1976).

In our setting, application of the direct approach would involve estimating $f(Z_i \mid W_i, \nu)$ in observed sites where determinations of clerks had established Z_i, and then applying the estimated value of ν to the current site with only W_i observed to estimate the probability of match for each candidate pair. If the previous sites differed only randomly from the current sites, or if the previous sites were a subsample of the current data selected on W_i, then this approach would be ideal. Also, if there were many previous sites and each could be described by relevant covariates, such as urban/rural and region of the country, then the direct approach could estimate the distribution of Z as a function of W and covariates and could use this for the current site. Limited experience of ours and of our colleagues at the Census Bureau, who investigated this possibility using 1990 Census data, has resulted in logistic regression being rejected as a method for estimating false-match rates in the census setting (W. E. Winkler 1993, personal communication).

But the indirect approach has distinct advantages when, as in our setting, there can be substantial differences among sites that are not easily modeled as a function of covariates and we have substantial information on the distribution of weights given true and false matches, $h(\cdot \mid \cdot)$. In particular,

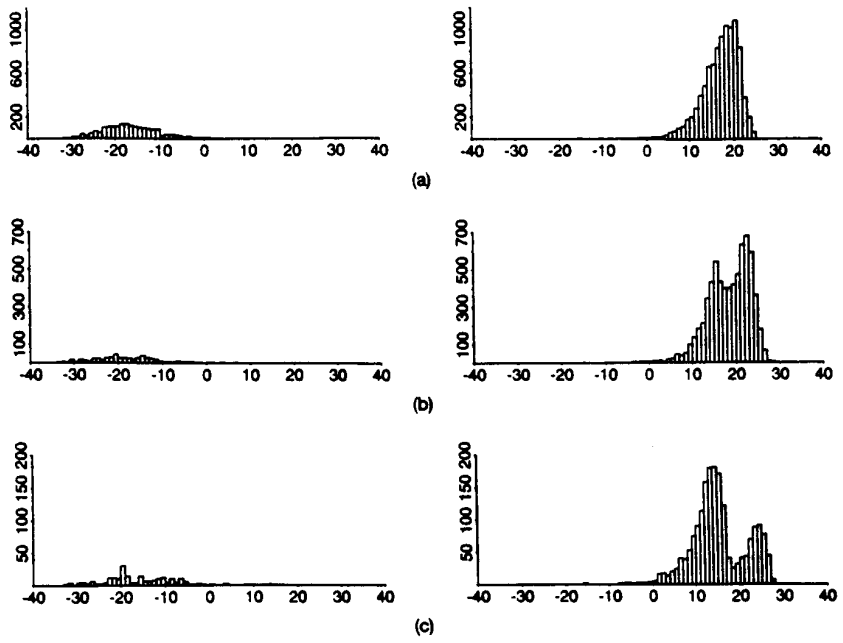


Figure 1. Histograms of Weights for True and False Matches by Site: (a) St. Louis; (b) Columbia; (c) Washington.

with the indirect approach, the observed marginal distribution of W_i in the current site is used to help estimate $P(Z_i \mid W_i)$ in this site, thereby allowing systematic site to site differences in $P(Z_i \mid W_i)$. In addition, there can be substantial gains in efficiency using the indirect approach when normality holds (Efron 1975), especially when $h(W_i \mid Z_i = 1, \phi)$ and $h(W_i \mid Z_i = 0, \phi)$ are well separated; (that is, when the number of standard deviations difference between their means is large).

Taking this idea one step further, suppose that previous validated data sets had shown that after a known transformation, the true-match weights were normally distributed, and that after a different known transformation, the false-match weights were normally distributed. Then, after inverting the transformations, $P(Z_i \mid W_i)$ could be estimated in the current site by fitting a normal mixture model, which would estimate the means and variances of the two normal components (i.e., ϕ) and the relative frequency of the two components (i.e., λ), and then applying Bayes's theorem. In this example, instead of assuming a common $P(Z_i \mid W_i)$ across all sites, only the normality after the fixed transformations would be assumed common across sites. If there were many sites with covariate descriptors, then (λ, ϕ) could be modeled as a function of these, for example, a linear model structure on the normal means.

To illustrate the application of our work, we use available test-census data consisting of records from three separate sites of the 1988 dress rehearsal Census and PES: St. Louis, Missouri, with 12,993 PES records; a region in East Central Missouri including Columbia, Missouri, with 7,855 PES records; and a rural area in eastern Washington state, with only 2,318 records. In each site records were reviewed by clerks, who made a final determination as to the actual match status of each record; for the purpose of our discussion, the clerks' determinations about the match status of record pairs are regarded as correct. The matching procedures used in the 1988 test Census have been documented by Brown et al. (1988), Jaro (1989), and Winkler (1989). Beyond differences in the sizes of the PES files, the types of street addresses in the areas offer considerably different amounts of information for matching purposes; for instance, rural route addresses, which were common in the Washington site but almost nonexistent in the St. Louis site, offer less information for matching than do most addresses commonly found in urban areas.

Figure 1 shows histograms of both true-match weights and false-match weights from each of the three sites. The bimodality in the true-match distribution for the Washington site appears to be due to some record pairs agreeing on address information and some not agreeing. This might generate concern, not so much for lack of fit in the center of the distribution as for lack of fit in the tails, which are essential to false-match rate estimation. Of course, it is not surprising that validated data dispel the assumption of normality for true-match weights and false-match weights. They do, however—at least at a coarse level in their apparent skewness—tend to support the idea of a similar nonnormal distributional shape for true-match weights across sites as well as a similar nonnormal distributional shape for false-match weights across sites. Moreover, although the locations of these distributions change from site to site, as do the relative frequencies of the true-match to the false-match components, the relative spread of the true to false components is similar across sites.

These observations lead us to formulate a transformed-normal mixture model for calibrating false-match rates in record-linkage settings. In this model, two power (or Box–Cox) transformations are used to normalize the false-match weights and the true-match weights, so that the observed raw weights in a current setting are viewed as a mixture of two transformed normal observations.

Mixture models have been used in a wide variety of statistical applications (see Titterington, Smith, and Makov 1985, pp. 16–21, for an extensive bibliography). Power transformations are also used widely in statistics, prominently in an effort to satisfy normal theory assumptions in regression settings (see, for example, Weisberg 1980, pp. 147–151). To our knowledge, neither of these techniques has been utilized in record linkage operations, nor have mixtures of transformed-normal distributions, with different transformations in the groups, appeared previously in the statistical literature, even though this extension is relatively straightforward. The most closely related effort to our own of which we are aware is that of Maclean, Morton, Elston, and Yee (1976), who used a common power transformation for different components of a mixture model, although their work focused on testing for the number of mixture components.

Section 2 describes the technology for fitting mixture models with components that are normally distributed after application of a power transformation, which provides the statistical basis for the proposed calibration method. This section also outlines the calibration procedure itself, including the calculation of standard errors for the predicted false-match rate. Section 3 demonstrates the performance of the method in the applied setting of matching the Census and PES, revealing it to be quite accurate. Section 4 summarizes a simulation experiment to gauge the performance of the calibration procedure in a range of hypothetical settings and this too supports the practical utility of the proposed calibration approach. Section 5 concludes the article with a brief discussion.

2. CALIBRATING FALSE-MATCH RATES IN RECORD LINKAGE USING TRANSFORMED-NORMAL MIXTURE MODELS

2.1 Strategy Based on Viewing Distribution of Weights as Mixture

We assume that a univariate composite weight has been calculated for each candidate pair in the record-linkage problem at hand, so that the distribution of observed weights is a mixture of the distribution of weights for true matches and the distribution of weights for false matches. We also assume the availability of at least one training sample in which match status (i.e., whether a pair of records is a true match or a false match) is known for all record pairs. In our applications, training samples come from other geographical locations previously studied. We implement and study the following strategy for calibrating the false-match rate in a current computer-matching problem:

1. Use the training sample to estimate "global" parameters, that is, the parameters of the transformations that normalize the true- and false-match weight distributions and the parameter that gives the ratio of variances between the two components on the transformed scale. The term "global" is used to indicate that these parameters are estimated by data from other sites and are assumed to be relatively constant from site to site, as opposed to "site-specific" parameters, which are assumed to vary from site to site and are estimated only by data from the current site.

2. Fix the values of the global parameters at the values estimated from the training sample and fit a mixture of transformed-normal distributions to the current site's weight data to obtain maximum likelihood estimates (MLE's) and associated standard errors of the component means, component variances, and mixing proportion. We use the EM algorithm (Dempster, Laird, and Rubin 1977) to obtain MLE's and the SEM algorithm (Meng and Rubin 1991) to obtain asymptotic standard errors.

3. For each possible cutoff level for weights, obtain a point estimate for the false-match rate based on the parameter estimates from the model and obtain an estimate of the standard error of the false-match rate. In calculating standard errors, we rely on a large-sample approximation that makes use of the estimated covariance matrix obtained from the SEM algorithm.

An appealing modification of this approach, which we later refer to as our "full strategy," reflects uncertainty in global parameters through giving them prior distributions. Then, rather than fixing the global parameters at their estimates from the training sample, we can effectively integrate over the uncertainty in the global parameters by modifying Step 2 to be:

2′. Draw values of the global parameters from their posterior distribution given training data, fix global parameters at their drawn values, and fit a mixture of transformed-normal distributions to the current weight data to obtain MLE's (and standard errors) of site-specific parameters;

and adding:

4. Repeat Steps 2′ and 3 a few or several times, obtaining false-match rate estimates and standard errors from each repetition, and combine the separate estimates and standard errors into a single point estimate and standard error that reflect uncertainty in the global parameters using the multiple imputation framework of Rubin (1987).

We now describe how to implement each of these steps.

2.2 Using a Training Sample to Estimate Global Parameters

Box and Cox (1964) offered two different parameterizations for the power family of transformations: one that ignores the scale of the observed data, and the other—which we will use—that scales the transformations by a function of the observed data so that the Jacobian is unity. We denote the family of transformations by

$$\psi(w_i; \gamma, \omega) = \frac{w_i^{\gamma} - 1}{\gamma \omega^{\gamma-1}} \quad \text{if } \gamma \neq 0,$$

$$= \omega \log(w_i) \quad \text{if } \gamma \neq 0, \qquad (1)$$

where ω is the geometric mean of the observations $w_1, \ldots, w_n$.

By "transformed-normal distribution," we mean that for some unknown values of γ and ω, the transformed observations $\psi(w_i; \gamma, \omega)$ $(i = 1, \ldots, n)$ are normally distributed. Although the sample geometric mean is determined by the data, we will soon turn to a setting involving a mixture of two components with different transformations to normality, in which even the sample geometric means of the two components are unknown; consequently, we treat ω as an unknown parameter, the population geometric mean.

When the transformations are not scaled by the geometric-mean factor, as Box and Cox (1964, p. 217) noted, "the general size and range of the transformed observations may depend strongly on $[\gamma]$." Of considerable interest in our setting is that when transformations are scaled, not only are the likelihoods for different values of γ directly comparable, at least asymptotically, but also, by implication, so are the residual sums of squares on the transformed scales for different values of γ. In other words, scaling the transformations by $\omega^{\gamma-1}$ has the effect asymptotically of unconfounding the estimated variance on the transformed scale from the estimated power parameter. This result is important in the context of fitting mixtures of transformed-normal distributions when putting constraints on component variances in the fitting of the mixture model; by using scaled transformations, we can constrain the variance ratio without reference to the specific power transformation that has been applied to the data.

Box and Cox (1964) also considered an unknown location parameter in the transformation, which may be needed because power transformations are defined only for positive random variables. Because the weights that arise from record-linkage procedures are often allowed to be negative, this issue is relevant in our application. Nevertheless, Belin (1991) reported acceptable results using an ad hoc linear transformation of record-linkage weights to a range from 1 to 1,000. Although this ad hoc shift and rescaling is assumed to be present, we suppress the parameters of this transformation in the notation.

In the next section we outline in detail a transformed-normal mixture model for record-linkage weights. Fitting this model requires separate estimates of γ and ω for the true-match and false-match distributions observed in the training data, as well as an estimate of the ratio of variances on the transformed scale. The γ's can, as usual, be estimated by performing a grid search of the likelihoods or of the respective posterior densities. A modal estimate of the variance ratio can be obtained as a by-product of the estimation of the γ's. We also obtain approximate large-sample variances by calculating for each parameter a second difference as numerical approximation to the second derivative of the log-likelihood in the neighborhood of the maximum (Belin 1991). In our work we have simply fixed the ω's at their sample values, which appeared to be adequate based on the overall success of the methodology on both real and simulated data; were it necessary to obtain a better fit to the data, this approach could be modified.

2.3 Fitting Transformed Normal Mixtures with Fixed Global Parameters

2.3.1 Background on Fitting Normal Mixtures Without Transformations. Suppose that f_1 and f_2 are densities that depend on an unknown parameter ϕ, and that the density f is a mixture of f_1 and f_2, i.e., $f(X|\phi, \lambda) = \lambda f_1(X|\phi) + (1 - \lambda) f_2(X|\phi)$ for some λ between 0 and 1. Given an iid sample $(X_1, X_2, \ldots, X_n)$ from $f(X|\phi, \lambda)$, the likelihood of $\theta = (\phi, \lambda)$ can then be written as

$$L(\theta|X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i|\theta)$$

$$= \prod_{i=1}^{n} [\lambda f_1(X_i|\phi) + (1 - \lambda) f_2(X_i|\phi)].$$

Following the work of many authors (e.g., Aitkin and Rubin 1985; Dempster et al. 1977; Little and Rubin 1987; Orchard and Woodbury 1972; Titterington et al. 1985), we formulate the mixture model in terms of unobserved indicators of component membership Z_i, $i = 1, \ldots, n$, where $Z_i = 1$ if X_i comes from component 1 and $Z_i = 0$ if X_i comes from component 2. The mixture model can then be expressed as a hierarchical model,

$$X_i|\{Z_i = 1\}, \theta \overset{\text{iid}}{\sim} f_1(\cdot|\phi)$$

$$X_i|\{Z_i = 0\}, \theta \overset{\text{iid}}{\sim} f_2(\cdot|\phi)$$

$$Z_i|\theta \overset{\text{iid}}{\sim} \text{Bernoulli}(\lambda).$$

The "complete-data" likelihood, which assumes that the "missing data" $Z_1, \ldots, Z_n$ are observed, can be written as

$$L(\phi, \lambda|X_1, \ldots, X_n; Z_1, \ldots, Z_n)$$

$$= \prod_{i=1}^{n} [\lambda f_1(X_i|\phi)]^{Z_i} [(1 - \lambda) f_2(X_i|\phi)]^{1-Z_i}.$$

Viewing the indicators for component membership as missing data motivates the use of the EM algorithm to obtain MLE's of (ϕ, λ). The E step involves finding the expected value of the Z_i's given the data and current parameter estimates $\phi^{(t)}$ and $\lambda^{(t)}$, where t indexes the current iteration. This is computationally straightforward both because the iid structure of the model implies that Z_i is conditionally independent of the rest of the data given X_i and because the Z_i's are indicator variables, so the expectation of Z_i is simply the posterior probability that Z_i equals 1. Using Bayes's theorem, the E step at the $(t + 1)$st iteration thus involves calculating

$$Z_i^{(t+1)} = E(Z_i|X_1, \ldots, X_n; \phi^{(t)}, \lambda^{(t)})$$

$$= \frac{\lambda^{(t)} f_1(X_i|\phi^{(t)})}{\lambda^{(t)} f_1(X_i|\phi^{(t)}) + (1 - \lambda^{(t)}) f_2(X_i|\phi^{(t)})} \quad (2)$$

for $i = 1, \ldots, n$.

The M step involves solving for MLE's of θ in the "complete-data" problem. In the case where f_1 corresponds to the $N(\mu_1, \sigma_1^2)$ distribution and f_2 corresponds to the $N(\mu_2, \sigma_2^2)$ distribution, so that $\phi = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, the M step at

iteration $(t+1)$ involves calculating

$$\mu_1^{(t+1)} = \frac{\sum_{i=1}^n Z_i^{(t+1)} X_i}{\sum_{i=1}^n Z_i^{(t+1)}}$$

$$\mu_2^{(t+1)} = \frac{\sum_{i=1}^n (1 - Z_i^{(t+1)}) X_i}{\sum_{i=1}^n (1 - Z_i^{(t+1)})} \quad (3)$$

and

$$\sigma_1^{2(t+1)} = \frac{\sum_{i=1}^n Z_i^{(t+1)} (X_i - \mu_1^{(t+1)})^2}{\sum_{i=1}^n Z_i^{(t+1)}}$$

$$\sigma_2^{2(t+1)} = \frac{\sum_{i=1}^n (1 - Z_i^{(t+1)}) (X_i - \mu_2^{(t+1)})^2}{\sum_{i=1}^n (1 - Z_i^{(t+1)})}. \quad (4)$$

The updated value of λ at the $(t+1)$st iteration is given by

$$\lambda^{(t+1)} = \frac{1}{n} \sum_{i=1}^n Z_i^{(t+1)}, \quad (5)$$

which holds no matter what the form of the component densities may be. Instabilities can arise in maximum likelihood estimation for normally distributed components with distinct variances, because the likelihood is unbounded at the boundary of the parameter space where either $\sigma_i^2 = 0$. Unless the starting values for EM are near a local maximum of the likelihood, EM can drift toward the boundary where the resulting fitted model suggests that one component consists of any single observation (with zero variance) and that the other component consists of the remaining observations (Aitkin and Rubin 1985).

When a constraint is placed on the variances of the two components, EM will typically converge to an MLE in the interior of the parameter space. Accordingly, a common approach in this setting is to find a sensible constraint on the variance ratio between the two components or to develop an informative prior distribution for the variance ratio. When the variance ratio $V = \sigma_1^2/\sigma_2^2$ is assumed fixed, the E step proceeds exactly as in (2) and the M step for $\mu_1^{(t+1)}$ and $\mu_2^{(t+1)}$ is given by (3); the M step for the scale parameters with fixed V is

$$\sigma_1^{2(t+1)} = \frac{1}{n}\left[\sum_{i=1}^n Z_i^{(t+1)} (X_i - \mu_1^{(t+1)})^2 + V(1 - Z_i^{(t+1)}) \times (X_i - \mu_2^{(t+1)})^2\right], \quad \sigma_2^{2(t+1)} = \frac{\sigma_1^{2(t+1)}}{V}. \quad (6)$$

2.3.2 Modifications to Normal Mixtures for Distinct Transformations of the Two Components. We now describe EM algorithms for obtaining MLE's of parameters in mixtures of transformed-normal distributions, where there are distinct transformations of each component. Throughout the discussion, we will assume that there are exactly two components; fitting mixtures of more than two components involves straightforward extensions of the arguments that follow (Aitkin and Rubin 1985).

We will also assume that the transformations are fixed; that is, we assume that the power parameters (the two γ_i's) and the "geometric-mean" parameters (the two ω_i's) are known in advance and are not to be estimated from the data. We can write the model for a mixture of transformed-normal components as follows:

$$X_i | \theta, Z_i = 1 \sim \text{Transformed-}N(\mu_1, \sigma_1^2, \gamma_1, \omega_1),$$

$$X_i | \theta, Z_i = 0 \sim \text{Transformed-}N(\mu_2, \sigma_2^2, \gamma_2, \omega_2),$$

$$Z_i | \theta \sim \text{Bernoulli}(\lambda),$$

where $\theta = \{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda, \gamma_1, \gamma_2, \omega_1, \omega_2\}$ and the expression "Transformed-N" with four arguments refers to the transformed-normal distribution with the four arguments being the location, scale, power parameter, and "geometric-mean" parameter of the transformed-normal distribution. The complete-data likelihood can be expressed as

$$L(\theta | X_1, \ldots, X_n; Z_1, \ldots, Z_n) = \prod_{i=1}^n \left[\frac{1}{\sigma_1\sqrt{2\pi}} e^{-1/2[(\psi(X_i;\gamma_1,\omega_1)-\mu_1)^2/\sigma_1^2]} \left|\left(\frac{X_i}{\omega_1}\right)^{\gamma_1 - 1}\right|\right]^{Z_i} \times \left[\frac{1}{\sigma_2\sqrt{2\pi}} e^{-1/2[(\psi(X_i;\gamma_2,\omega_2)-\mu_2)^2/\sigma_2^2]} \left|\left(\frac{X_i}{\omega_2}\right)^{\gamma_2 - 1}\right|\right]^{(1-Z_i)} J_1 J_2,$$

where J_1 and J_2 are the Jacobians of the scaled transformations $X \to \psi$. If ω_1 and ω_2 were not fixed a priori but instead were the geometric means of the X_i for the respective components, then $J_1 = J_2 = 1$. In our situation, however, because the Z_i's are missing, J_1 and J_2 are functions of $\{X_i\}$, $\{Z_i\}$, and θ, and are not generally equal to 1. Still, J_1 and J_2 are close to 1 when the estimated geometric mean of the sample X_i in component k is close to ω_k. We choose to ignore this minor issue; that is, although we model ω_1 and ω_2 as known from prior considerations, we still assume $J_1 = J_2 = 1$. To do otherwise would greatly complicate our estimation procedure with, we expect, no real benefit; we do not blindly believe such fine details of our model in any case, and we would not expect our procedures to be improved by the extra analytic work and computational complexity.

To keep the distinction clear between the parameters assumed fixed in EM and the parameters being estimated in EM, we partition the parameter into $\theta = \{\theta_{fix}, \theta_{est}\}$, where $\theta_{fix} = \{\gamma_1, \gamma_2, \omega_1, \omega_2, V\}$ and $\theta_{est} = \{\mu_1, \mu_2, \sigma_1^2, \lambda\}$, and where the variance ratio $V = \sigma_1^2/\sigma_2^2$. Based on this formulation, MLE's of θ_{est} can be obtained from the following EM algorithm:

E step. For $i = 1, \ldots, n$, calculate $Z_i^{(t+1)}$ as in (2), where

$$f_g(X_i | \theta^{(t)}) = \frac{1}{\sigma_g\sqrt{2\pi}} e^{-1/2[(\psi(X_i;\gamma_g,\omega_g)-\mu_g)^2/\sigma_g^2]} \left|\left(\frac{X_i}{\omega_g}\right)^{\gamma_g - 1}\right|, \quad g = 1, 2. \quad (7)$$

M step. Calculate $\mu_1^{(t+1)}$ and $\mu_2^{(t+1)}$ as in (3), $\lambda^{(t+1)}$ as in (5), and $\sigma_1^{2(t+1)}$ and $\sigma_2^{2(t+1)}$ as in (6), with X_i replaced by $\psi(X_i; \gamma_g, \omega_g)$ for $g = 1, 2$; if the variance ratio V were not fixed but were to be estimated, then (4) would be used in place of (6).

2.3.3 Transformed-Normal Mixture Model for Record-Linkage Weights. Let the weights associated with record pairs in a current data set be denoted by W_i, $i = 1, \ldots, n$,

where as before $Z_i = 1$ implies membership in the false-match component and $Z_i = 0$ implies membership in the true-match component. We assume that we have already obtained, from a training sample, (a) values of the power transformation parameters, denoted by γ_F for the false-match component and by γ_T for the true-match component, (b) values of the "geometric mean" parameters in the transformations, denoted by ω_F for the false-match component and by ω_T for the true-match component, and (c) a value for the ratio of the variances between the false-match and true-match components, denoted by V. Our model then becomes

$$W_i|\theta, \{Z_i = 1\} \sim \text{Transformed-}N(\mu_F, \sigma_F^2, \gamma_F, \omega_F),$$

$$W_i|\theta, \{Z_i = 0\} \sim \text{Transformed-}N(\mu_T, \sigma_T^2, \gamma_T, \omega_T),$$

$$Z_i|\theta \sim \text{Bernoulli}(\lambda),$$

where $\theta = \{\mu_F, \mu_T, \sigma_F^2, \sigma_T^2, \lambda, \gamma_F, \gamma_T, \omega_F, \omega_T\}$. We work with $\theta_{fix} = \{\gamma_F, \gamma_T, \omega_F, \omega_T, V = \sigma_F^2/\sigma_T^2\}$ and $\theta_{est} = \{\mu_F, \mu_T, \sigma_F^2, \lambda\}$. The algorithm of Section 2.3.2, with "F" and "T" substituted for "1" and "2," describes the EM algorithm for obtaining MLE's of θ_{est} from $\{W_i; i = 1, \ldots, n\}$ with $\{Z_i; i = 1, \ldots, n\}$ missing and global parameters γ_F, γ_T, ω_F, ω_T, and V fixed at specified values.

2.4 False-Match Rate Estimates and Standard Errors with Fixed Global Parameters

2.4.1 Estimates of the False-Match Rate. Under the transformed-normal mixture model formulation, the false-match rate associated with a cutoff C can be expressed as a function of the parameters θ as

$$\varepsilon(C|\theta) = \frac{\lambda\left[1 - \Phi\left(\frac{\psi_F(C; \gamma_F, \omega_F) - \mu_F}{\sigma_F}\right)\right]}{\lambda\left[1 - \Phi\left(\frac{\psi_F(C; \gamma_F, \omega_F) - \mu_F}{\sigma_F}\right)\right] + (1 - \lambda)\left[1 - \Phi\left(\frac{\psi_T(C; \gamma_T, \omega_T) - \mu_T}{\sigma_T}\right)\right]}. \quad (8)$$

Substitution of MLE's for the parameters in this expression provides a predicted false-match rate associated with cutoff C.

Because there is a maximum possible weight associated with perfect agreement in most record-linkage procedures, one could view the weight distribution as truncated above. According to this view, the contribution of the tail above the upper truncation point (say, B), should be discarded by substituting $\Phi([\psi_g(B; \gamma_g, \omega_g) - \mu_g]/\sigma_g)$ for the 1s inside the bracketed expressions (g = F, T as appropriate). Empirical investigation suggests that truncation of the extreme upper tail makes very little difference in predictions. The results in Sections 3 and 4 reflect false-match rate predictions without truncation of the extreme upper tail.

2.4.2 Obtaining an Asymptotic Covariance Matrix for Mixture-Model Parameters From SEM Algorithm. The SEM algorithm (Meng and Rubin 1991) provides a method for obtaining standard errors of parameters in models that are fit using the EM algorithm. The technique uses estimates of the fraction of missing information derived from successive EM iterates to inflate the complete-data variance–covariance matrix to provide an appropriate observed-data variance-covariance matrix. Details on the implementation of the SEM algorithm in our mixture-model setting are deferred to the Appendix.

Standard arguments lead to large-sample standard errors for functions of parameters. For example, the false-match rate $\varepsilon(C|\theta)$ can be expressed as a function of the four components of $\theta_{est} = (\mu_F, \mu_T, \sigma_F^2, \lambda)$ by substituting $\sigma_F/\sqrt{V}$ for σ_T in (8). Then the squared standard error of the estimated false-match rate is given by $SE^2(\varepsilon) \approx \mathbf{d}^T\mathbf{A}\mathbf{d}$, where $\mathbf{A}$ is the covariance matrix for θ_{est} obtained by SEM and the νth component of $\mathbf{d}$ is $d_\nu = \partial\varepsilon/\partial\theta_F$.

2.4.3 Estimates of the Probability of False Match for a Record Pair With a Given Weight. The transformed-normal mixture model also provides a framework for estimating the probability of false match associated with various cutoff weights. To be clear, we draw a distinction between the "probability of false match" and what we refer to as the "neighborhood false-match rate" to avoid any confusion caused by (1) our using a continuous mixture distribution to approximate the discrete distribution of weights associated with a finite number of record pairs, and (2) the fact that there are only finitely many possible weights associated with many record-linkage weighting schemes. The "neighborhood false-match rate around W" is the number of false matches divided by the number of declared matches among pairs of records with composite weights in a small neighborhood of W; with a specific model, the neighborhood false-match rate is the "probability of false match" implied by the relative density of the true-match and false-match components at W.

In terms of the mixture-model parameters, the false-match rate among record pairs with weights between W and $W + h$ is given by

$$\xi(W, h|\theta) = \frac{\lambda p_F(W, h|\theta)}{\lambda p_F(W, h|\theta) + (1 - \lambda)p_T(W, h|\theta)},$$

where

$$p_g(W, h|\theta) = \Phi\left(\frac{\psi_g(W + h; \gamma_g, \omega_g) - \mu_g}{\sigma_g}\right) - \Phi\left(\frac{\psi_g(W; \gamma_g, \omega_g) - \mu_g}{\sigma_g}\right), \quad g = F, T,$$

and $\theta = \{\mu_F, \mu_T, \sigma_F^2, \sigma_T^2, \gamma_F, \gamma_T, \omega_F, \omega_T, \lambda\}$. Although the number of false matches is not a smooth function of the number of declared matches, $\xi(W, h|\theta)$ is a smooth function

of h. The probability of false match under the transformed-normal mixture model is the limit as $h \to 0$ of $\xi(W, h|\theta)$, which we denote as $\eta(W|\theta)$; we obtain

$$\eta(W|\theta) = \lim_{h \to 0} \left\{ \frac{\lambda\left(\frac{\partial p_F(W, h|\theta)}{\partial h}\right)}{\lambda\left(\frac{\partial p_F(W, h|\theta)}{\partial h}\right) + (1-\lambda)\left(\frac{\partial p_T(W, h|\theta)}{\partial h}\right)} \right\} = \frac{\lambda\phi_F^*(W|\theta)}{\lambda\phi_F^*(W|\theta) + (1-\lambda)\phi_T^*(W|\theta)}, \quad (9)$$

where

$$\phi_g^*(W|\theta) = \frac{\lambda}{\sigma_g\sqrt{2\pi}} e^{-1/2(\psi_g(W;\gamma_g,\omega_g)-\mu_g/\sigma_g)^2}\left(\frac{W}{\omega_g}\right)^{\gamma_g - 1}, \quad g = F, T.$$

Estimates of neighborhood false-match rates are thus routinely obtained by substituting the fixed global parameter values and MLE's of μ_F, μ_T, σ_F, σ_T, and λ into (9).

Because the neighborhood false-match rate captures the trade-off between the number of false matches and the number of declared matches, the problem of setting cutoffs can be cast in terms of the question "Approximately how many declared matches are needed to make up for the cost of a false match?" If subject matter experts who are using a record-linkage procedure can arrive at an answer to this question, then a procedure for setting cutoffs could be determined by selecting a cutoff weight where the estimated neighborhood false-match rate equals the appropriate ratio. Alternatively, one could monitor changes in the neighborhood false-match rate (instead of specifying a "tolerable" neighborhood false-match rate in advance) and could set a cutoff weight at a point just before the neighborhood false-match rate accelerates.

2.5 Reflecting Uncertainty in Global Parameters

When there is more than one source of training data, the information available about both within-site and between-site variability in global parameters can be incorporated into the prior specification. For example, with two training sites, we could combine the average within-site variability in a global parameter with a 1 df estimate of between-site variability to represent prior uncertainty in the parameter. With many sites with covariate descriptors, we could model the multivariate regression of global parameters on covariates.

The procedure we used in the application to census/PES data offers an illustration in the simple case with two training sites available to calibrate a third site. For each of the training-data sites and each of the components (true-match and false-match), joint MLE's were found for $(\gamma_g, \mu_g, \sigma_g^2)$, $g = F, T$, using a simple grid-search over the power parameters. This yielded two estimates of the power parameters, γ_F and γ_T, and two estimates of the variance ratio V between the false-match and true-match components. Additionally, an estimated variance–covariance matrix for these three parameters was obtained by calculating second differences of the log-likelihood at grid points near the maximum.

Values of each parameter for the mixture-model fitting were drawn from separate truncated-normal distributions with mean equal to the average of the estimates from the two training sites and variance equal to the sum of the squared differences between the individual site parameter values and their mean (i.e., the estimated "between" variance), plus the average squared standard error from the two prior fittings (i.e., the average "within" variance). The truncation ensured that the power parameter for the false-match component was less than 1, that the power parameter for the true-match component was greater than 1, and that the variance ratio was also greater than 1. These constraints on the power parameters were based on the view that because there is a maximum possible weight corresponding to complete agreement and a minimum possible weight corresponding to complete disagreement, the true-match component will have a longer left tail than right tail and the false-match component will have a longer right tail than left tail. The truncation for the variance ratio was based on an assumption that false-match weights will exhibit more variability than true-match weights for these data on the transformed scale as well as on the original scale.

For simplicity, the geometric-mean terms in the transformations (ω_F and ω_T) were simply fixed at the geometric mean of the component geometric means from the two previous sites. If the methods had not worked as well as they did with test and simulated data, then we would have also reflected uncertainty in these parameters.

Due to the structure of our problem, in which the role of the prior distribution is to represent observable variability in global parameters from training data, we presume that the functional form of the prior is not too important as long as variability in global parameters is represented accurately. That is, we anticipate that any one of a number of methods that reflect uncertainty in the parameters estimated from training data will yield interval estimates with approximately the correct coverage properties, i.e., nominal $(1-\alpha) \times 100\%$ interval estimates will cover the true value of the estimand approximately $(1-\alpha) \times 100\%$ or more of the time. Alternative specifications for the prior distribution were described by Belin (1991).

When we fit multiple mixture models to average over uncertainty in the parameters estimated by prior data (i.e., when we use the "full strategy" of Section 2.1), the multiple-imputation framework of Rubin (1987) can be invoked to combine estimates and standard errors from the separate models to provide one inference. Suppose that we fit m mixture models corresponding to m separate draws of the global parameters from their priors and thereby obtain false-match rate estimates $e_1, e_2, \ldots, e_m$ and variance estimates $u_1, u_2, \ldots, u_m$, where $u_i = SE^2(e_i)$ is obtained by the method of Section 2.4.2. Following Rubin (1987, p. 76), we can estimate the false-match rate by

$$\bar{e} = \frac{1}{m}\sum_{i=1}^{m} e_i$$

and its squared standard error by

$$SE^2(\bar{\varepsilon}) = \frac{m+1}{m}\left[\frac{1}{m-1}\sum_{i=1}^{m}(\varepsilon_i - \bar{\varepsilon})^2\right] + \frac{1}{m}\sum_{i=1}^{m} u_i.$$

Monte Carlo evaluations documented by Rubin (1987, secs. 4.6–4.8) illustrate that even a few imputations ($m = 2$, 3, or 5) are enough to produce very reasonable coverage properties for interval estimates in many cases. The combination of estimation procedures that condition on global parameters with a multiple-imputation procedure to obtain inferences that average over those global parameters is a powerful technique that can be applied quite generally.

3. PERFORMANCE OF CALIBRATION PROCEDURE ON CENSUS COMPUTER-MATCHING DATA

3.1 Results From Test-Census Data

We use the test-census data described in Section 1.3 to illustrate the performance of the proposed calibration procedure, where determinations by clerks are the best measures available for judging true-match and false-match status. With three separate sites available, we were able to apply our strategy three times, with two sites serving as training data and the mixture-model procedure applied to the data from the third site.

We display results from each of the three tests in Figure 2. The dotted curve represents predicted false-match rates obtained from the mixture-model procedure, with accompanying 95% intervals represented by the dashed curves. Also plotted are the observed false-match rates, denoted by the "O" symbol, associated with each of several possible choices of cutoff values between matches and nonmatches.

We call attention to several features of these plots. First, it is clearly possible to match large proportions of the files with little or no error. Second, the quality of candidate matches becomes dramatically worse at some point where the false-match rate accelerates. Finally, the calibration procedure performs very well in all three tests from the standpoint of providing predictions that are close to the true values and interval estimates that include the true values.

In Figure 3 we take a magnifying glass to the previous displays to highlight the behavior of the calibration procedure in the region of interest where the false-match rate accelerates. That the predicted false-match rate curves bend upward close to the points where the observed false-match rate curves rise steeply is a particularly encouraging feature of the calibration method.

For comparison with the logistic-regression approach, we report in Table 2 (p. 704) the estimated false-match rates across the various sites for records with weights in the interval $[-5, 0]$, which in practice contains both true matches and false matches. Two alternative logistic regression models—one in which $\text{logit}(\eta)$ is modeled as a linear function of matching weight and the other in which $\text{logit}(\eta)$ is modeled as a quadratic function of matching weight, where η is the probability of false match—were fitted to data from two sites to predict false-match rates in the third site. A predictive standard error to reflect binomial sampling variability, as well as uncertainty in parameter estimation, was calculated using

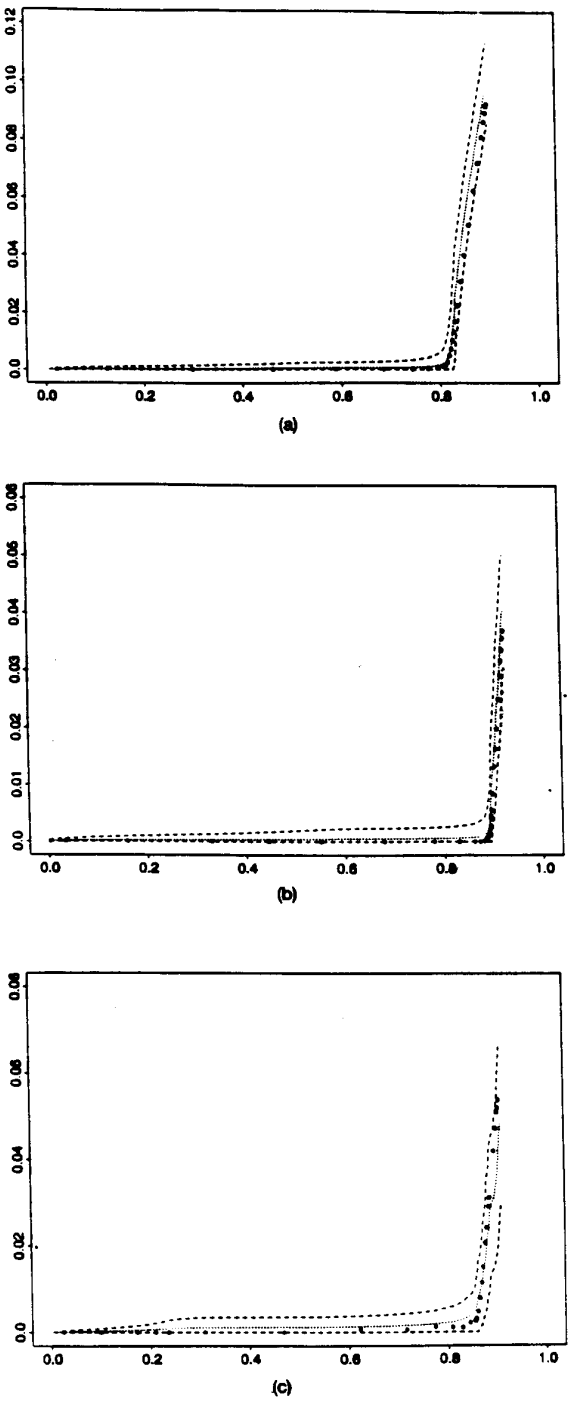


Figure 2. Performance of Calibration Procedure on Test-Census Data: (a) St. Louis, Using Columbia and Washington as Training Sample; (b) Columbia, Using St. Louis and Washington as Training Sample; (c) Washington, Using St. Louis and Columbia as Training Sample. O = observed false-match rate; · · · = predicted false-match rate; --- = upper and lower 95% bounds.

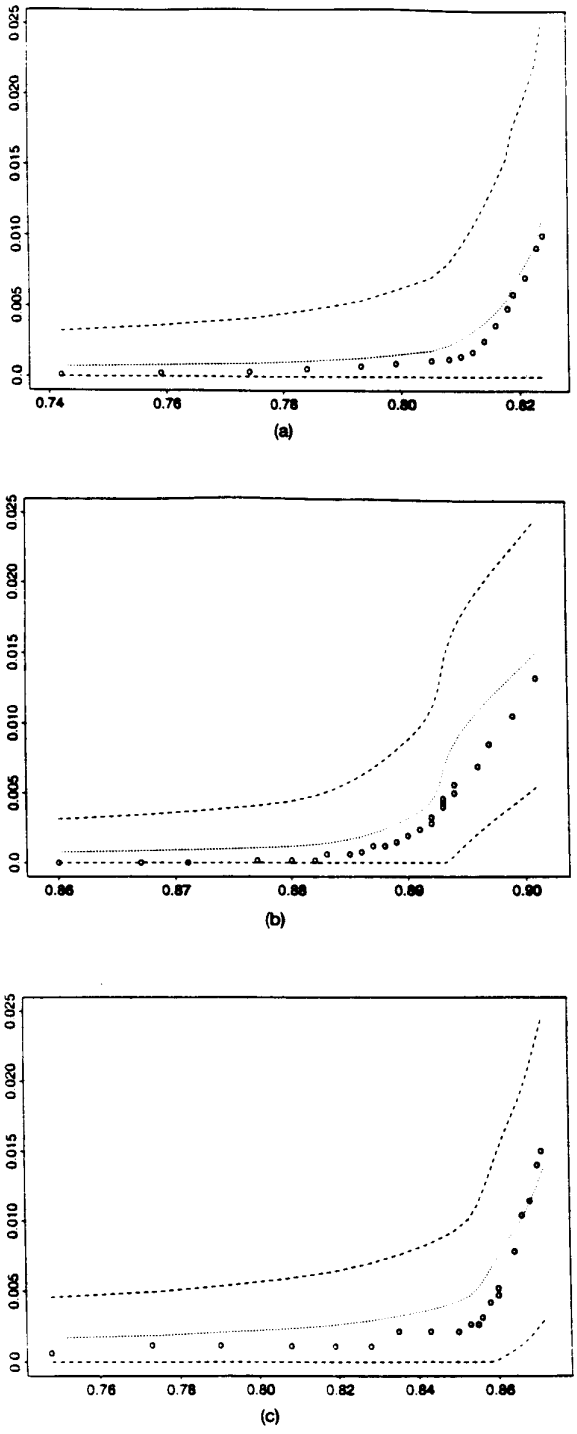


Figure 3. Performance of Calibration Procedure in Region of Interest: (a) St. Louis, Using Columbia and Washington as Training Sample; (b) Columbia, Using St. Louis and Washington as Training Sample; (c) Washington, Using St. Louis and Columbia as Training Sample. O = observed false-match rate; · · · = predicted false-match rate; --- = upper and lower 95% bounds.

$$\mathrm{SE} = \frac{1}{n^*}\sqrt{\sum_{i=1}^{n^*}\hat{\eta}_i(1-\hat{\eta}_i)[1+\mathbf{w}_i^T\mathrm{var}(\hat{\boldsymbol{\beta}})\mathbf{w}_i]},$$

where n^* is the number of record pairs with weights in the target interval and $\hat{\eta}_i$ is the predicted probability of false match for candidate pair i. For the linear logistic model, we have $\mathbf{w}_i = (1, w_i)^T$, where $\mathbf{w}_i$ is the weight for record pair i and $\mathrm{var}(\hat{\boldsymbol{\beta}})$ is the 2×2 covariance matrix of the regression parameters, whereas for the quadratic logistic model we have $\mathbf{w}_i = (1, w_i, w_i^2)^T$, where $\mathrm{var}(\hat{\boldsymbol{\beta}})$ is the 3×3 covariance matrix of the regression parameters.

As can be seen from Table 2, the predicted false-match rates from two alternative logistic regression models often were not in agreement with the observed false-match rates; in fact, they were often several standard errors apart. Because weights typically depend on site-specific data, this finding was not especially surprising. It is also noteworthy that the estimate of the quadratic term β_2 in the quadratic models was more than two standard errors from zero using the St. Louis data ($p = .029$) but was near zero using the Columbia and Washington data sets individually ($p = .928$ and $p = .719$). Using the mixture-model calibration approach, in two of the three sites the observed false-match rate is within two standard errors of the predicted false-match rate, and in the other site (St. Louis) the mixture-model approach is conservative in that it overstates the false-match rate. We regard this performance as superior to that of logistic regression—not surprising in light of our earlier discussion in Section 1.3 of why we eschewed logistic regression in this setting. Refining the mixture-model calibration approach through a more sophisticated prior distribution for global parameters (e.g., altering the prior specification so that urban sites are exchangeable with one another but not with rural sites) may result in even better performance by reflecting key distinctions in distributional shapes across sites.

3.2 A Limitation in the Extreme Tails

For small proportions of the records declared matched, counter-intuitive false-match rate estimates arise, with false-match rate estimates increasing as the proportion declared matched goes to zero. Such effects are explained by the false-match component being more variable than the true-match component, so that in the extreme upper tail of the component distributions the false-match component density is greater than the true-match component density. To avoid nonsensical results (since we know that the false-match rate should go to zero as the weight approaches the maximum possible weight), we find the points along the false-match-rate curve and the upper interval-estimate curve, if any, where the curves depart from a monotone decreasing pattern as the proportion declared matched approaches zero. From the point at which the monotone pattern stops, we linearly interpolate false-match rate estimates between that point and zero. We are not alarmed to find that the model does not fit well in the far reaches of the upper tails of component distributions, and other smoothing procedures may be preferred to the linear-interpolation procedure used here.

Table 2. Performance of Mixture-Model Calibration Procedure on Test-census Matching Weights in the Interval [−5, 0]

Site to be predicted	Observed false-match rate among cases with weights in [−5, 0]	Predicted false-match rate (SE) for cases with weights in [−5, 0] under linear logistic model; that is logit(η) = α + β (Wt)	Predicted false-match rate (SE) for cases with weights in [−5, 0] under quadratic logistic model; that is logit(η) = α + β_1 (Wt) + β_2 (Wt)^2	Predicted false-match rate (SE) for cases with weights in [−5, 0] based on mixture-model calibration method
St. Louis	.646 (= 73/113)	.417 (.045)	.429 (.045)	.852 (.033)
Columbia	.389 (= 14/36)	.584 (.079)	.613 (.079)	.524 (.083)
Washington	.294 (= 5/17)	.573 (.115)	.597 (.115)	.145 (.085)

4. SIMULATION EXPERIMENT TO INVESTIGATE PROPERTIES OF CALIBRATION PROCEDURE

4.1 Description of Simulation Study

Encouraged by the success of our method with real data, we conducted a large simulation experiment to enhance our understanding of the calibration method and to study statistical properties of the procedure (e.g., bias in estimates of the false-match rate, bias in estimates of the probability of false match, coverage of nominal 95% interval estimates for false-match rates, etc.) when the data generating mechanism was known. The simulation procedure involved generating data from two overlapping component distributions, with potential "site-to-site" variability from one weight distribution to another, to represent plausible weights, and replicating the calibration procedure on these samples.

Beta distributions were used to represent the components of the distribution of weights in the simulation experiment. Simulated weights thus were generated from component distributions that generally would be skewed and that have a functional form other than the transformed-normal distribution used in our procedure. The choice of beta-distributed components was convenient in that simple computational routines were available (Press, Flannery, Teukolsky, and Vetterling 1986) to generate beta-random deviates and to evaluate tail probabilities of beta densities.

Table 3 lists factors and outcome variables that were included in the experiment. Here we report only broad descriptive summaries from the simulation study. Greater detail on the design of the experiment and a discussion of the strategy for relating simulation outcomes to experimental factors have been provided by Belin (1991). The calibration procedure was replicated 6,000 times, with factors selected in a way that favored treatments that were less costly in terms of computer time (for details, again see Belin 1991).

4.2 Results

Figure 4 displays the average relative bias from the various simulation replicates in a histogram. Due to the way that we have defined relative bias (see Table 3), negative values correspond to situations where the predicted false-match rate is greater than the observed false-match rate; that is, negative relative bias corresponds to conservative false-match rate estimates. It is immediately apparent that the calibration procedure is on target in the vast majority of cases, with departures falling mostly on the "conservative" side, where the procedure overstates false-match rates and only a few cases where the procedure understates false-match rates.

The few cases in which the average relative bias was unusually large and negative were examined more closely, and all of these had one or more cutoffs where the observed false-match rate was zero and the expected false-match rate was small. In such instances the absolute errors are small, but relative errors can be very large. Clearly, however, errors between a predicted false-match rate of .001 or .002 and observed false-match rate of 0 presumably are not of great concern in applications.

There was a single replicate that had a positive average relative bias substantially larger than that of the other replicates. Further investigation found that a rare event occurred in that batch, with one of the eight highest-weighted records being a false match, which produced a very high relative error. In this replicate, where the predicted false-match rate under the mixture model was .001, the observed false-match rate was .125 and the expected (beta) false-match rate was .00357. Belin (1991) reported that other percentiles of the distribution in this replicate were fairly well calibrated (e.g., predicted false-match rates of .005, .01, .10, .50, and .90 corresponded to expected beta false-match rates of .007, .011, .130, .455, and .896); thus it was apparently a rare event rather than a breakdown in the calibration method that led to the unusual observation.

With respect to the coverage of interval estimates, we focus on simulation results when the SEM algorithm was applied to calculate a distinct covariance matrix for each fitted mixture model (n = 518). In the other simulation replicates, shortcuts in standard error calculations were taken so as to avoid computation; Belin (1991) reported that these shortcut methods performed moderately worse in terms of coverage. For nominal two-sided 95% intervals, the calculated intervals covered observed false-match rates 88.58% of the time (SE = 1.94%); for nominal one-sided 97.5% intervals, the calculated intervals covered observed false-match rates 97.27% of the time (SE = 1.45%). Thus the calibration method does not achieve 95% coverage for two-sided interval estimates, but when it errs it tends to err on the side of *overstating* false-match rates, so that the one-sided interval estimates perform well.

Table 3. Description of Factors and Outcomes in Simulation Experiment

Experimental factors	Comments
1. Number of sources of training data	1. Possible values = {3, 4, . . . , 30}
2. Sizes of training samples	2–3. Possible values = {2,000, 2,001, . . . , 9,999}
3. Size of current data base	4. Possible values in [.01, .5]
4. Mixing proportion between false-match and true-match components	5. $Be(a_F, b_F)$, $a_F \sim U(1.75, 2)$ $b_F \sim U(3, 6)$
5. Shape of the false-match component	6. $Be(a_T, b_T)$, $a_T \sim U(8, 12)$ $b_T \sim U(1.75, 2)$
6. Shape of true-match component	7–8. Separate draws from 5–6 for separate sites, or common shapes across sites
7. Amount of site-to-site variability in mixing proportion	9. Perform SEM for each mixture model being fit, or a short-cut method to save computing time based on approximations
8. Amount of site-to-site variability in shapes of component distributions	10. Possible values = {3, 5, 10}
9. Method for calculating standard errrors	
10. Number of mixture models fit	

Outcomes measured in simulation	Comments
1. Average relative bias in false-match rate estimates across 20 prespecified predicted false-match rates	1. Prespecified false-match rates = {.001, .002, . . . , .015, .02, .025, .03, .04, .05} Relative bias = [(observed false-match rate) − (predicted false-match rate)]/√(expected false-match rate) "predicted false-match rate" = estimated false-match rate calculated under transformed-normal mixture model "observed false-match rate" = {# false matches}/{# declared matches} at given cutoff "expected false-match rate" = {tail area of Beta false-match component}/{sum of tail areas of Beta component distributions}
2. Average two-sided coverage rate = average of 20 indicators for whether interval estimates of false-match rate covered observed false-match rate	2–3. Same prespecified false-match rates as in 1.
3. Average one-sided coverage rate = average of 20 indicators for whether interval estimates of false-match rate covered or overstated false-match rate	4. Estimated probabilities of false match = {.00125, .0025, .005, .01, .02, .10, .25, .50, .75, .90}
4. Expected probability of false match for 10 predicted probabilities of false match	

Turning to the performance of the estimated probabilities of false match (i.e., neighborhood false-match rates) obtained from the fitted mixture models, Table 4 provides the mean, standard deviation, minimum, and maximum of the true underlying probabilities being estimated by the calibration procedure. Although in specific sites the calibration procedure substantially understates or overstates false-match rates, the procedure appears to have good properties in the aggregate.

5. DISCUSSION

Previous attempts at estimating false-match rates in record linkage were either unreliable or too cumbersome for practical use. Although our method involves fitting nonstandard models, other researchers have used software that we developed to implement the technique in at least two other settings (Ishwaran, Berry, Duan, and Kanouse 1991; Scheuren and Winkler 1991, 1993). This software is available on request from the first author.

Analyses by Belin (1991) have revealed that the deficiencies in the calibration procedure typically occurred where the split in the proportion of records between the two components was very extreme. For example, after excluding a few dozen replicates where 99% or more of the records were declared matched above the point where the procedure predicted a false-match rate of .005, there was no evidence that sample sizes of the data bases being matched had an impact on the accuracy of estimated probabilities of false match, implying that breakdown of the calibration procedure appears to be a threshold phenomenon.

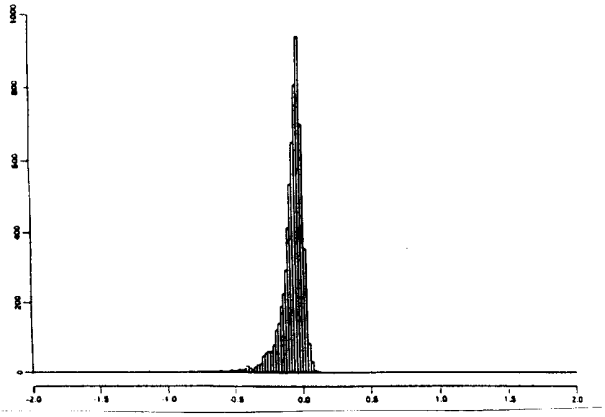


Figure 4. Histogram of Average Relative Bias Across Simulation Replicates.

Table 4. Performance of Estimated Probabilities of False-Match in Predicting True Underlying Probabilities

Estimated probabilty	Mean of actual probabilities	Std. deviation of actual probabilities	Minimum of actual probabilities	Maximum of actual probabilities
.00125	.00045	.00075	.00000	.0105
.0025	.0012	.00143	.00000	.0170
.005	.0030	.00266	.00000	.0295
.01	.0072	.00470	.00002	.059
.02	.0160	.00818	.00019	.1183
.10	.0813	.03231	.00155	.4980
.25	.2086	.07385	.00549	.8094
.50	.4555	.11996	.02812	.9551
.75	.7459	.10768	.20013	.9988
.90	.9244	.05222	.56845	1.000

Finally, on several occasions when we have discussed these techniques and associated supporting evidence, we have been questioned about the validity of using determinations of clerks as a proxy for true-match or false-match status. Beyond pointing to the results of the simulations, we note that (a) clerical review is as close to truth as one is likely to get in many applied contexts, and (b) possible inaccuracy in assignment of match status by clerks is no criticism of the calibration procedure. This methodology provides a way to calibrate false-match rates to whatever approach is used to identify truth and falsehood in the training sample and appears to be a novel technique that is useful in applied contexts.

APPENDIX: IMPLEMENTATION OF THE SEM ALGORITHM

The SEM algorithm is founded on the identity that the observed-data observed information matrix $\mathbf{I}_{obs}$ for a k-dimensional parameter θ can be expressed in terms of the conditional expectation of the "complete-data" observed information matrix evaluated at the MLE ($\mathbf{I}_{com}$) as

$$\mathbf{I}_{obs} = (\mathbf{I} - \mathbf{DM})\mathbf{I}_{com},$$

where $\mathbf{I}$ is the $k \times k$ identity matrix and $\mathbf{DM}$ is the Jacobian of the mapping defined by EM (i.e., of the mapping that updates parameter values on the tth iteration to those on the $(t+1)$st iteration) evaluated at the MLE $\hat{\theta}$. Taking inverses of both sides of the previous equation yields the identity

$$\mathbf{V}_{obs} = \mathbf{I}_{obs}^{-1} = \mathbf{V}_{com} + \Delta\mathbf{V},$$

where $\mathbf{V}_{com} = \mathbf{I}_{com}^{-1}$ and $\Delta\mathbf{V} = \mathbf{V}_{com}\mathbf{DM}(\mathbf{I} - \mathbf{DM})^{-1}$, the latter reflecting the increase in variance due to the missing information.

The SEM procedure attempts to evaluate the $\mathbf{DM}$ matrix of partial derivatives numerically. First, EM is run to convergence and the MLE $\hat{\theta}$ is obtained. To compute partial derivatives, all but one of the components of the parameter are fixed at their MLE's, and the remaining component is set to its value at the tth iteration, say $\theta^{(t)}(i)$. Then, after taking a "forced EM step" by using this parameter value as a start to a single iteration (E step and M step), the new estimates, say $\tilde{\theta}_j^{(t+1)}(i)$ for $j = 1, \ldots, k$, yield the following estimates of partial derivatives:

$$r_{ij}^{(t)} = \frac{\tilde{\theta}_j^{(t+1)}(i) - \hat{\theta}_j}{\theta_i^{(t)}(i) - \hat{\theta}_i}.$$

It is necessary to perform k forced EM steps at every iteration of SEM—although Meng and Rubin (1991) pointed out that once convergence is reached for each component of the vector $(r_{i'1}, r_{i'2}, \ldots, r_{i'k})$, it is no longer necessary to perform the forced EM step for component $i = i'$.

Because we regard the variance ratio as fixed when fitting our mixture models, we are actually estimating four parameters in the calibration mixture-model setting: the locations of the two components, one unknown scale parameter, and a mixing parameter. We can calculate the complete-data information matrix for $(\mu_F, \mu_T, \sigma_F^2, \lambda)$ as

$$\mathbf{A} = \begin{pmatrix} a_{11} & 0 & a_{13} & 0 \\ 0 & a_{22} & a_{23} & 0 \\ a_{31} & a_{32} & a_{33} & 0 \\ 0 & 0 & 0 & a_{44} \end{pmatrix},$$

where

$$a_{11} = \frac{\sum_{i=1}^{n} Z_i}{\sigma_F^2} \qquad a_{22} = \frac{V\sum_{i=1}^{n}(1 - Z_i)}{\sigma_F^2} \qquad a_{33} = \frac{\sum_{i=1}^{n} Z_i}{\sigma_F^6}$$

$$a_{44} = \frac{\sum_{i=1}^{n} Z_i}{\lambda^2} + \frac{\sum_{i=1}^{n}(1 - Z_i)}{(1-\lambda)^2}$$

$$a_{31} = a_{13} = \frac{\sum_{i=1}^{n} Z_i(w_i - \mu_F)}{\sigma_F^4}$$

$$a_{32} = a_{23} = \frac{V\sum_{i=1}^{n}(1 - Z_i)(w_i - \mu_T)}{\sigma_F^4}.$$

The missing information in our problem arises from the fact that the Z_i's are unknown.

Because the covariance matrix is 4×4, every iteration of the SEM algorithm takes roughly four times as long as an iteration of EM. It also should be pointed out that the SEM algorithm relies on precise calculation of MLE's. Although it may only be necessary to run EM for 10 or 20 iterations to obtain accuracy to two decimal places in MLE's, it might take 100 or more iterations to obtain accuracy to, say, six decimal places. These aspects of the SEM algorithm can make it computationally expensive.

The $\mathbf{DM}$ matrix containing the r_{ij}'s will not generally be symmetric, but of course the resulting covariance matrix should be symmetric. If the resulting covariance matrix is not symmetric even though several digits of numerical precision are obtained for the MLE and the r_{ij}'s, this reflects an error in the computer code used to implement the forced SEM steps or perhaps in the code for the E step and M step themselves. The symmetry or lack thereof in the resulting covariance matrix thus provides a diagnostic check on the correctness of the program.

Experience with the SEM algorithm suggests that convergence of the numerical approximations to the partial derivatives of the mapping often occurs in the first few iterations and further reveals that beyond a certain number of iterations, the approach can give nonsensical results owing to limitations in numerical precision, just as with any numerical differentiation procedure. Meng and Rubin (1991) suggested specifying a convergence criterion for the r_{ij}'s and ceasing to calculate these terms once the criterion is satisfied for all $j = 1, \ldots, k$. An alternative (used in producing the results in Secs. 3 and 4) involves running the SEM algorithm for eight iterations, estimating all partial derivatives of the mapping on each iteration, assessing which two of the eight partial derivative estimates are closest to one another, and taking the second of the two as our estimate of the derivative. Experience with this approach suggests that it yields acceptable results for practice in that the off-diagonal elements of the resulting covariance matrix agree with one another to a few decimal places.

[Received February 1993. Revised November 1993.]

REFERENCES

Aitkin, M., and Rubin, D. B. (1985), "Estimation and Hypothesis Testing in Finite Mixture Models," *Journal of the Royal Statistical Society*, Ser. B, 47, 67–75.

Belin, T. R. (1989a), "Outline of Procedure for Evaluating Computer Matching in a Factorial Experiment," unpublished memorandum, U.S. Bureau of the Census, Statistical Research Division.

——— (1989b), "Results from Evaluation of Computer Matching," unpublished memorandum, U.S. Bureau of the Census, Statistical Research Division.

——— (1990), "A Proposed Improvement in Computer Matching Techniques," in *Statistics of Income and Related Administrative Record Research: 1988–1989*, Washington, DC: U.S. Internal Revenue Service, pp. 167–172.

——— (1991), "Using Mixture Models to Calibrate Error Rates in Record-Linkage Procedures, with Application to Computer Matching for Census Undercount Estimation," Ph.D. thesis, Harvard University, Dept. of Statistics.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 26, 206–252.

Brown, P., Laplant, W., Lynch, M., Odell, S., Thibaudeau, Y., and Winkler, W. (1988), *Collective Documentation for the 1988 PES Computer-Match Processing and Printing*, Vols. I–III, Washington, DC: U.S. Bureau of the Census, Statistical Research Division.

Copas, J., and Hilton, F. (1990), "Record Linkage: Statistical Models for Matching Computer Records," *Journal of the Royal Statistical Society*, Ser. A, 153, 287–320.

Dawid, A. P. (1976), "Properties of Diagnostic Data Distributions," *Biometrics*, 32, 647–658.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Efron, B. (1975), "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," *Journal of the American Statistical Association*, 70, 892–898.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183–1210.

Halperin, M., Blackwelder, W. C., and Verter, J. I. (1971), "Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches," *Journal of Chronic Diseases*, 24, 125–158.

Hogan, H. (1992), "The 1990 Post-Enumeration Survey: An Overview," *The American Statistician*, 46, 261–269.

Ishwaran, H., Berry, S., Duan, N., and Kanouse, D. (1991), "Replicate Interviews in the Los Angeles Women's Health Risk Study: Searching for the Three-Faced Eve," technical report, RAND Corporation.

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 84, 414–420.

Kelley, R. P. (1986), "Robustness of the Census Bureau's Record Linkage System," in *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 620–624.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley.

Maclean, C. J., Morton, N. E., Elston, R. C., and Yee, S. (1976), "Skewness in Commingled Distributions," *Biometrics*, 32, 695–699.

Mantel, N., and Brown, C. (1974), "Alternative Tests for Comparing Normal Distribution Parameters Based on Logistic Regression," *Biometrics*, 30, 485–497.

Meng, X. L., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance–Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, 86, 899–909.

Newcombe, H. B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford, U.K.: Oxford University Press.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science*, 130, 954–959.

Orchard, T., and Woodbury, M. A. (1972), "A Missing Information Principle: Theory and Applications," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 697–715.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1986), *Numerical Recipes: The Art of Scientific Computing*, Cambridge, U.K.: Cambridge University Press.

Rogot, E., Sorlie, P. D., and Johnson, N. J. (1986), "Probabilistic Methods in Matching Census Samples to the National Death Index," *Journal of Chronic Diseases*, 39, 719–734.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

Scheuren, F., and Winkler, W. E. (1991), "An Error Model for Regression Analysis of Data Files That are Computer Matched," in *Proceedings of the 1991 Annual Research Conference, U.S. Bureau of the Census*, pp. 669–687.

——— (1993). "Regression Analysis of data files that are computer matched," *Siney Methodology*, 19, 39–58.

Tepping, B. J. (1968), "A Model for Optimum Linkage of Records," *Journal of the American Statistical Association*, 63, 1321–1332.

Thibaudeau, Y. (1989), "Fitting Log-Linear Models in Computer Matching," in *Proceedings of the American Statistical Association, Section on Statistical Computing*, pp. 283–288.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: John Wiley.

Weisberg, S. (1980), *Applied Linear Regression*, New York: John Wiley.

Winkler, W. E. (1989), "Near-Automatic Weight Computation in the Fellegi–Sunter Model of Record Linkage," in *Proceedings of the Fifth Annual Research Conference, U.S. Bureau of the Census*, pp. 145–155.

Chapter 4

Modeling Issues and the Use of Experience in Record Linkage

Michael D. Larsen, Harvard University

Abstract

The goal of record linkage is to link quickly and accurately records corresponding to the same person or entity. Fellegi and Sunter (1969) proposed a statistical model for record linkage that assumes pairs of entries, one from each of two files, either are matches corresponding to a single person or nonmatches arising from two different people. Certain patterns of agreements and disagreements on variables in the two files are more likely among matches than among nonmatches. The observed patterns can be viewed as arising from a mixture distribution.

Mixture models, which for discrete data are generalizations of latent-class models, can be fit to comparison patterns in order to find matching and nonmatching pairs of records. Mixture models, when used with data from the U.S. Decennial Census — Post Enumeration Survey, quickly give accurate results.

A critical issue in new record-linkage problems is determining when the mixture models consistently identify matches and nonmatches, rather than some other division of the pairs of records. A method that uses information based on experience, identifies records to review, and incorporates clerically-reviewed data is proposed.

Introduction

Record linkage entails comparing records in one or more data files and can be implemented for unduplication or to enable analyses of relationships between variables in two or more files. Candidate records being compared really arise from a single person or from two different individuals. Administrative data bases are large and clerical review to find matching and nonmatching pairs is expensive in terms of human resources, money, and time. Automated linkage involves using computers to perform matching operations quickly and accurately.

Mixture models can be used when the population is composed of underlying and possibly unidentified subpopulations. The clerks manually identify matches and nonmatches, while mixture models can be fit to unreviewed data in the hopes of finding the same groups. However, mixture models applied to some variables can produce groups that fit the data but do not correspond to the desired divisions. A critical issue in this application is determining when the model actually is identifying matches and nonmatches.

A procedure is proposed in this paper that when applied to Census data seems to work well. The more that is known about a particular record linkage application, the better the procedure should work. The size of the two files being matched, the quality of the information recorded in the two files, and any clerical review that has already been completed are incorporated into the procedure. Additionally, the procedure

should help clerks be more efficient because it can direct their efforts and increase the value of reviewed data through use in the model.

The paper defines mixture models and discusses estimation of parameters, clustering, and error rates. Then previous theoretical work on record linkage is described. Next, the paper explains the proposed procedure. A summary of the application of the procedure to five Census data sets is given. The paper concludes with a brief summary of results and reiteration of goals.

Mixture Models

An observation $\mathbf{y}_i$ (possibly multivariate) arising from a finite mixture distribution with G classes has probability density

$$p(\mathbf{y}_i \mid \Pi, \Theta) = \Sigma_{g=1,G}\, \pi_g\, p_g(\mathbf{y}_i \mid \theta_g), \qquad (1)$$

where π_g $(\Sigma_{g=1,G}\, \pi_g = 1)$, p_g, and θ_g are the proportion, the density of observations, and the distributional parameters, respectively, in class g, and Π and Θ are abbreviated notation for the collections of proportions and parameters, respectively. The likelihood for π and θ based on a set of n observations is a product with index i=1,...,n of formula (1).

The variables considered in this paper are dichotomous and define a table of counts, which can have its cells indexed by i. In the application, each observation is ten dimensional, so n=1024. The mixture classes are in effect subtables, which when combined yield the observed table. The density $p_g(\bullet \mid \bullet)$ in mixture class g can be defined by a log-linear model on the expected counts in the cells of the subtable. The relationship among variables described by the log linear model can be the same or different in the various classes. When the variables defining the table in all classes are independent conditional on the class, the model is the traditional latent-class model. Sources for latent-class models include Goodman (1974) and Haberman (1974, 1979).

Maximum likelihood estimates of π and θ can be obtained using the EM (Dempster, Laird, Rubin 1977) and ECM (Meng and Rubin, 1993) algorithms. The ECM algorithm is needed when the log linear model in one or more of the classes can not be fit in closed form, but has to be estimated using iterative proportional fitting.

The algorithms treat estimation as a missing data problem. The unobserved data are the counts in each pattern in each class and can be represented by a matrix $\mathbf{z}$ with n rows and G columns, where entry z_{ig} is the number of observations with pattern i in class g. If the latent counts were known, the density would be

$$p(\mathbf{y}, \mathbf{z} \mid \Pi, \Theta) = \Pi_{i=1,n}\, \Pi_{g=1,G}\, (\pi_g\, p_g(\mathbf{y}_i \mid \theta_g))^{z_{ig}}. \qquad (2)$$

Classified data can be used along with unclassified data in algorithms for estimating parameters. The density then is a combination of formulas (2) and a product over i of (1). Known matches and nonmatches, either from a previous similar matching problem or from clerk-reviewed data in a new problem, can be very valuable since subtables tend to be similar to the classified data.

Probabilities of group membership for unclassified data can be computed using Bayes' Theorem. For the k^{th} observation in the i^{th} cell, the probability of being in class g $(z_{igk}=1)$ is

$$p(z_{igk} = 1 \mid \mathbf{y}_i, \pi, \theta) = \pi_g\, p_g(\mathbf{y}_i \mid \theta_g) \,/\, \Pi_{h=1,G}\, \pi_h\, p_h(\mathbf{y}_i \mid \theta_h)\ . \qquad (3)$$

Probability (3) is the same for all observations in cell i. Probabilities of class membership relate to probabili-

ties of being a match and nonmatch only to the degree that mixture classes are similar to matches and nonmatches.

The probabilities of group membership can be used to cluster the cells of the table by sorting the cells of the table according to descending probability of membership in a selected class. An estimated error rate at a given probability cut-off is obtained by dividing the expected number of observations not in a class by the total number of observations assigned to a class. As an error rate is reduced by assigning fewer cells to a class, the number of observations in a nebulous group not assigned to a class increases.

Before the match and nonmatch status is determined by clerks tentative declarations as probable match and probable nonmatch can be made using mixture models. It is necessary to choose a class or classes to be used as probable matches and probable nonmatches, which usually can be done by looking a probabilities of agreement on fields in the mixture classes. The estimated error rates from the mixture model correspond to the actual rate of misclassification of matches and nonmatches only to the degree that the mixture classes correspond to match and nonmatch groups.

Record Linkage Theory

Fellegi and Sunter (1969) proposed a statistical model for record linkage that assumes pairs of entries, one from each of two files, either are matches corresponding to a single person or nonmatches arising from two different people. Patterns of agreements and disagreements on variables have probabilities of occurring among matches and among nonmatches. If the pairs of records are ordered according to the likelihood ratio for being a match versus being a nonmatch and two cut-off points are chosen, one above which pairs are declared matches and one below which pairs are declared nonmatches, the procedure is optimal in the sense of minimizing the size of the undeclared set at given error levels.

Fellegi and Sunter (1969) suggested methods to estimate the unknown probabilities involved in the likelihood ratio. Some of their simplifying assumptions, such as the independence of agreement on fields of information within matches and nonmatches, have continued to be used extensively in applications.

In the methods proposed in this paper, the likelihood ratio is estimated using the mixture model. If the first class is the class of probable matches, the likelihood ratio is for pattern i is

$$p(g=1 \mid y_i, \Pi, \Theta)/ p(g\neq 1 \mid \mathbf{y}_i, \Pi, \Theta) = \pi_1 p_1(\mathbf{y}_i \mid \theta_1) / \Sigma_{g=2,G} \pi_g p_g(\mathbf{y}_i \mid \theta_g). \quad (4)$$

The success depends on the relationship between the implied latent groups and the match and nonmatch categories.

The choice of cutoff values for declaring observations matches and nonmatches is critical, as demonstrated by Belin (1993). Belin and Rubin (1995) have shown that previous applications of the Fellegi-Sunter procedure do not always have their specified error levels. In applications, the cutoff values often are determined by manual review of observations in the "gray area" or likely to be sent to clerical review.

In the current paper, a cutoff can be chosen using mixture model results to achieve a specified error level, but the actual error level might or might not be close to the estimated level.

Winkler (1988, 1989a, 1989b, 1989c, 1990, 1992, 1993, 1994) and Thibaudeau (1989, 1993) have used mixture models of the type used in this article in record-linkage applications at the Census Bureau. The new procedure in this article addresses the critical question of when a mixture-model approach is appropriate for a new record-linkage situation.

Belin and Rubin (1995) developed a procedure for estimating error rates in some situations using the Fellegi-Sunter algorithm and applied it to Census data. Their method needs clerically-reviewed training data from similar record linkages and works well when the distributions of likelihood values (4) for matches and nonmatches are well separated. The new approach of this paper does not require training data, but could use it as classified observations, and provides its own estimates of error rates as described for mixture models.

Other applications of record linkage have used more information, such as frequency of names and string comparator metrics, than simple binary agree/disagree comparisons of fields. While there obviously is value in more detailed comparisons, this paper uses only multivariate dichotomous data and leaves development of model-based alternatives to current methods for more complicated data to later.

Procedure

The procedure for applying mixture models to record-linkage vector comparisons is specified below. It has many informal aspects some of which correspond to decisions often made in practical applications. Later work will investigate formalizing the procedure.

- Fit a collection of mixture models that have been successful in previous similar record-linkage problems to the data.
- Select a model with a class having (a) high probabilities of agreement on important fields, (b) probability of class membership near the expected percent of pairs that are matches, and (c) probabilities of class membership for individual comparison patterns near 0 or 1.
- Identify a set of records for clerks to review using the mixture model results.
- Refit the mixture model using both the classified and unclassified data.
- Cycle through the two previous steps as money and time allow, or until satisfied with results.

Models that can be used in step (1) are illustrated below. Some searching through other model possibilities might have to be done. In step (2), from the observed, unclassified data, it is possible to compute the probability of agreement on fields and combinations of fields. The probabilities should be higher in the probable match class than overall. The percent of pairs that are matches is limited by the size of the smaller of the two lists contributing to candidate pairs. If a class is much larger than the size of the smaller list, it must contain several nonmatches. Of course no single model may be clearly preferable given the informal statement of criteria.

In step (3), records to review can be identified by first accumulating pairs into the probable match class according to probability of membership until a certain point and then reviewing pairs at the chosen boundary. The boundary used in this paper is the minimum of the estimated proportion in the probable mixture class and the size of the small list divided by the total number of pairs.

In the next section, the procedure is applied to five Census data sets and produces good results. Many aspects of the procedure parallel successful applications of the Fellegi-Sunter approach to record linkage. Different mixture models give slightly different estimates of the likelihood ratio just as different estimation methods currently used in practice lead to different orderings of pairs.

Application

In 1988, a trial census and post-enumeration survey (PES) were conducted. Data from two urban

sites are referred to as D88a and D88b in this section. In 1990, following the census, a PES was conducted in three locations. D90a and D90b are the data sets from urban sites, and D90c are data from a rural site. Table 1 contains summaries of the five data sets. Not all possible pairs of records from the census and PES were compared. Candidate match pairs had to agree on a minimal set of characteristics. Sites vary in size, proportion that are matches, and probabilities of agreeing on fields. D90c is rural, and its address information is not very precise. Thus, relatively more pairs are compared, yielding lower probabilities of agreement and a lower proportion of matches.

Table 1. -- Summary of Five Census/Post-Enumeration Survey Data Sets, Including Probabilities of Agreements on Fields Overall (and for Matches)

Data set	D88a	D88b	D90a	D90b	D90c
Census size	12072	9794	5022	4539	2414
PES size	15048	7649	5212	4859	4187
Total pairs	116305	56773	37327	38795	39214
Matches	11092	6878	3596	3488	1261
Nonmatch	105213	49895	33731	35307	37953
Last name	.32(.98)	.41(.99)	.31(.98)	.29(.98)	.26(.98)
First name	.11(.95)	.14(.98)	.12(.96)	.11(.95)	.06(.95
House #	.28(.97)	.18(.50)	.30(.95)	.27(.94)	.06(.42)
Street	.60(.96)	.28(.49)	.37(.67)	.59(.95)	.11(.44)
Phone #	.19(.71)	.31(.83)	.19(.69)	.18(.66)	.06(.45)
Age	.16(.85)	.23(.94)	.19(.89)	.17(.88)	.11(.89)
Relation to head of household	.19(.48)	.20(.54)	.16(.46)	.19(.48)	.25(.56)
Martial status	.41(.84)	.44(.89)	.36(.78)	.42(.85)	.42(.88)
Sex	.53(.96)	.53(.98)	.52(.97)	.52(.96)	.50(.96)
Race	.91(.97)	.93(.98)	.80(.93)	.83(.91)	.80(.86)

Mixture models considered in this application have either two or three classes. Models for the variables within each class include either main effects only, all two-way interactions, all three-way interactions, a five-way interaction between typically household variables (last name, house number, street name, phone number, and race) and a five-way interaction between typically personal variables (the other five), and a set of interactions described by Armstrong and Mayda (1993). The actual models are described in Table 2.

Table 2. -- Mixture Models Considered for Each Data Set

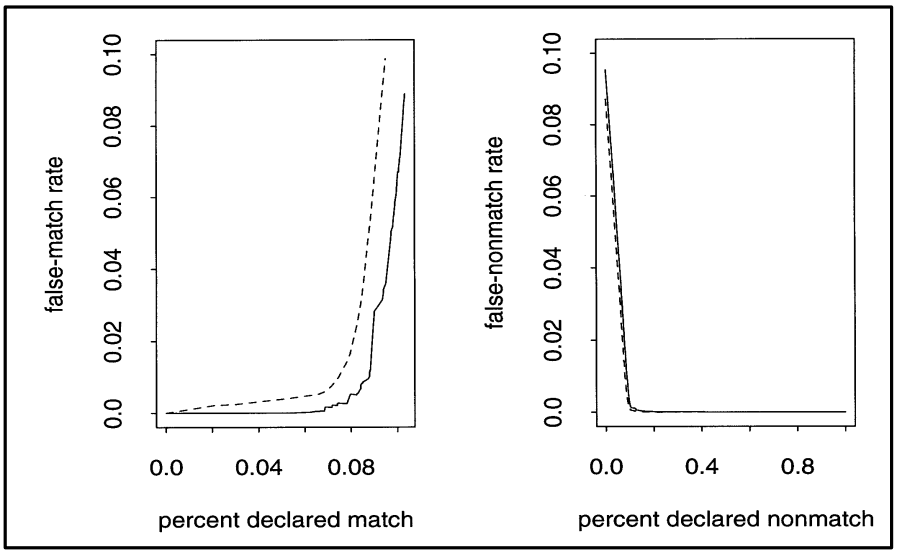
Abbreviation	Model Class 1	Model Class 2	Model Class 3
2C CI	Independent	Independent	
2C CI-2way	Independent	2way interactions	
2C CI-3way	Independent	3way interactions	
3C CI	Independent	Independent	Independent
3C CI-2way	Independent	2way interactions	2way interactions

2C CI-HP	Independent	5way interactions (Household, Personal)	
2C HP	5way interactions	5way interactions	
3C CI-HP	Independent	5way interactions	5way interactions
2C AM	Independent	Armstrong and Mayda, 1993	
3C AM	Independent	Armstrong and Mayda, 1993	Armstrong and Mayda, 1993

To illustrate the results from fitting a mixture model, the three-class conditional independence model (model 4) was fit to D88a. Figure 1 contains plots of the estimated and actual false-match and false-nonmatch rate. At an error rate of .005, using the estimated false-match curve, 7462 matches and 3 nonmatches are declared matches, giving an actual error rate of .0004. At an estimated error rate of .01, 8596 matches and 23 nonmatches are declared matches, giving an actual error rate of .0027.

Figure 1. -- False-Match and False-Nonmatch Rates From Fitting a Three-Class Conditional Independence Mixture to D88a

(The solid lines are actual and the dashed lines are estimated error rates)



The three-class conditional independence model works for the D88a data, because one of the classes tends to consist of pairs that agree on most comparisons. The medium-sized mixture class tends to agree on variables defining households, but to disagree on personal variables. This class can be called the same household-different person class. The third class tends to disagree on all comparisons. The three-class conditional-independence model also produces good results for the other data sets, except for D90c data. The difference could be caused by the fact that D90c is from a rural area, while the others have a lot of the population in urban settings with better household identifiers.

The search procedure was applied to each of the five data sets. The models selected to start with are given in the second row of Table 3. The number of matches and nonmatches declared matches with estimated false-match rates .005 and .01 are given in lines three and four of Table 3. The number of matches and nonmatches declared

nonmatches with estimated false-nonmatch rates .001 and .005 are given in the next two lines. Most models are successfully separating matches and nonmatches. However, in some cases, the rapid rise of estimated false-match rates means few observations can safely be declared matches.

Table 3. -- Initial Model Selected for Each Data Set, Along with Matches and Nonmatches Declared Matches and Nonmatches for Two False-Match (FMR) and False-Nonmatch (FNMR) Rates

Parentheses enclose (match, nonmatch) counts

Data set	**D88a**	**D88b**	**D90a**	**D90b**	**D90c**
Model	3C CI	2C CI-2way	3C CI	3C CI	3C CI-HP
.005 FMR	(7442,2)	(0,0)	(2802,27)	(2421,12)	(766,99)
.01 FMR	(8596,23)	(24,0)	(3083,50)	(2812,25)	(997,112)
.001 FNMR	(260, 104587)	(3455, 49855)	(124, 33244)	(69, 34507)	(32, 36900)
.005 FNMR	(1021, 105117)	(3858, 49882)	(248, 33469)	(234, 35126)	(61, 37571)
Total Counts	(11092, 105213)	(6878, 49895)	(3596, 33731)	(3488, 35307)	(1261, 37953)

The models used for D88a, D88b, and D90c were clearly the best candidates among the proposed models for trying to identify matches. In the cases of D90a and D90b, the model with two classes, one with conditional independence between the variables and the other with all two-way interactions, were close competitors to the three-class conditional-independence model. The models chosen for D90a and D90b had estimated error rates that grew slowly until approximately the proportion in the smallest class. The models not chosen had rapidly rising estimated error rates right away.

Pairs were identified to be reviewed by clerks. For the data set D88a, 5000 pairs were reviewed and error rates reestimated. 1000 pairs were reviewed and then the model was refit 5 times. Then 5000 more pairs were reviewed, 1000 at a time. Table 4 contains results for all 5 data sets. For the smaller data sets, fewer observations were reviewed. Note that in Table 4, the reported estimated false-match rates have been reduced. After about ten percent of the pairs are reviewed, most of the matches and nonmatches can be identified with few errors.

Table 4. -- Matches and Nonmatches Declared Matches and Nonmatches for Two False-Match (FMR) and False-Nonmatch Rates (FNMR) After Reviewing Some Records and Refitting Models

Parentheses enclose (match, nonmatch) counts

Data set	**D88a**	**D88b**	**D90a**	**D90b**	**D90c**
Model	3C CI	2C CI-2way	3C CI	3C CI	3C CI-HP
Reviewed	5000	2500	2000	2000	2000
.001 FMR	(10764, 0)	(2703, 1)	(2620, 10)	(2562, 8)	(48, 2)
.005 FMR	(10917, 27)	(3105, 8)	(3447, 26)	(3347, 17)	(393,5)

.001 FNMR	(58, 102728)	(3339, 49694)	(104, 33657)	(76, 35227)	(40, 37633)
.005 FNMR	(255 , 105212)	(3448, 49866)	(316, 33718)	(206, 35298)	(121, 37863)
Reviewed	10000	5000	4000	4000	4000
.001 FMR	(10916, 13)	(5057, 1)	(3439, 1)	(3341, 3)	(1019, 5)
.005 FMR	(10917, 27)	(6479, 17)	(3456, 18)	(3352, 9)	(1217, 5)
.001 FNMR	(58, 102728)	(246, 49857)	(106, 33688)	(76, 35236)	(32, 37994)
.005 FNMR	(255, 105212)	(433, 49881)	(194, 33731)	(206, 35307)	(186, 37948)
Total counts	(11092, 105213)	(6878, 49895)	(3596, 33731)	(3488, 35307)	(1261, 37953)

Figure 2 (on the next page) illustrates the impact of the addition of clerk-reviewed data on false-match rate estimates for data set D90c. The method performs better on the other data sets with their models than on D90c.

Conclusion

The development of theory related to applications can be useful for several reasons. The mixture-modeling approach of this paper hopefully can provide some insight into adjustments that are made in applications to make current theory work. Aspects of the new procedure with models parallel actual practice without models. The modeling approach also could improve efficiency by helping clerks identify valuable records to review and then using the additional information through the model to learn more about unclassified observations. More formal model selection procedures and models that allow more complex comparison data will increase the usefulness of the theory.

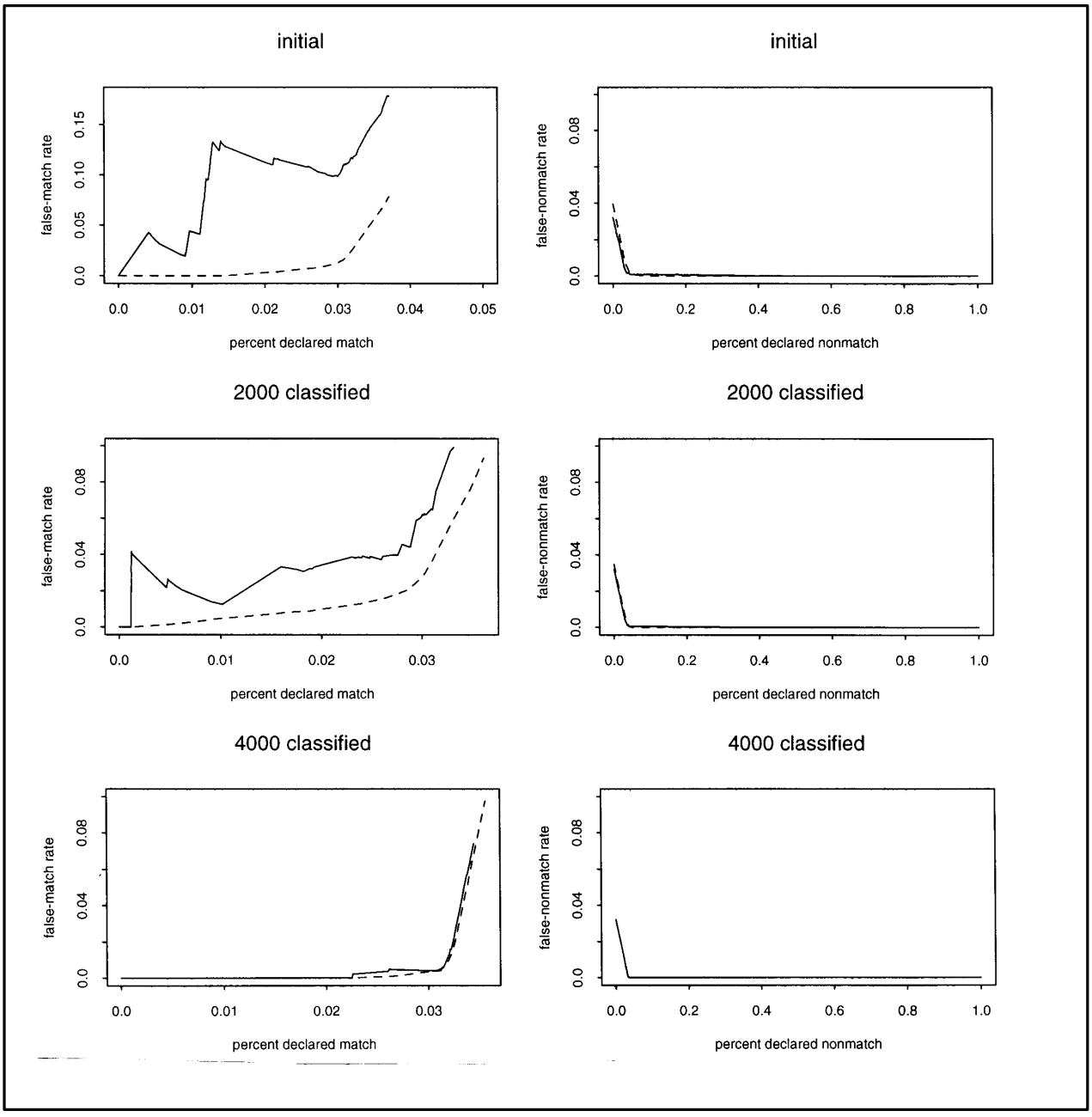
The goal of this paper has been to demonstrate methods that could be used in new record-linkage situations with big lists where accuracy, automation, and efficiency are needed. The procedure identifies matches and nonmatches, directs clerks in their work, and provides cut-offs and estimates of error rates on five Census data sets.

Acknowledgments

The author wishes to thank William E. Winkler and Donald B. Rubin for their support and guidance in this project. Thanks also to William E. Winkler and Fritz Scheuren for organizing the Record Linkage Workshop.

Figure 2. -- False-Match (FMR) and False-Nonmatch (FNMR) Rates for D90c: Initial Estimates, Estimates After Reviewing 2000, and Estimates After Reviewing 4000 Pairs

(Note that the initial FMR plot has different axes than the others)



References

Armstrong, J. B. and Mayda, J. E. (1993). Model-Based Estimation of Record Linkage Error Rates, *Survey Methodology*, 19, 137-147.

Belin, Thomas R. (1993). Evaluation of Sources of Variation in Record Linkage Through a Factorial Experiment, *Survey Methodology*, 19, 13-29.

Belin, Thomas R. and Rubin, Donald B. (1995). A Method for Calibrating False-Match Rates in Record Linkage, *Journal of the American Statistical Association*, 90, 694-707.

Dempster, A. P.; Laird, N. M.; and Rubin, Donald B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm, *Journal of the Royal Statistical Society*, Series B, 39, 1-22, (C/R: 22-37).

Fellegi, Ivan P. and Sumter, Alan B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, 64, 1183-1210.

Goodman, Leo A. (1974). Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models, *Biometrika*, 61, 215-231.

Haberman, Shelby J. (1974). Log-Linear Models for Frequency Tables Derived by Indirect Observation: Maximum Likelihood Equations, *The Annals of Statistics*, 2, 911-924.

Haberman, Shelby J. (1979). *Analysis of Qualitative Data*, Vol. 2, New York: Academic Press.

Meng, Xiao-Li and Rubin, Donald B. (1993). Maximum Likelihood Estimation Via the ECM Algorithm: A General Framework, *Biometrika*, 80, 267-278.

Thibaudeau, Yves. (1989). Fitting Log-Linear Models in Computer Matching, *Proceedings of the Section on Statistical Computing, American Statistical Association*, 283-288.

Thibaudeau, Yves. (1993). The Discrimination Power of Dependency Structures in Record Linkage, *Survey Methodology*, 19, 31-38.

Winkler, William E. (1988). Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage, *Proceedings of the Survey Research Methods Section, American Statistical Association*, 667-671.

Winkler, William E. (1989a). Frequency-Based Matching in the Fellegi-Sunter Model of Record Linkage, *Proceedings of the Survey Research Methods Section, American Statistical Association*, 778- 783.

Winkler, William E. (1989b). Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage, *Proceedings of the Bureau of the Census Annual Research Conference*, 5, 145- 155.

Winkler, William E. (1989c). Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage, *Survey Methodology*, 15, 101-117.

Winkler, William E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, *Proceedings of the Survey Research Methods Section, American Statistical Association*, 354-359.

Winkler, William E. (1992), Comparative Analysis of Record Linkage Decision Rules, *Proceedings of the Survey Research Methods Section, American Statistical Association,* 829- 834.

Winkler, William E. (1993). Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage, *Proceedings of the Survey Research Methods Section, American Statistical Association,* 274- 279.

Winkler, William E. (1994). Advanced Methods for Record Linkage, *Proceedings of the Survey Research Methods Section, American Statistical Association,* 467-472.

Chapter
4

Regression Analysis of Data Files that Are Computer Matched – Part I

Fritz Scheuren, Ernst and Young, LLP
William E. Winkler, Bureau of the Census

ABSTRACT

This paper focuses on how to deal with record linkage errors when engaged in regression analysis. Recent work by Rubin and Belin (1991) and by Winkler and Thibaudeau (1991) provides the theory, computational algorithms, and software necessary for estimating matching probabilities. These advances allow us to update the work of Neter, Maynes, and Ramanathan (1965). Adjustment procedures are outlined and some successful simulations are described. Our results are preliminary and intended largely to stimulate further work.

KEY WORDS: Record linkage; Matching error; Regression analysis.

1. INTRODUCTION

Information that resides in two separate computer data bases can be combined for analysis and policy decisions. For instance, an epidemiologist might wish to evaluate the effect of a new cancer treatment by matching information from a collection of medical case studies against a death registry in order to obtain information about the cause and date of death (*e.g.*, Beebe 1985). An economist might wish to evaluate energy policy decisions by matching a data base containing fuel and commodity information for a set of companies against a data base containing the values and types of goods produced by the companies (*e.g.*, Winkler 1985). If unique identifiers, such as verified social security numbers or employer identification numbers, are available, then matching data sources can be straightforward and standard methods of statistical analysis may be applicable directly.

When unique identifiers are not available (*e.g.*, Jabine and Scheuren 1986), then the linkage must be performed using information such as company or individual name, address, age, and other descriptive items. Even when typographical variations and errors are absent, name information such as "Smith" and "Robert" may not be sufficient, by itself, to identify an individual. Furthermore, the use of addresses is often subject to formatting errors because existing parsing or standardization software does not effectively allow comparison of, say, a house number with a house number and a street name with a street name. The addresses of an individual we wish to match may also differ because one is erroneous or because the individual has moved.

Over the last few years, there has been an outpouring of new work on record linkage techniques in North America (*e.g.*, Jaro 1989; and Newcombe, Fair and Lalonde 1992). Some of these results were spurred on by a series of conferences beginning in the mid-1980s (*e.g.*, Kilss and Alvey 1985; Howe and Spasoff 1986; Coombs and Singh 1987; Carpenter and Fair 1989); a further major stimulus in the U.S. has been the effort to study undercoverage in the 1990 Decennial Census (*e.g.*, Winkler and Thibaudeau 1991). The new book by Newcombe (1988) has also had an important role in this ferment. Finally, efforts elsewhere have also been considerable (*e.g.*, Copas and Hilton 1990).

What is surprising about all of this recent work is that the main theoretical underpinnings for computer-oriented matching methods are quite mature. Sound practice dates back at least to the 1950s and the work of Newcombe and his collaborators (*e.g.*, Newcombe *et al.* 1959). About a decade later, the underlying theory for these basic ideas was firmly established with the papers of Tepping (1968) and, especially, Fellegi and Sunter (1969).

Part of the reason for the continuing interest in record linkage is that the computer revolution has made possible better and better techniques. The proliferation of machine readable files has also widened the range of application. Still another factor has been the need to build bridges between the relatively narrow (even obscure) field of computer matching and the rest of statistics (*e.g.*, Scheuren 1985). Our present paper falls under this last category and is intended to look at what is special about regression analyses with matched data sets.

By and large we will not discuss linkage techniques here. Instead, we will discuss what happens *after* the link status has been determined. The setting, we will assume, is the typical one where the linker does his or her work separately from the analyst. We will also suppose that the analyst (or user) may want to apply conventional statistical techniques - regression, contingency tables, life tables, *etc.* - to the linked file. A key question we want to explore then is "What should the linker do to help the analyst?" A

Reprinted with permission from *Survey Methodology* (1993), 19, 1, pp. 39-58.

related question is "What should the analyst know about the linkage and how should that information be used?"

In our opinion it is important to conceptualize the linkage and analysis steps as part of a single statistical system and to devise appropriate strategies accordingly. Obviously the quality of the linkage effort may directly impact on any analyses done. Despite this, rarely are we given direct measures of that impact (*e.g.*, Scheuren and Oh 1975). Rubin (1990) has noted the need to make inferential statements that are designed to summarize evidence in the data being analyzed. Rubin's ideas were presented in the connotation of data housekeeping techniques like editing and imputation, where nonresponse can often invalidate standard statistical procedures that are available in existing software packages. We believe Rubin's perspective applies at least with equal force in record linkage work.

Organizationally, our discussion is divided into four sections. First, we provide some background on the linkage setting, because any answers - even partial ones - will depend on the files to be linked and the uses of the matched data. In the next section we discuss our methodological approach, focusing, as already noted, just on regression analysis. A few results are presented in section 4 from some exploratory simulations. These simulations are intended to help the reader weigh our ideas and get a feel for some of the difficulties. A final section consists of preliminary conclusions and ideas for future research. A short appendix containing more on theoretical considerations is also provided.

2. RECORD LINKAGE BACKGROUND

When linking two or more files, an individual record on one file may not be linked with the correct corresponding record on the other file. If a unique identifier for corresponding records on two files is not available - or is subject to inaccuracy - then the matching process is subject to error. If the resultant linked data base contains a substantial proportion of information from pairs of records that have been brought together erroneously or a significant proportion of records that need to be brought together are erroneously left apart, then statistical analyses may be sufficiently compromised that results of standard statistical techniques could be misleading. For the bulk of this paper we will only be treating the situation of how erroneous links affect analyses. The impact of problems caused by erroneous nonlinks (an implicit type of sampling that can yield selection biases) is discussed briefly in the final section.

2.1 Fellegi-Sunter Record Linkage Model

The record linkage process attempts to classify pairs in a product space $\mathbf{A} \times \mathbf{B}$ from two files A and B into M, the set of true links, and U, the set of true nonlinks. Making rigorous concepts introduced by Newcombe (*e.g.*, Newcombe *et al.* 1959), Fellegi and Sunter (1969) considered ratios of probabilities of the form:

$$R = Pr(\gamma \in \Gamma \mid M)/Pr(\gamma \in \Gamma \mid U), \qquad (2.1)$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on surname, first name, and age. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific surnames, such as Smith or Zabrinsky, occur. The fields that are compared (surname, first name, age) are referred to as *matching variables*.

The decision rule is given by:

If $R > Upper$, then designate pair as a link.

If $Lower \leq R \leq Upper$, then designate pair as a possible link and hold for clerical review. (2.2)

If $R < Lower$, then designate pair as a nonlink.

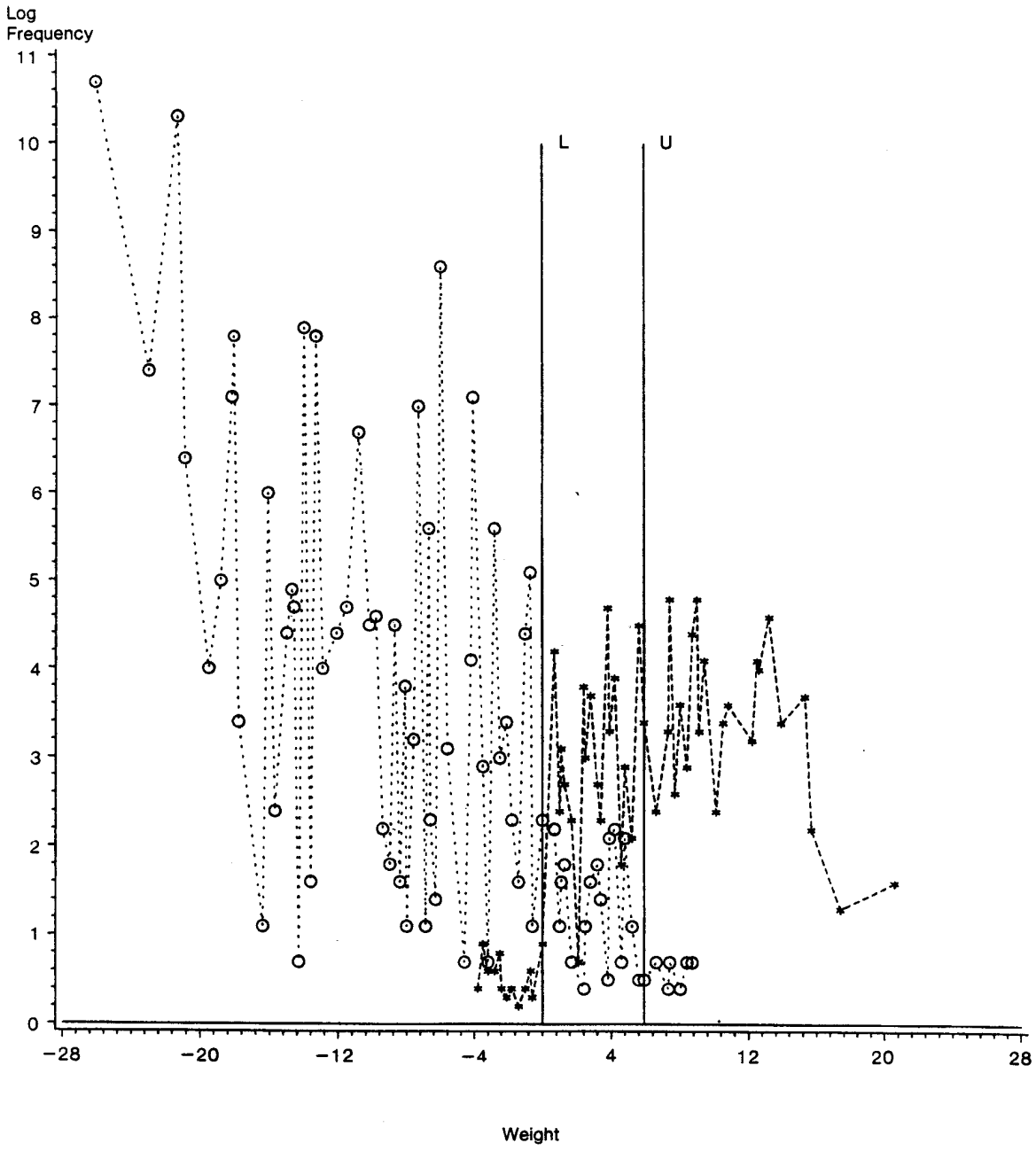
Fellegi and Sunter (1969) showed that the decision rule is optimal in the sense that for any pair of fixed bounds on R, the middle region is minimized over all decision rules on the same comparison space Γ. The cutoff thresholds *Upper* and *Lower* are determined by the error bounds. We call the ratio R or any monotonely increasing transformation of it (such as given by a logarithm) a *matching weight* or *total agreement weight*.

In actual applications, the optimality of the decision rule (2.2) is heavily dependent on the accuracy of the estimates of the probabilities given in (2.1). The probabilities in (2.1) are called *matching parameters*. Estimated parameters are (nearly) *optimal* if they yield decision rules that perform (nearly) as well as rule (2.2) does when the true parameters are used.

The Fellegi-Sunter approach is basically a direct extension of the classical theory of hypothesis testing to record linkage. To describe the model further, suppose there are two files of size n and m where - without loss of generality - we will assume that $n \leq m$. As part of the linkage process, a comparison might be carried out between all possible $n \times m$ pairs of records (one component of the pair coming from each file). A decision is, then, made as to whether or not the members of each comparison-pair represent the same unit or whether there is insufficient evidence to determine link status.

Schematically, it is conventional to look at the $n \times m$ pairs arrayed by some measure of the probability that the pair represent records for the same unit. In Figure 1, for example, we have plotted two curves. The curve on the right is a hypothetical distribution of the n true links by the "matching weight" (computed from (2.1) but in natural logarithms). The curve on the left is the remaining of the $n \times (m - 1)$ pairs - the true nonlinks - plotted by their matching weights again in logarithms.

Figure 1. Log Frequency *vs.* Weight, Links and Nonlinks



Typically, as Figure 1 indicates, the link and nonlink distributions overlap. At the extremes the overlap is of no consequence in arriving at linkage decisions; however, there is a middle region of potential links, say between "L" and "U", where it would be hard, based on Figure 1 alone, to distinguish with any degree of accuracy between links and nonlinks.

The Fellegi-Sunter model is valid on any set of pairs we consider. However, for computational convenience, rather than consider all possible pairs in $\mathbf{A} \times \mathbf{B}$, we might consider only a subset of pairs where the records from both files agree on key or "blocking" information that is thought to be highly accurate. Examples of the *logical blocking criteria* include items such as a geographical identifier like Postal (*e.g.*, ZIP) code or a surname identifier such as a Soundex or NYSIIS code (see *e.g.*, Newcombe 1988, pp. 182-184). Incidentally, the Fellegi-Sunter Model does not presuppose (as Figure 1 did) that among the $n \times m$ pairs there will be n links but rather, if there are no duplicates on A or B, that there will be at most n links.

2.2 Handling Potential Links

Even when a computer matching system uses the Fellegi-Sunter decision rule to designate some pairs as almost certain *true links* or *true nonlinks*, it could leave a large subset of pairs that are only potential links. One way to address potentially linked pairs is to clerically review them in an attempt to delineate true links correctly. A way to deal with erroneously nonlinked pairs is to perform additional (again possibly clerical) searches. Both of these approaches are costly, time-consuming, and subject to error.

Not surprisingly, the main focus of record linkage research since the beginning work of Newcombe has been how to reduce the clerical review steps caused by the potential links. Great progress has been made in improving linkage rules through better utilization of information in pairs of records and at estimating error rates via probabilistic models.

Record linkage decision rules have been improved through a variety of methods. To deal with minor typographical errors such as "Smith" versus "Smoth", Winkler and Thibaudeau (1991) extended the string comparator metrics introduced by Jaro (1989). Alternatively, Newcombe *et al.* (1989) developed methods for creating and using partial agreement tables. For certain classes of files, Winkler and Thibaudeau (1991) (see also Winkler 1992; Jaro 1989) developed Expectation-Maximization procedures and *ad hoc* modelling procedures based on *a priori* information that automatically yielded the optimal parameters in (2.1) for use in the decision rules (2.2).

Rubin and Belin (1991) introduced a method for estimating error rates, when error rates could not be reliably estimated via conventional methods (Belin 1991, pp. 19-20). Using a model that specified that the curves of weights versus log frequency produced by the matching process could be expressed as a mixture of two curves (links and nonlinks), Rubin and Belin estimated the curves which, in turn, gave estimates of error rates. To apply their method, Rubin and Belin needed a training sample to yield an *a priori* estimate of the shape of the two curves.

While many linkage problems arise in retrospective, often epidemiological settings, occasionally linkers have been able to designate what information is needed in both data sets to be linked based on known analytic needs. Requiring better matching information, such as was done with the 1990 Census Post-Enumeration Survey (see *e.g.*, Winkler and Thibaudeau 1991), assured that sets of potential links were minimized.

Despite these strides, eventually, the linker and analyst still may have to face a possible clerical review step. Even today, the remaining costs in time, money and hidden residual errors can still be considerable. Are there safe alternatives short of a full review? We believe so and this belief motivates our perspective in section 3, where we examine linkage errors in a regression analysis context. Other approaches, however, might be needed for different analytical frameworks.

3. REGRESSION WITH LINKED DATA

Our discussion of regression will presuppose that the linker has helped the analyst by providing a combined data file consisting of pairs of records - one from each input file - along with the match probability and the link status of each pair. Link, nonlink, and potential links would all be included and identified as such. Keeping likely links and potential links seems an obvious step; keeping likely nonlinks, less so. However, as Newcombe has pointed out, information from likely nonlinks is needed for computing biases. We conjecture that it will suffice to keep no more than two or three pairs of matches from the B file for each record on the A file. The two or three pairs with the highest matching weights would be retained.

In particular, we will assume that the file of linked cases has been augmented so that every record on the smaller of the two files has been paired with, say, the *two* records on the larger file having the highest matching weights. As $n \leq m$, we are keeping $2n$ of the $n \times m$ possible pairs. For each record we keep the linkage indicators and the probabilities associated with the records to which it is paired. Some of these cases will consist of (link, nonlink) combinations or (nonlink, nonlink) combinations. For simplicity's sake, we are not going to deal with settings where more than one true link could occur; hence, (link,link) combinations are by definition ruled out.

As may be quite apparent, such a data structure allows different methods of analysis. For example, we can partition

the file back into three parts - identified links, nonlinks, and potential links. Whatever analysis we are doing could be repeated separately for each group or for subsets of these groups. In the application here, we will use nonlinks to adjust the potential links, and, thereby, gain an additional perspective that could lead to reductions in the Mean Square Error (MSE) over statistics calculated only from the linked data.

For statistical analyses, if we were to use only data arising from pairs of records that we were highly confident were links, then we might be throwing away much additional information from the set of potentially linked pairs, which, as a subset, could contain as many true links as the set of pairs which we designate as links. Additionally, we could seriously bias results because certain subsets of the true links that we might be interested in might reside primarily in the set of potential links. For instance, if we were considering affirmative action and income questions, certain records (such as those associated with lower income individuals) might be more difficult to match using name and address information and, thus, might be heavily concentrated among the set of potential links.

3.1 Motivating Theory

Neter, Maynes, and Ramanathan (1965) recognized that errors introduced during the matching process could adversely affect analyses based on the resultant linked files. To show how the ideas of Neter *et al.* motivate the ideas in this paper, we provide additional details of their model. Neter *et al.* assumed that the set of records from one file (1) always could be matched, (2) always had the same probability p of being correctly matched, and (3) had the same probability q of being mismatched to any remaining records in the second file (*i.e.* $p + (N - 1)q = 1$ where N is file size). They generalized their basic results by assuming that the sets of pairs from the two files could be partitioned into classes in which (1), (2) and (3) held.

Our approach follows that of Neter *et al.* because we believe their approach is sensible. We concur with their results showing that if matching errors are moderate then regression coefficients could be severely biased. We do not believe, however, that condition (3) - which was their main means of simplifying computational formulas - will ever hold in practice. If matching is based on unique identifiers such as social security numbers subject to typographical error, it is unlikely that a typographical error will mean that a given record has the same probability of being incorrectly matched to all remaining records in the second file. If matching variables consist of name and address information (which is often subject to substantially greater typographical error), then condition (3) is even more unlikely to hold.

To fix ideas on how our work builds on and generalizes results of Neter *et al.* we consider a special case. Suppose we are conducting ordinary least squares using a simple regression of the form,

$$y = a_0 + a_1 x + \epsilon. \tag{3.1}$$

Next, assume mismatches have occurred, so that the y variables (from one file) and the x variables (from another file) are *not* always for the *same unit*.

Now in this setting, the unadjusted estimator of a_1 would be biased; however, under assumptions such as that x and y are independent when a mismatch occurs, it can be shown that, if we know the mismatch rate, h, that an unbiased adjusted estimator can be obtained by simply correcting the ordinary estimator by multiplying it by $(1/(1 - h))$. Intuitively, the erroneously linked pairs lead to an understatement of the true correlation (positive or negative) between x and y. The adjusted coefficient removes this understatement. With the adjusted slope coefficent $\hat{a}_1$, the proper intercept can be obtained from the usual expression $\hat{a}_0 = \bar{y} - \hat{a}_1\bar{x}$, where $\hat{a}_1$ has been adjusted.

Methods for estimating regression standard errors can also be devised in the presence of matching errors. Rather than just continuing to discuss this special case, though, we will look at how the idea of making a multiplicative adjustment can be generalized. Consider

$$Y = X\beta + \epsilon, \tag{3.2}$$

the ordinary univariate regression model, for which error terms all have mean zero and are independent with constant variance σ^2. If we were working with a data base of size n, Y would be regressed on X in the usual manner. Now, given that each case has two matches, we have $2n$ pairs altogether. We wish to use (X_i, Y_i), but instead use (X_i, Z_i). Z_i could be Y_i, but may take some other value, Y_j, due to matching error.

For $i = 1, \ldots, n$,

$$Z_i = \begin{cases} Y_i \text{ with probability } p_i \\ Y_j \text{ with probability } q_{ij} \text{ for } j \neq i, \end{cases} \tag{3.3}$$

$p_i + \sum_j q_{ij} = 1$.

The probability p_i may be zero or one. We define $h_i = 1 - p_i$ and divide the set of pairs into n mutually exclusive classes. The classes are determined by records from one of the files. Each class consists of the independent x-variable X_i, the true value of the dependent y-variable, the values of the y-variables from records in the second file to which the record in the first file containing X_i have been paired, and computer matching probabilities (or weights). Included are links, nonlinks, and potential links. Under an assumption of one-to-one matching, for each $i = 1, \ldots, n$, there exists at most one j such that $q_{ij} > 0$. We let ϕ be defined by $\phi(i) = j$.

The intuitive idea of our approach (and that of Neter *et al.*) is that we can, under the model assumptions, express each observed data point pair (X,Z) in terms of the true values (X,Y) and a bias term (X,b). All equations needed for the usual regression techniques can then be obtained. Our computational formulas are much more complicated than those of Neter *et al.* because their strong assumption (3) made considerable simplification possible in the computational formulas. In particular, under their model assumptions, Neter *et al.* proved that both the mean and variance of the observed Z-values were necessarily equal the mean and variance of the true Y-values.

Under the model of this paper, we observe (see Appendix) that

$$E(Z) = (1/n)\sum_i E(Z|i) = (1/n)\sum_i (Y_i p_i + \sum_j Y_j q_{ij})$$

$$= (1/n)\sum_i Y_i + (1/n)\sum_i [Y_i(-h_i) + Y_{\phi(i)}h_i]$$

$$= \bar{Y} + B. \qquad (3.4)$$

As each X_i, $i = 1, \ldots, n$, can be paired with either Y_i or $Y_{\phi(i)}$, the second equality in (3.4) represents $2n$ points. Similarly, we can represent σ_{zy} in terms of σ_{xy} and a bias term B_{xy}, and σ_z^2 in terms of σ_y^2 and a bias term B_{yy}. We neither assume that the bias terms have expectation zero nor that they are uncorrelated with the observed data.

With the different representations, we can adjust the regression coefficients β_{zx} and their associated standard errors back to the true values β_{yx} and their associated standard errors. Our assumption of one-to-one matching (which is not needed for the general theory) is done for computational tractability and to reduce the number of records and amount of information that must be tracked during the matching process.

In implementing the adjustments, we make two crucial assumptions. The first is that, for $i = 1, \ldots, n$, we can accurately estimate the true probabilities of a match p_i. See Appendix for the method of Rubin and Belin (1991). The second is that, for each $i = 1, \ldots, n$, the true value Y_i associated with independent variable X_i is the pair with the highest matching weight and the false value $Y_{\phi(i)}$ is associated with the second highest matching weight. (From the simulations conducted it appears that at least the first of these two assumptions matters greatly when a significant portion of the pairs are potential links.)

3.2 Simulated Application

Using the methods just described, we attempted a simulation with real data. Our basic approach was to take two files for which true linkage statuses were known and re-link them using different matching variables - or really versions of the same variables with different degrees of distortion introduced, making it harder and harder to distinguish a link from a nonlink. This created a setting where there was enough discrimination power for the Rubin-Belin algorithm for estimating probabilities to work, but not so much discriminating power that the overlap area of potential links becomes insignificant.

The basic simulation results were obtained by starting with a pair of files of size 10,000 that had good information for matching and for which true match status was known. To conduct the simulations a range of error was introduced into the matching variables, different amounts of data were used for matching, and greater deviations from optimal matching probabilities were allowed.

Three matching scenarios were considered: (1) *good*, (2) *mediocre*, and (3) *poor*. The good matching scenario consisted of using most of the available procedures that had been developed for matching during the 1990 U.S. Census (*e.g.*, Winkler and Thibaudeau 1991). Matching variables consisted of last name, first name, middle initial, house number, street name, apartment or unit identifier, telephone, age, marital status, relationship to head of household, sex, and race. Matching probabilities used in crucial likelihood ratios needed for the decision rules were chosen close to optimal.

The mediocre matching scenario consisted of using last name, first name, middle initial, two address variations, apartment or unit identifier, and age. Minor typographical errors were introduced independently into one seventh of the last names and one fifth of the first names. Matching probabilities were chosen to deviate from optimal but were still considered to be consistent with those that might be selected by an experienced computer matching expert.

The poor matching scenario consisted of using last name, first name, one address variation, and age. Minor typographical errors were introduced independently into one fifth of the last names and one third of the first names. Moderately severe typographical errors were made in one fourth of the addresses. Matching probabilities were chosen that deviated substantially from optimal. The intent was for them to be selected in a manner that a practitioner might choose after gaining only a little experience.

With the various scenarios, our ability to distinguish between true links and true nonlinks differs significantly. For the good scenario, we see that the scatter for true links and nonlinks is almost completely separated (Figure 2). With the mediocre scheme, the corresponding sets of points overlap moderately (Figure 3); and, with the poor, the overlap is substantial (Figure 4).

We primarily caused the good matching scenario to degenerate to the poor matching error (Figures 2-4) by using less matching information and inducing typographical error in the matching variables. Even if we had kept the same matching variables as in the good matching scenario (Figure 2), we could have caused curve overlap (as in Figure 4) merely by varying the matching

Figure 2. Log of Frequency *vs.* Weight Good Matching Scenario, Links and Nonlinks

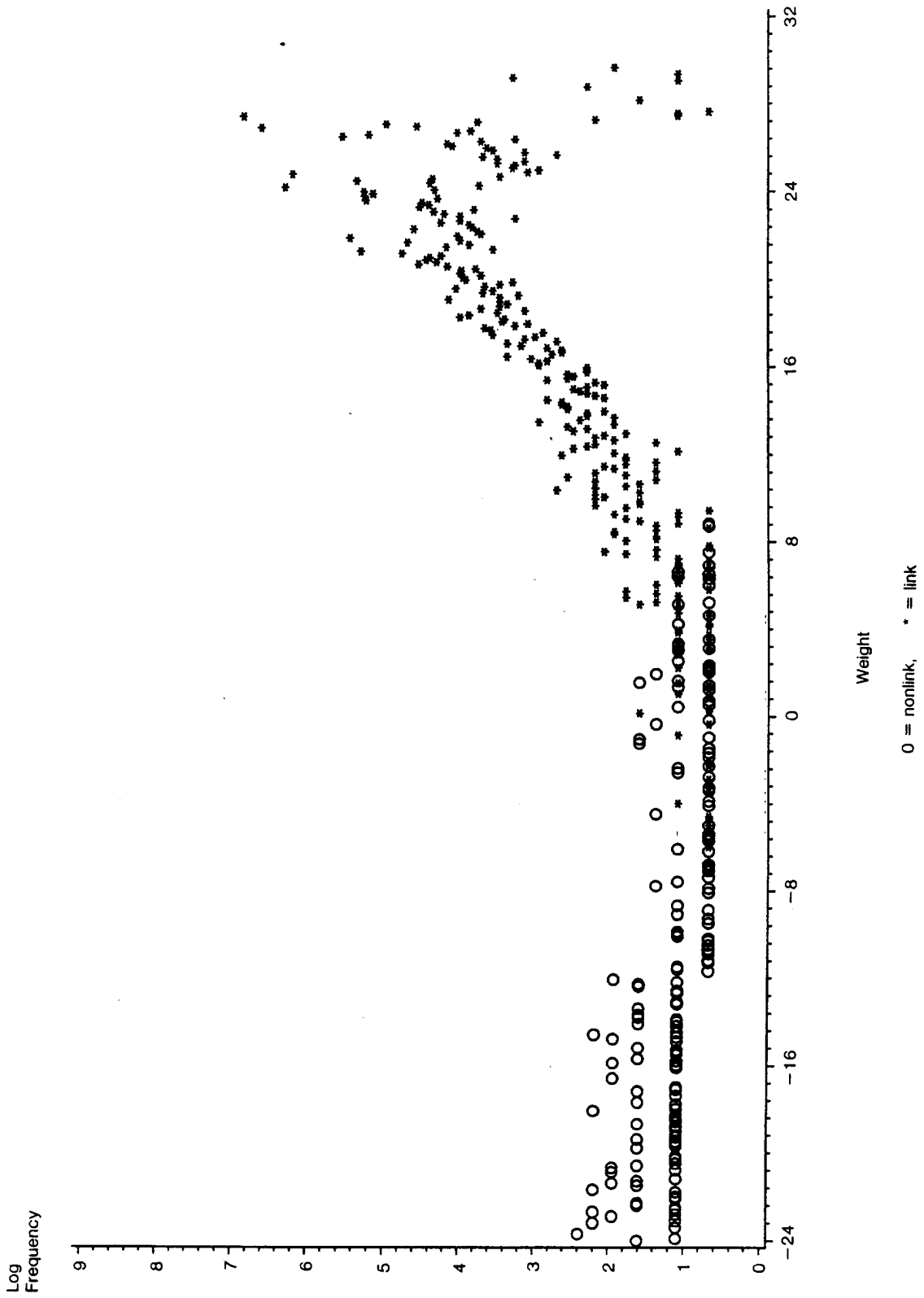


Figure 3. Log of Frequency *vs.* Weight Mediocre Matching Scenario, Links and Nonlinks

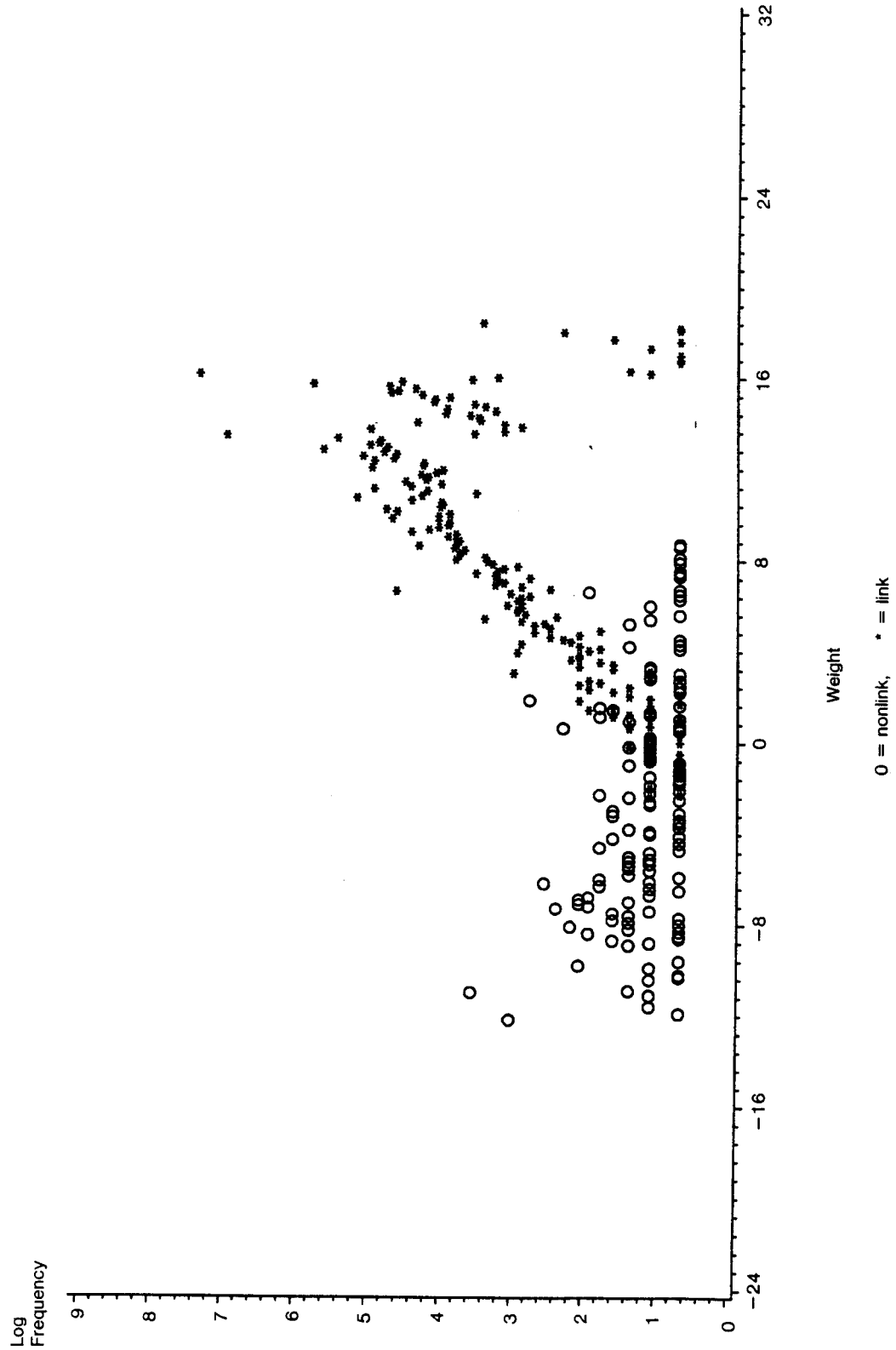


Figure 4. Log Frequency *vs.* Weight Poor Matching Scenario, Links and Nonlinks

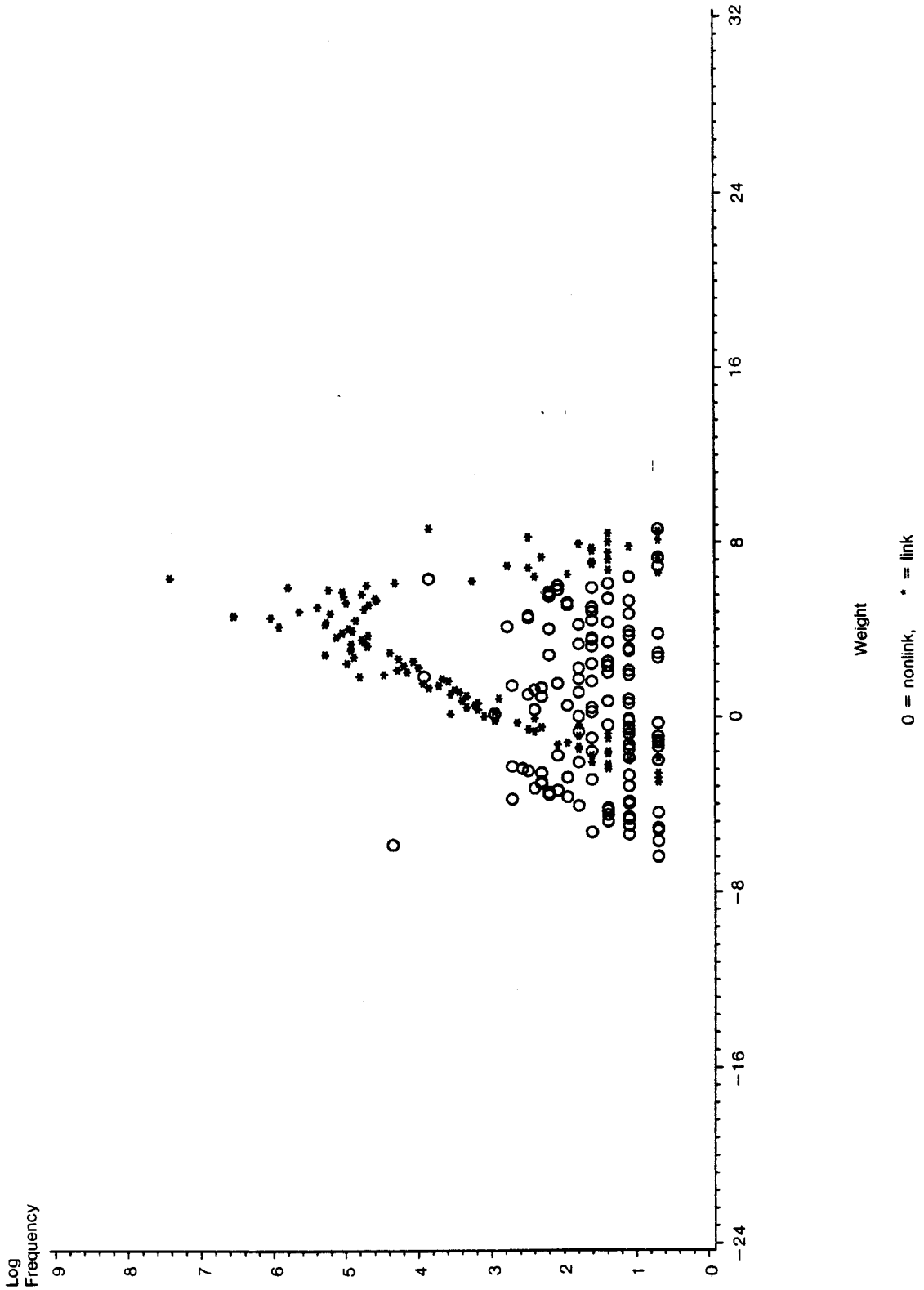


Table 1

Counts of True Links and True Nonlinks and Probabilities of an Erroneous Link in Weight Ranges for Various Matching Cases; Estimated Probabilities via Rubin-Belin Methodology

Weight	False match rates											
	Good				Mediocre				Poor			
	True		Prob		True		Prob		True		Prob	
	Link	NL	True	Est	Link	NL	True	Est	Link	NL	True	Est
15+	9,176	0	.00	.00	2,621	0	.00	.00	0	1	.00	.00
14	111	0	.00	.00	418	0	.00	.00	0	1	.00	.00
13	91	0	.00	.01	1,877	0	.00	.00	0	1	.00	.00
12	69	0	.00	.02	1,202	0	.00	.00	0	1	.00	.00
11	59	0	.00	.03	832	0	.00	.00	0	1	.00	.00
10	69	0	.00	.05	785	0	.00	.00	0	1	.00	.00
9	42	0	.00	.08	610	0	.00	.00	0	1	.00	.00
8	36	2	.05	.13	439	3	.00	.00	65	1	.02	.00
7	30	1	.03	.20	250	4	.00	.01	39	1	.03	.00
6	14	7	.33	.29	265	9	.03	.03	1,859	57	.03	.03
5	28	4	.12	.40	167	8	.05	.06	1,638	56	.03	.03
4	6	3	.33	.51	89	6	.06	.11	2,664	62	.02	.05
3	12	7	.37	.61	84	5	.06	.20	1,334	31	.02	.11
2	8	6	.43	.70	38	7	.16	.31	947	30	.03	.19
1	7	13	.65	.78	33	34	.51	.46	516	114	.18	.25
0	7	4	.36	.83	13	19	.59	.61	258	65	.20	.28
−1	3	5	.62	.89	7	20	.74	.74	93	23	.20	.31
−2	0	11	.99	.91	3	11	.79	.84	38	23	.38	.41
−3	4	6	.60	.94	4	19	.83	.89	15	69	.82	.60
−4	4	3	.43	.95	0	15	.99	.94	1	70	.99	.70
−5	4	4	.50	.97	0	15	.99	.96	0	25	.99	.68
−6	0	5	.99	.98	0	27	.99	.98	0	85	.99	.67
−7	1	6	.86	.98	0	40	.99	.99			.99	.99
−8	0	8	.99	.99	0	41		.99			.99	.99
−9	0	4	.99	.99	0	4		.99			.99	.99
−10−	0	22			0	22		.99			.99	.99

Notes: In the first column, weight 10 means weight range from 10 to 11. Weight ranges 15 and above and weight ranges −9 and below are added together. Weights are log ratios that are based on estimated agreement probabilities. **NL** is nonlinks and **Prob** is probability.

parameters given by equation (2.1). The poor matching scenario can arise when we do not have suitable name parsing software that allows comparison of corresponding surnames and first names or suitable address parsing software that allows comparison of corresponding house numbers and street names. Lack of proper parsing means that corresponding matching variables associated with many true links will not be properly utilized.

Our ability to estimate the probability of a match varies significantly. In Table 1 we have displayed these probabilities, both true and estimated, by weight classes. For the good and mediocre matching scenarios, estimated probabilities were fairly close to the true values. For the poor scenario, in which most pairs are potential links, deviations are quite substantial.

For each matching scenario, empirical data were created. Each data base contained a computer matching weight, true and estimated matching probabilities, the independent x-variable for the regression, the true dependent y-variable, the observed y-variables in the record having the highest match weight, and the observed y-variable from the record having the second highest matching weight.

The independent x-variables for the regression were constructed using the SAS RANUNI procedure, so as to be uniformly distributed between 1 and 101. For this paper, they were chosen independently of any matching variables. (While we have considered the situation for which regression variables are dependent on one or more matching variables (Winkler and Scheuren 1991), we do not present any such results in this paper.)

Three regression scenarios were then considered. They correspond to progressively lower R^2 values: (1) R^2 between 0.75 and 0.80; (2) between 0.40 and 0.45; and (3) between 0.20 and 0.22. The dependent variables were generated with independent seeds using the SAS RANNOR procedure. Within each matching scenario (good, mediocre, or poor), all pairing of records obtained by the matching process and, thus, matching error was fixed.

It should be noted that there are two reasons why we generated the (x,y)-data used in the analyses. First, we wanted to be able to control the regression data sufficiently well to determine what the effect of matching error was. This was an important consideration in the very large Monte Carlo simulations reported in Winkler and Scheuren (1991). Second, there existed no available pairs of data files in which highly precise matching information is available and which contain suitable quantitative data.

In performing the simulations for our investigation, some of which are reported here, we created more than 900 data bases, corresponding to a large number of variants of the three basic matching scenarios. Each data base contained three pairs of (x,y)-variables corresponding to the three basic regression scenarios. An examination of these data bases was undertaken to look at some of the matching sensitivity of the regressions and associated adjustments to the sampling procedure. The different data bases determined by different seed numbers are called *different samples*.

The regression adjustments were made separately for each weight class shown in Table 1, using both the estimated and true probabilities of linkage. In Table 1, weight class 10 refers to pairs having weights between 10 and 11 and weight class -1 refers to pairs having weights between -0 and -1. All pairs having weights 15 and above are combined into class 15+ and all pairs having weights -9 and below are combined into class $-10-$. While it was possible with the Rubin-Belin results to make individual adjustments for linkage probabilities, we chose to make average adjustments, by each weight class in Table 1. (See Czajka *et al.* 1992, for discussion of a related decision. Our approach has some of the flavor of the work on propensity scores (*e.g.*, Rosenbaum and Rubin 1983, 1985). Propensity scoring techniques, while proposed for other classes of problems, may have application here as well.

4. SOME HIGHLIGHTS AND LIMITATIONS OF THE SIMULATION RESULTS

Because of space limitations, we will present only a few representative results from the simulations conducted. For more information, including an extensive set of tables, see Winkler and Scheuren (1991).

The two outcome measures from our simulation that we consider are the relative bias and relative standard error. We will only discuss the mediocre matching scenario in detail and only for the case R^2 between 0.40 and 0.45. Figures 5-7 shows the relative bias results from a single representative sample. An overall summary, though, for the other scenarios is presented in Table 2. Some limitations on the simulation are also noted at the end of this section.

4.1 Illustrative Results for Mediocre Matching

Rather than use all pairs, we only consider pairs having weights 10 or less. Use of the smaller subset of pairs allows us to examine regression adjustment procedures for weight classes having low to high proportions of true nonlinks. We note that the eliminated pairs (having weight 10 and above) are associated only with true links. Figures 5 and 6 present our results for adjusted and unadjusted regression data, respectively. Results obtained with unadjusted data are based on conventional regression formulas (*e.g.*, Draper and Smith 1981). The weight classes displayed are cumulative beginning with pairs having the highest weight. Weight class w refers to all pairs having weights between w and 10.

We observe the following:

- The *accumulation* is by decreasing matching weight (*i.e.* from classes most likely to consist almost solely of true links to the classes containing increasing higher proportions of true nonlinks). In particular, for weight class $w = 8$, the first data point shown in Figures 5-7, there were 3 nonlinks and 439 links. By the time, say, we had cumulated the data through weight class $w = 5$, there were 24 nonlinks; the links, however, had grown to 1,121 - affording us a much larger overall sample size with a corresponding reduction in the regression standard error.
- Relative *biases* are provided for the original and adjusted slope coefficient $\hat{a}_1$ by taking the ratio of the true coefficient (about 2) and the calculated one for each cumulative weight class.
- Adjusted regression results are shown employing both estimated and true match probabilities. In particular, Figure 5 corresponds to the results obtained using estimated probabilities (all that would ordinarily be available in practice). Figure 7 corresponds to the unrealistic situation for which we knew the true probabilities.
- Relative *root mean square errors* (not shown) are obtained by calculating MSEs for each cumulative weight class. For each class, the bias is squared, added to the square of the standard errors, and square roots taken.

Observations on the results we obtained are fairly straightforward and about what we expected. For example, as sample size increased, we found the relative root mean square errors decreasd substantially for the adjusted coefficients. If the regression coefficients were not adjusted,

Figure 5. Relative Bias for Adjusted Estimators, Estimated Probabilities

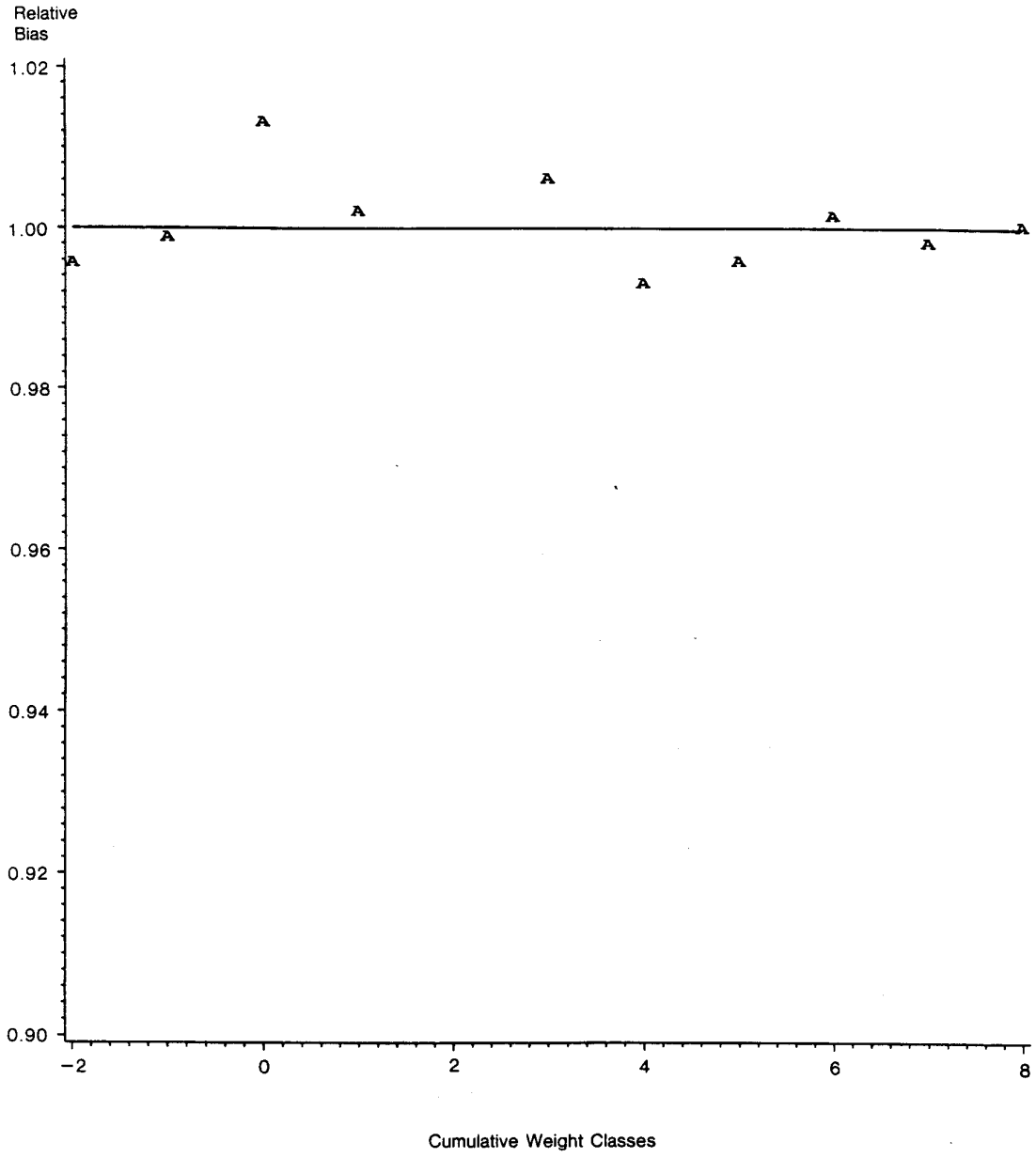


Figure 6. Relative Bias for Unadjusted Estimators

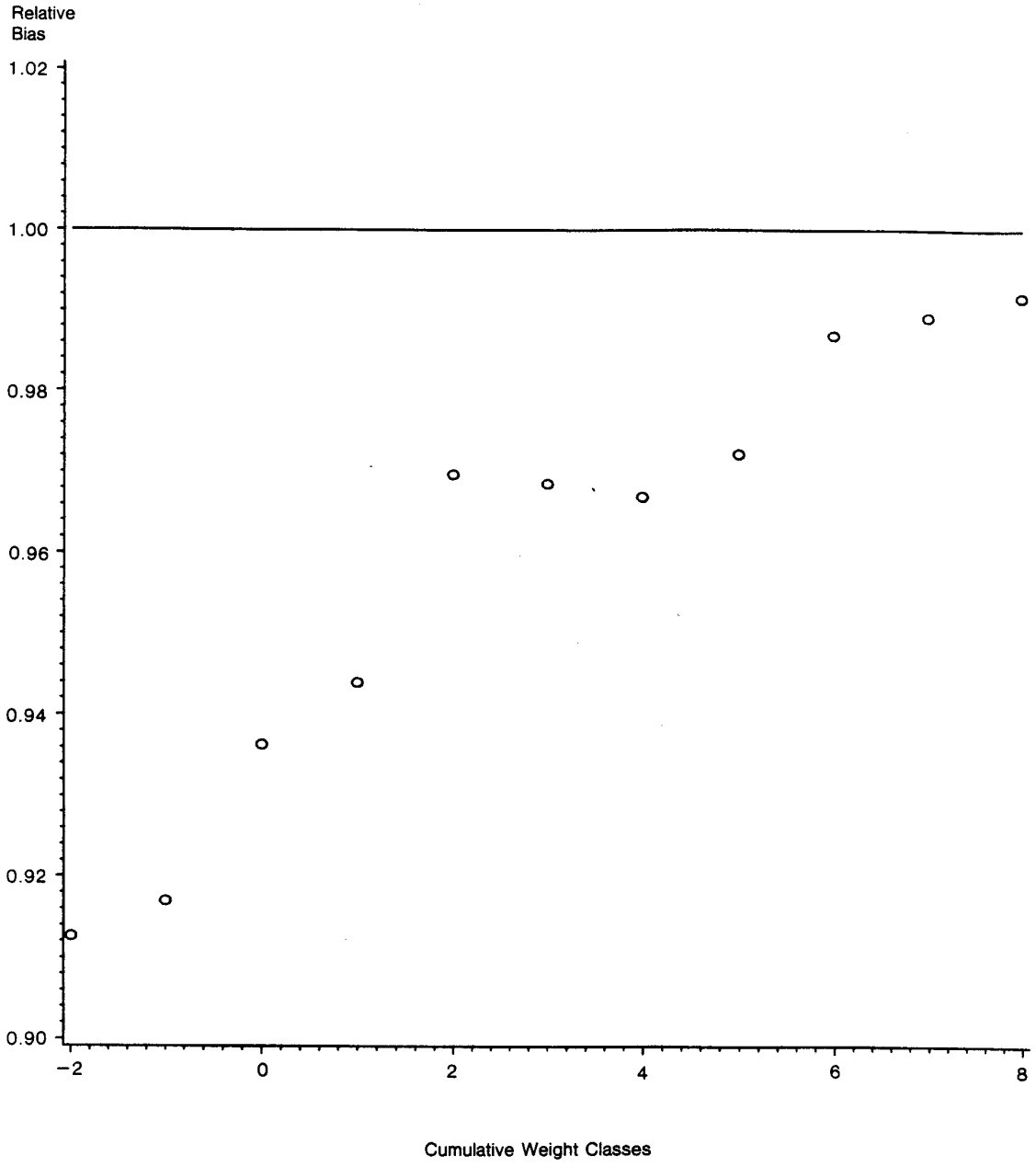
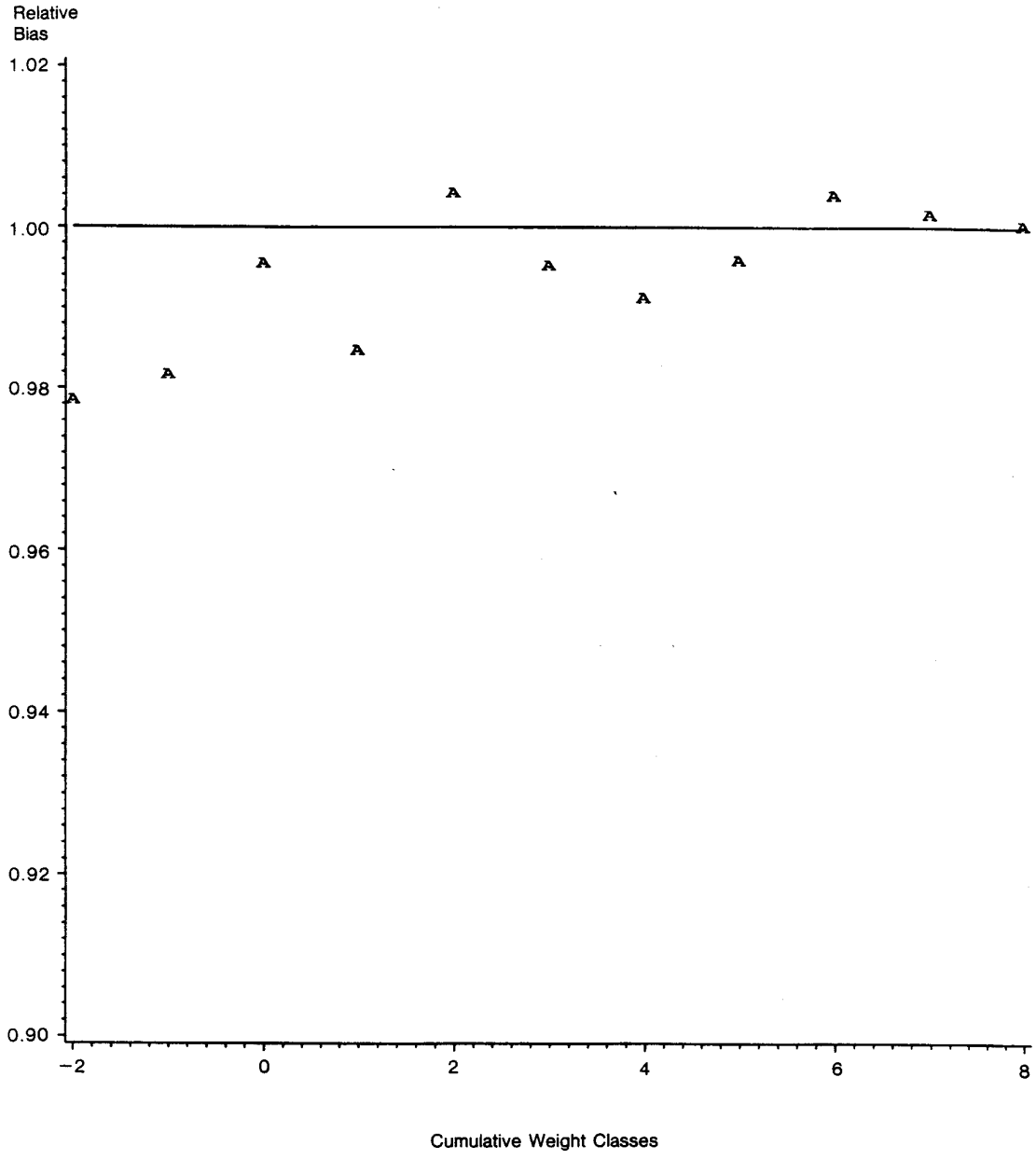


Figure 7. Relative Bias for Adjusted Estimators, True Probabilities



standard errors still decreased as the sample size grew, but at an unacceptably high price in increased bias.

One point of concern is that our ability to accurately estimate matching probabilities critically affects the accuracy of the coefficient estimates. If we can accurately estimate the probabilities (as in this case), then the adjustment procedure works reasonably well; if we cannot (see below), then the adjustment could perform badly.

4.2 Overall Results Summary

Our results varied somewhat for the three different values of R^2 - being better for larger R^2 values. These R^2 differences, however, do not change our main conclusions; hence, Table 2 does not address them. Notice that, for the good matching scenario, attempting to adjust does little good and may even cause some minor harm. Certainly it is pointless, in any case, and we only included it in our simulations for the sake of completeness. At the other extreme, even for poor matches, we obtained satisfactory results, but only when using the true probabilities - something not possible in practice.

Table 2

Summary of Adjustment Results for Illustrative Simulations

Basis of adjustments	Matching scenarios		
	Good	Mediocre	Poor
True probabilities	Adjustment was not helpful because it was not needed	Good results like those in Section 4.1	Good results like those in Section 4.1
Estimated probabilities	Same as above	Same as above	Poor results because Rubin-Belin could not estimate the probabilities

Any statistical estimation procedure will have difficulty with the poor matching scenario because of the extreme overlap of the curves. See Figure 4. We believe the mediocre scenario covers a wide range of typical settings. Nonetheless, the poor matching scenario might arise fairly often too, especially with less experienced linkers. Either new estimation procedures will have to be developed for the poor case or the Rubin-Belin probability estimation procedure - which was not designed for this situation - will have to be enhanced.

4.3 Some Simulation Limitations

The simulation results are subject to a number of limitations. Some of these are of possible major practical significance; others less so. A partial list follows:

- In conducting simulations for this paper, we assumed that the highest weight pair was a true link and the second highest a true nonlink. This assumption fails because, sometimes, the second highest is the true link and the highest a true nonlink. (We do not have a clear sense of how important this issue might be in practice. It would certainly have to be a factor in poor matching scenarios.)
- A second limitation of the data sets employed for the simulations is that the truly linked record may not be present at all in the file to which the first file is being matched. (This could be important. In many practical settings, we would expect the "logical blocking criteria" also to cause both pairs used in the adjustment to be false links.)
- A third limitation of our approach is that no use has been made of conventional regression diagnostic tools. (Depending on the environment, outliers created because of nonlinks could wreak havoc with underlying relationships. In our simulations this did not show up as much of a problem, largely, perhaps, because the X and Y values generated were bounded in a moderately narrow range.)

5. CONCLUSIONS AND FUTURE WORK

The theoretical and related simulation results presented here are obviously somewhat contrived and artificial. A lot more needs to be done, therefore, to validate and generalize our beginning efforts. Nonetheless, some recommendations for current practice stand out, as well as areas for future research. We will cover first a few of the topics that intrigued us as worthy of more study to improve the adjustment of potential links. Second, some remarks are made about the related problem of what to do with the (remaining) nonlinks. Finally, the section ends with some summary ideas and a revisitation of our perspective concerning the unity of the tasks that linkers and analysts do.

5.1 Improvements in Linkage Adjustment

An obvious question is whether our adjustment procedures could borrow ideas from general methods for errors-in-variables (*e.g.*, Johnston 1972). We have not explored this, but there may be some payoffs.

Of more interest to us are techniques that grow out of conventional regression diagnostics. A blend of these with our approach has a lot of appeal. Remember we are making adjustments, weight class by weight class. Suppose we looked ahead of time at the residual scatter in a particular weight class, where the residuals were calculated around the regression obtained from the cumulative weight classes above the class in question. Outliers, say, could then be identified and might be treated as nonlinks rather than potential links.

We intend to explore this possibility with simulated data that is heavier-tailed than what was used here. Also we will explore consciously varying the length of the weight classes and the minimum number of cases in each class. We have an uneasy feeling that the number of cases in each class may have been too small in places. (See Table 1.) On the other hand, we did not use the fact that the weight classes were of equal length nor did we study what would have happened had they been of differing lengths.

One final point, as noted already: we believe our approach has much in common with propensity scoring, but we did not explicitly appeal to that more general theory for aid and this could be something worth doing. For example, propensity scoring ideas may be especially helpful in the case where the regression variables and the linkage variables are dependent. (See Winkler and Scheuren (1991) for a report on the limited simulations undertaken and the additional difficulties encountered.)

5.2 Handling Erroneous Nonlinks

In the use of record linkage methods the general problem of selection bias arises because of erroneous nonlinks. There are a number of ways to handle this. For example, the links could be adjusted by the analyst for lack of representativeness, using the approaches familiar to those who adjust for unit or, conceivably, item nonresponse (*e.g.*, Scheuren *et al.* 1981).

The present approach for handling potential links could help reduce the size of the erroneous nonlink problem but, generally, would not eliminate it. To be specific, suppose we had a linkage setting where, for resource reasons, it was infeasible to follow up on the potential links. Many practitioners might simply drop the potential links, thereby, increasing the number of erroneous nonlinks. (For instance, in ascertaining which of a cohort's members is alive or dead, a third possibility - unascertained - is often used.)

Our approach to the potential links would have *implicitly* adjusted for that portion of the erroneous nonlinks which were potentially linkable (with a followup step, say). Other erroneous nonlinks would generally remain and another adjustment for them might still be an issue to consider.

Often we can be faced with linkage settings where the files being linked have subgroups with matching information of varying quality, resulting in differing rates of erroneous links and nonlinks. In principle, we could employ the techniques in this paper to each subgroup separately. How to handle very small subgroups is an open problem and the effect on estimated differences between subgroups, even when both are of modest size, while seemingly straightforward, deserves study.

5.3 Concluding Comments

At the start of this paper we asked two "key" questions. Now that we are concluding, it might make sense to reconsider these questions and try, in summary fashion, to give some answers.

- "*What should the linker do to help the analyst?*" If possible, the linker should play a role in designing the datasets to be matched, so that the identifying information on both is of high quality. Powerful algorithms exist now in several places to do an excellent job of linkage (*e.g.*, at Statistics Canada or the U.S. Bureau of the Census, to name two). Linkers should resist the temptation to design and develop their own software. In most cases, modifying or simply using existing software is highly recommended (Scheuren 1985). Obviously, for the analyst's sake, the linker needs to provide as much linkage information as possible on the files matched so that the analyst can make informed choices in his or her work. In the present paper we have proposed that the links, nonlinks, and potential links be provided to the analyst - not just links. We strongly recommend this, even if a clerical review step has been undertaken. We do *not* necessarily recommend the particular choices we made about the file structure, at least not without further study. We would argue, though, that our choices are serviceable.
- "*What should the analyst know about the linkage and how should this be used?*" The analyst needs to have information like link, nonlink, and potential link status, along with linkage probabilities, if available. Many settings could arise where simply doing the data analysis steps separately by link status will reveal a great deal about the sensitivity of one's results. The present paper provides some initial ideas about how this use might be approached in a regression context. There also appears to be some improvements possible using the adjustments carried out here, particularly for the mediocre matching scenario. How general these improvements are remains to be seen. Even so, we are relatively pleased with our results and look forward to doing more. Indeed, there are direct connections to be made between our approach to the regression problem and other standard techniques, like contingency table loglinear models.

Clearly, we have not developed complete, general answers to the questions we raised. We hope, though, that this paper will at least stimulate interest on the part of others that could lead us all to better practice.

ACKNOWLEDGMENTS AND DISCLAIMERS

The authors would like to thank Yahia Ahmed and Mary Batcher for their help in preparing this paper and two referees for detailed and discerning comments. Fruitful discussions were held with Tom Belin. Wendy Alvey also provided considerable editorial assistance.

The usual disclaimers are appropriate here: in particular, this paper reflects the views of the authors and not necessarily those of their respective agencies. Problems, like a lack of clarity in our thinking or in our exposition, are entirely the authors' responsibility.

APPENDIX

The appendix is divided into four sections. The first provides details on how matching error affects regression models for the simple univariate case. The approach most closely resembles the approach introduced by Neter *et al.* (1965) and provides motivation for the generalizations presented in appendix sections two and three. Computational formulas are considerably more complicated than those presented by Neter *et al.* because we use a more realistic model of the matching process. In the second section, we extend the univariate model to the case for which all independent variables arise from one file, while the dependent variable comes from the other, and, in the third, we extend the second case to that in which some independent variables come from one file and some come from another. The fourth section summarizes methods of Rubin and Belin (1991) (see also Belin 1991) for estimating the probability of a link.

A.1. Univariate Regression Model

In this section we address the simplest regression situation in which we match two files and consider a set of numeric pairs in which the independent variable is taken from a record in one file and the dependent variable is taken from the corresponding matched record from the other file.

Let $Y = X\beta + \epsilon$ be the ordinary univariate regression model for which error terms are independent with expectation zero and constant variance σ^2. If we were working with a single data base, Y would be regressed on X in the usual manner. For $i = 1, \ldots, n$, we wish to use (X_i,Y_i) but we will use (X_i,Z_i), where Z_i is usually Y_i but it may take some other value Y_j due to matching error.

That is, for $i = 1, \ldots, n$,

$$z_i = \begin{cases} Y_i & \text{with probability } p_i \\ Y_j & \text{with probability } q_{ij} \text{ for } j \neq i, \end{cases}$$

where $p_i + \sum_{j \neq i} q_{ij} = 1$.

The probability p_i may be zero or one. We define $h_i = 1 - p_i$. As in Neter *et al.* (1965), we divide the set of pairs into n mutually exclusive classes. Each class consists of exactly one (X_i,Z_i) and, thus, there are n classes. The intuitive idea of our procedure is that we basically adjust Z_i in each (X_i,Z_i) for the bias induced by the matching process. The accuracy of the adjustment is heavily dependent on the accuracy of the estimates of the matching probabilities in our model.

To simplify the computational formulas in the explanation, we assume one-to-one matching; that is, for each $i = 1, \ldots, n$, there exists at most one j such that $q_{ij} > 0$. We let ϕ be defined by $\phi(i) = j$. Our model still applies if we do not assume one-to-one matching.

As intermediate steps in estimating regression coefficients and their standard errors, we need to find $\mu_z \equiv E(Z)$, σ_z^2, and σ_{zx}. As in Neter *et al.* (1965),

$$\begin{aligned} E(Z) &\equiv (1/n)\sum_i E(Z|i) \equiv (1/n)\sum_i (Y_i p_i + \sum_{j \neq i} Y_j q_{ij}) \\ &= (1/n)\sum_i Y_i \\ &\quad + (1/n)\sum_i [Y_i(-h_i) + Y_{\phi(i)} h_i] \\ &\equiv \bar{Y} + B. \end{aligned} \tag{A.1.1}$$

The first and second equalities are by definition and the third is by addition and subtraction. The third inequality is the first time we apply the one-to-one matching assumption. The last term on the right hand side of the equality is the bias which we denote by B. Note that the overall bias B is the statistical average (expectation) of the individual biases $[Y_i(-h_i) + Y_{\phi(i)} h_i]$ for $i = 1, \ldots, n$. Similarly, we have

$$\begin{aligned} \sigma_z^2 &\equiv E(Z - EZ)^2 = E(Z - (\bar{Y} + B))^2 \\ &= (1/n)\sum_i (Y_i - \bar{Y})^2 p_i + (1/n)\sum_{j \neq i} \\ &\quad (Y_j - \bar{Y})^2 q_{ij} - 2B\,E(Z - \bar{Y}) + B^2 \\ &= (1/n)S_{yy} + B_{yy} - B^2 = \sigma_y^2 + B_{yy} - B^2, \end{aligned} \tag{A.1.2}$$

where $B_{yy} = (1/n)\sum_i [(Y_i - \bar{Y})^2(-h_i) + (Y_{\phi(i)} - \bar{Y})^2 h_i]$, $S_{yy} = \sum_i (Y_i - \bar{Y})^2$ and $\sigma_y^2 = (1/n)S_{yy}$.

$$\begin{aligned} \sigma_{zx} &\equiv E[(Z - EZ)(X - EX)] \\ &= (1/n)\sum_i (Y_i - \bar{Y})(X_i - \bar{X})p_i \\ &\quad + (1/n)\sum_{j \neq i} (Y_j - \bar{Y})(X_i - \bar{X})\, q_{ij} \\ &= (1/n)S_{yx} + B_{yx} = \sigma_{yx} + B_{yx}, \end{aligned} \tag{A.1.3}$$

where $B_{yx} = (1/n)\sum_i[(Y_i - \bar{Y})(X_i - \bar{X})(-h_i) + (Y_{\phi(i)} - \bar{Y})(X_i - \bar{X})h_i]$, $S_{yx} = \sum_i (Y_i - \bar{Y})(X_i - \bar{X})$, and $\sigma_{yx} = (1/n)S_{yx}$. The term B_{yy} is the bias for the second moments and the term B_{yx} is the bias for the cross-product of Y and X. Formulas (A.1.1), (A.1.2), and (A.1.3), respectively, correspond to formulas (A.1), (A.2), and (A.3) in Neter *et al.* The formulas necessarily differ in detail because we use a more general model of the matching process.

The regression coefficients are related by

$$\beta_{zx} \equiv \sigma_{zx}/\sigma_x^2 = \sigma_{yx}/\sigma_x^2 + B_{yx}/\sigma_x^2 = \beta_{yx} + B_{yx}/\sigma_x^2. \quad \text{(A.1.4)}$$

To get an estimate of the variance of β_{yx}, we first derive an estimate s^2 for the variance σ^2 in the usual manner.

$$(n-2)\,s^2 = \sum_i (y_i - \hat{y}_i)^2 = S_{yy} + \beta_{yx}\,S_{xy}$$
$$= n\,\sigma_y^2 - n\,\beta_{yx}\,\sigma_x^2. \quad \text{(A.1.5)}$$

Using (A.1.2) and (A.1.3) allows us to express s^2 in terms of the observable quantities σ_z^2 and σ_{zx} and the bias terms B_{yy}, B_{yx}, and B that are computable under our assumptions. The estimated variance of β_{yx} is then computed by the usual formula (*e.g.*, Draper and Smith 1981, 18-20)

$$\text{Var}(\beta_{yx}) = s^2/(n\,\sigma_x^2).$$

We observe that the first equality in (A.1.5) involves the usual regression assumption that the error terms are independent with identical variance.

In the numeric examples of this paper we assumed that the true independent value X_i associated with each Y_i was from the record with the highest matching weight and the false independent value was taken from the record with the second highest matching weight. This assumption is plausible because we have only addressed simple regression in this paper and because the second highest matching weight was typically much lower than the highest. Thus, it is much more natural to assume that the record with the second highest matching weight is false. In our empirical examples we use straightforward adjustments and make simplistic assumptions that work well because they are consistent with the data and the matching process. In more complicated regression situations or with other models such as loglinear we will likely have to make additional modelling assumptions. The additional assumptions can be likened to the manner in which simple models for nonresponse require additional assumptions as the models progress from ignorable to nonignorable (see Rubin 1987).

In this section, we chose to adjust independent x-values and leave dependent y-values as fixed in order to achieve consistency with the reasoning of Neter *et al.* We could have just as easily adjusted dependent y-values leaving x-values as fixed.

A.2. Multiple Regression with Independent Variables from One File and Dependent Variables from the Other File

At this point we pass to the usual matrix notation (*e.g.*, Graybill 1976). Our basic model is

$$Y = X\beta + \epsilon,$$

where Y is a $n \times 1$ array, X is a $n \times p$ array, β is a $p \times 1$ array, and ϵ is a $n \times 1$ array.

Analogous to the reasoning we used in (A.1.1), we can represent

$$Z = Y + B, \quad \text{(A.2.1)}$$

where Z, Y, and B are $n \times 1$ arrays having terms that correspond, for $i = 1, \ldots, n$, via

$$z_i = y_i + p_i y_i + h_i y_{\phi(i)}.$$

Because we observe Z and X only, we consider the equation

$$Z = XC + \epsilon. \quad \text{(A.2.2)}$$

We obtain an estimate $\hat{C}$ by regressing on the observed data in the usual manner. We wish to adjust the estimate $\hat{C}$ to an estimate $\hat{\beta}$ of β in a manner analogous to (A.1.1).

Using (A.2.1) and (A.2.2) we obtain

$$(X^TX)^{-1}X^TY + (X^TX)^{-1}XB = \hat{C}. \quad \text{(A.2.3)}$$

The first term on the left hand side of (A.2.3) is the usual estimate $\hat{\beta}$. The second term on the left hand side of (A.2.3) is our bias adjustment. X^T is the transpose of X.

The usual formula (Graybill 1976, p. 176) allows estimation of the variance σ^2 associated with the i.i.d. error components of ϵ,

$$(n-p)\,\hat{\sigma}^2 = (Y - X\beta)^T(Y - X\beta)$$
$$= Y^TY - \hat{\beta}\,X^TY, \quad \text{(A.2.4)}$$

where $\hat{\beta} = (X^TX)^{-1}X^TY$.

Via (A.2.1) $\hat{\beta}\,X^TY$ can be represented in terms of the observable Z and X in a manner similar to (A.1.2) and (A.1.3). As

$$Y^TY = Z^TZ - B^TZ - Z^TB + B^TB, \quad \text{(A.2.5)}$$

we can obtain the remaining portion of the right hand side of (A.2.4) that allows estimation of σ^2.

Via the usual formula (*e.g.*, Graybill 1976, p. 276), the covariance of $\hat{\beta}$ is

$$\text{cov}[\hat{\beta}] = \sigma^2\,(X^TX)^{-1}, \quad \text{(A.2.6)}$$

which we can estimate.

A.3. Multiple Regression with Independent Variables from Both Files

When some of the independent variables come from the same file as Y we must adjust them in a manner similar to the way in which we adjust Y in equations (A.1.1) and (A.2.1). Then data array X can be written in the form

$$X_d = X + D, \qquad \text{(A.3.1)}$$

where D is the array of bias adjustments taking those terms of X arising from the same file as Y back to their true values that are represented in X_d. Using (A.2.1) and (A.2.2), we obtain

$$Y + B = (X_d - D)C. \qquad \text{(A.3.2)}$$

With algebra (A.3.2) becomes

$$\begin{aligned}(X_d^T X_d)^{-1} X_d^T Y &= (X_d^T X_d)^{-1} X_d^T(-B) \\ &\quad + (X_d^T X_d)^{-1} X_d^T (X_d + D)C \\ &= (X_d^T X_d)^{-1} X_d^T(-B) \\ &\quad + (X_d^T X_d)^{-1} X_d^T DC + C. \qquad \text{(A.3.3)}\end{aligned}$$

If D is zero (*i.e.*, all independent x-values arise from a single file), then (A.3.3) agrees with (A.2.3). The first term on the left hand side of (A.2.3) is the estimate of $\hat{\beta}$. The estimate $\hat{\sigma}^2$ is obtained analogously to the way (A.2.3), (A.2.4) and (A.2.5) were used. The covariance of $\hat{\beta}$ follows from (A.2.6).

A.4. Rubin-Belin Model

To estimate the probabilty of a true link within any weight range, Rubin and Belin (1991) consider the set of pairs that are produced by the computer matching program and that are ranked by decreasing weight. They assume that the probability of a true link is a montone function of the weight; that is, the higher the weight, the higher the probability of a true link. They assume that the distribution of the observed weights is a mixture of the distributions for true links and true nonlinks.

Their estimation procedure is:

1. Model each of the two components of the mixture as normal with unknown mean and variance after separate power transformations.
2. Estimate the power of the two transformations from a training sample.
3. Taking the two transformations as known, fit a normal mixture model to the current weight data to obtain maximum likelihood estimates (and standard errors).
4. Use the parameters from the fitted model to obtain point estimates of the false-link rate as a function of cutoff level and obtain standard errors for the false-link rate using the delta-method approximation.

While the Rubin-Belin method requires a training sample, the training sample is primarily used to get the shape of the curves. That is, if the power transformation is given by

$$\psi(w_i; \delta, \omega) = \begin{cases} (w_i^\delta - 1)/(\delta\, \omega^{\delta-1}) & \text{if } \delta \neq 0 \\ \omega \log(w_i) & \text{if } \delta = 0, \end{cases}$$

where ω is the geometric mean of the weights w_i, $i = 1, \ldots, n$, then ω and δ can be estimated for the two curves. For the examples of this paper and a large class of other matching situations (Winkler and Thibaudeau 1991), the Rubin-Belin estimation procedure works well. In some other situations a different method (Winkler 1992) that uses more information than the Rubin-Belin method and does not require a training sample yields accurate estimates, while software (see *e.g.*, Belin 1991) based on the Rubin-Belin method fails to converge even if new calibration data are obtained. Because the calibration data for the good and mediocre scenarios of this paper are appropriate, the Rubin-Belin method provides better estimates than the method of Winkler.

REFERENCES

BEEBE, G. W. (1985). Why are epidemiologists interested in matching algorithms? In *Record Linkage Techniques - 1985*. U.S. Internal Revenue Service.

BELIN, T. (1991). Using Mixture Models to Calibrate Error Rates in Record Linkage Procedures, with Application to Computer Matching for Census Undercount Estimation. Harvard Ph.D. Thesis.

CARPENTER, M., and FAIR, M.E. (Editors) (1989). *Proceedings of the Record Linkage Sessions and Workshop*, Canadian Epidemiological Research Conference, Statistics Canada.

COOMBS, J.W., and SINGH, M.P. (Editors) (1987). *Proceedings: Symposium on Statistical Uses of Administrative Data*, Statistics Canada.

COPAS, J.B., and HILTON, F.J. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society A*, 153, 287-320.

CZJAKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., and RUBIN, D.B. (1992). Evaluation of a new procedure for estimating income and tax aggregrates from advance data. *Journal of Business and Economic Statistics*, 10, 117-131.

DRAPER, N.R., and SMITH, H. (1981). *Applied Regression Analysis*, 2nd Edition. New York: J. Wiley.

FELLEGI, I.P., and SUNTER, A. (1969). A theory of record linkage. *Journal of the of the American Statistical Association*, 64, 1183-1210.

GRAYBILL, F.A. (1976). *Theory and Application of the Linear Model*. Belmont, CA: Wadsworth.

HOWE, G., and SPASOFF, R.A. (Editors) (1986). *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto, Ontario, Canada: University of Toronto Press.

JABINE, T.B., and SCHEUREN, F.J. (1986). Record linkages for statistical purposes: methodological issues. *Journal of Official Statistics*, 2, 255-277.

JARO, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.

JOHNSTON, J. (1972). *Econometric Methods*, 2nd Edition. New York: McGraw-Hill.

KILSS, B., and ALVEY, W. (Editors) (1985). *Record Linkage Techniques - 1985*. U.S. Internal Revenue Service, Publication 1299, 2-86.

NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business*. Oxford: Oxford University Press.

NEWCOMBE, H.B., FAIR, M.E., and LALONDE, P. (1992). The use of names for linking personal records. *Journal of the Americal Statistical Association*, 87, 1193-1208.

NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., and JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.

NETER, J., MAYNES, E.S., and RAMANATHAN, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.

ROSENBAUM, P., and RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

ROSENBAUM, P., and RUBIN, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39, 33-38.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley.

RUBIN, D.B. (1990). Discussion (of Imputation Session). *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, 676-678.

RUBIN, D., and BELIN, T. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 657-668.

SCHEUREN, F. (1985). Methodologic issues in linkage of multiple data bases. *Record Linkage Techniques - 1985*. U.S. Internal Revenue Service.

SCHEUREN, F., and OH, H.L. (1975). Fiddling Around with Nonmatches and Mismatches. *Proceedings of the Social Statistics Section, American Statistical Association*, 627-633.

SCHEUREN, F., OH, H.L., VOGEL, L., and YUSKAVAGE, R. (1981). Methods of Estimation for the 1973 Exact Match Study. Studies from Interagency Data Linkages, U.S. Department of Health and Human Services, Social Security Administration, Publication 13-11750.

TEPPING, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.

WINKLER, W.E. (1985). Exact matching list of businesses: blocking, subfield identification, and information theory. In *Record Linkage Techniques - 1985*, (Eds. B. Kilss and W. Alvey). U.S. Internal Revenue Service, Publication 1299, 2-86.

WINKLER, W.E. (1992). Comparative analysis of record linkage decision rules. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear.

WINKLER, W.E., and SCHEUREN, F. (1991). How Computer Matching Error Effects Regression Analysis: Exploratory and Confirmatory Analysis. U.S. Bureau of the Census, Statistical Research Division Technical Report.

WINKLER, W.E., and THIBAUDEAU, Y. (1991). An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census. U.S. Bureau of the Census, Statistical Research Division Technical Report.

Chapter 4

Regression Analysis of Data Files that are Computer Matched – Part II*

Fritz Scheuren, Ernst and Young, LLP
William E. Winkler, Bureau of the Census

Abstract

Many policy decisions are best made when there is supporting statistical evidence based on analyses of appropriate microdata. Sometimes all the needed data exist but reside in multiple files for which common identifiers (e.g., SIN's, EIN's, or SSN's) are unavailable. This paper demonstrates a methodology for analyzing two such files: when there is common nonunique information subject to significant error and when each source file contains noncommon quantitative data that can be connected with appropriate models. Such a situation might arise with files of businesses only having difficult-to-use name and address information in common, one file with the energy products consumed by the companies, and the other file containing the types and amounts of goods they produce. Another situation might arise with files on individuals in which one file has earnings data, another information about health-related expenses, and a third information about receipts of supplemental payments. The goal of the methodology presented is to produce valid statistical analyses; appropriate microdata files may or may not be produced.

Introduction

Application Setting

To model the energy economy properly, an economist might need company-specific microdata on the fuel and feedstocks used by companies that are only available from Agency A and corresponding microdata on the goods produced for companies that is only available from Agency B. To model the health of individuals in society, a demographer or health science policy worker might need individual-specific information on those receiving social benefits from Agencies B1, B2, and B3, corresponding income information from Agency I, and information on health services from Agencies H1 and H2. Such modeling is possible if analysts have access to the microdata and if unique, common identifiers are available (*e.g.*, Oh and Scheuren, 1975; Jabine and Scheuren, 1986). If the only common identifiers are error-prone or nonunique or both, then probabilistic matching techniques (*e.g.*, Newcombe et al., 1959, Fellegi and Sunter, 1969) are needed.

Relation To Earlier Work

In earlier work (Scheuren and Winkler, 1993), we provided theory showing that elementary regression analyses could be accurately adjusted for matching error, employing knowledge of the quality of the matching. In that work, we relied heavily on an error-rate estimation procedure of Belin and Rubin (1995). In later research (Winkler and Scheuren, 1995, 1996), we showed that we could make further improvements by using noncommon quantitative data from the two files to improve matching and adjust statistical analyses for matching error. The main requirement -- even in heretofore seemingly impossible situations -- was that there

*Reprinted with permission. To appear in *Survey Methodology* (1997), 23, 2.

exist a reasonable model for the relationships

among the noncommon quantitative data. In the empirical example of this paper, we use data for which a very small subset of pairs can be accurately matched using name and address information only and for which the noncommon quantitative data is at least moderately correlated. In other situations, researchers might have a small microdata set that accurately represents relationships of noncommon data across a set of large administrative files or they might just have a reasonable guess at what the relationships among the noncommon data are. We are not sure, but conjecture that, with a reasonable starting point, the methods discussed here will succeed often enough to be of general value.

Basic Approach

The intuitive underpinnings of our methods are based on now well-known probabilistic record linkage (**RL**) and edit/imputation (**EI**) technologies. The ideas of modern **RL** were introduced by Newcombe (Newcombe *et al.*, 1959) and mathematically formalized by Fellegi and Sunter (1969). Recent methods are described in Winkler (1994, 1995). **EI** has traditionally been used to clean up erroneous data in files. The most pertinent methods are based on the **EI** model of Fellegi and Holt (1976).

To adjust a statistical analysis for matching error, we employ a four-step recursive approach that is very powerful. We begin with an enhanced **RL** approach (*e.g.*, Winkler, 1994; Belin and Rubin, 1995) to delineate a subset of pairs of records in which the matching error rate is estimated to be very low. We perform a regression analysis, **RA,** on the low-error-rate linked records and partially adjust the regression model on the remainder of the pairs by applying previous methods (Scheuren and Winkler, 1993). Then, we refine the **EI** model using traditional outlier-detection methods to edit and impute outliers in the remainder of the linked pairs. Another regression analysis (**RA**) is done and this time the results are fed back into the linkage step so that the **RL** step can be improved (and so on). The cycle continues until the analytic results desired cease to change. Schematically, these analytic linking methods take the form

Structure of What Follows

Beginning with this introduction, the paper is divided into five sections. In the second section, we undertake a short review of Edit/Imputation (**EI**) and Record Linkage (**RL**) methods. Our purpose is not to describe them in detail but simply to set the stage for the present application. Because Regression Analysis (**RA**) is so well known, our treatment of it is covered only in the particular simulated application (Section 3). The intent of these simulations is to use matching scenarios that are more difficult than what most linkers typically encounter. Simultaneously, we employ quantitative data that is both easy to understand but hard to use in matching. In the fourth section, we present results. The final section consists of some conclusions and areas for future study.

EI and RL Methods Reviewed

Edit/Imputation

Methods of **editing** microdata have traditionally dealt with logical inconsistencies in data bases. Software consisted of **if-then-else** rules that were data-base-specific and very difficult to maintain or modify, so as to keep current. Imputation methods were part of the set of **if-then-else** rules and could yield revised records that still failed edits. In a major theoretical advance that broke with prior statistical methods, Fellegi and Holt (1976)

introduced operations-research-based methods that both provided a means of checking the logical consistency of an edit system and assured that an edit-failing record could always be updated with imputed values, so that the revised record satisfies all edits. An additional advantage of Fellegi-Holt systems is that their edit methods tie directly with current methods of **imputing** microdata (*e.g.*, Little and Rubin 1987).

Although we will only consider continuous data in this paper, **EI** techniques also hold for discrete data and combinations of discrete and continuous data. In any event, suppose we have continuous data. In this case a collection of edits might consist of rules for each record of the form

$$\mathbf{c_1X < Y < c_2X} \ .$$

In words,

> **Y can be expected to be greater than c_1X and less than c_2X; hence, if Y less than c_1X and greater than c_2X, then the data record should be reviewed (with resource and other practical considerations determining the actual bounds used).**

Here **Y** may be total wages, **X** the number of employees, and $\mathbf{c_1}$ and $\mathbf{c_2}$ constants such that $\mathbf{c_1 < c_2}$. When an (**X, Y**) pair associated with a record fails an edit, we may replace, say, Y with an estimate (or prediction).

Record Linkage

A record linkage process attempts to classify pairs in a product space **A H B** from two files **A** and **B** into **M**, the set of true links, and **U**, the set of true nonlinks. Making rigorous concepts introduced by Newcombe (*e.g.*, Newcombe *et al.*, 1959; Newcombe *et al,* 1992), Fellegi and Sunter (1969) considered ratios **R** of probabilities of the form

$$\mathbf{R = Pr((\gamma, \Gamma) \mid M) / Pr((\gamma, \Gamma) \mid U)}$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on surname, first name, and age. Alternatively, each (γ, Γ) might additionally account for the relative frequency with which specific surnames, such as Scheuren or Winkler, occur. The fields compared (surname, first name, age) are called *matching variables*. The decision rule is given by

> **If R > *Upper*, then designate pair as a link.**
>
> **If *Lower* # R # *Upper*, then designate pair as a possible link and hold for clerical review.**
>
> **If R < *Lower*, then designate pair as a nonlink.**

Fellegi and Sunter (1969) showed that this decision rule is optimal in the sense that for any pair of fixed bounds on **R**, the middle region is minimized over all decision rules on the same comparison space Γ. The cutoff thresholds, *Upper* and *Lower*, are determined by the error bounds. We call the ratio **R** or any monotonely increasing transformation of it (typically a logarithm) a *matching weight* or *total agreement weight*.

With the availability of inexpensive computing power, there has been an outpouring of new work on record linkage techniques (*e.g.*, Jaro, 1989, Newcombe, Fair, Lalonde, 1992, Winkler, 1994, 1995). The new computer-intensive methods reduce, or even sometimes eliminate, the need for clerical review when name, address, and other information used in matching is of reasonable quality. The proceedings from a recently concluded international conference on record linkage showcases these ideas and might be the best single reference (Alvey and Jamerson, 1997).

Simulation Setting

Matching Scenarios

For our simulations, we considered a scenario in which matches are virtually indistinguishable from nonmatches. In our earlier work (Scheuren and Winkler, 1993), we considered three matching scenarios in which matches are more easily distinguished from nonmatches than in the scenario of the present paper.

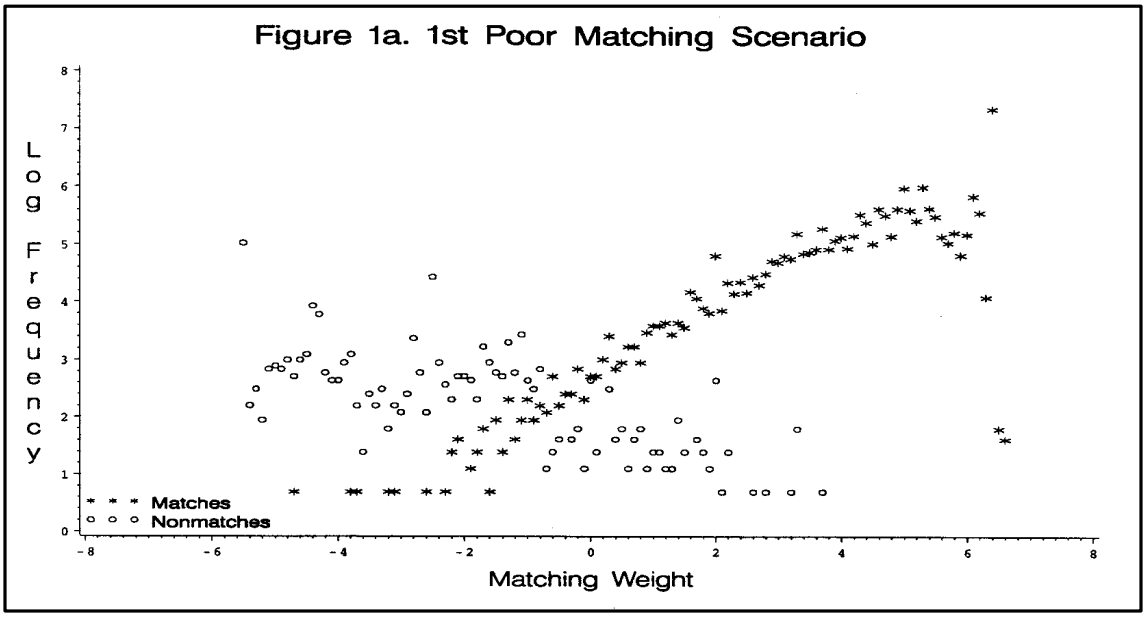
In both papers, the basic idea is to generate data having known distributional properties, adjoin the data to two files that would be matched, and then to evaluate the effect of increasing amounts of matching error on analyses. Because the methods of this paper work better than what we did earlier, we only consider a matching scenario that we label "Second Poor," because it is more difficult than the poor (most difficult) scenario we considered previously.

We started here with two population files (sizes 12,000 and 15,000), each having good matching information and for which true match status was known. The settings were examined: high, medium and low -- depending on the extent to which the smaller file had cases also included in the larger file. In the high file inclusion situation, about 10,000 cases are on both files for an file inclusion or intersection rate on the smaller or base file of about 83%. In the medium file intersection situation, we took a sample of one file so that the intersection of the two files being matched was approximately 25%. In the low file intersection situation, we took samples of both files so that the intersection of the files being matched was approximately 5%. The number of intersecting cases, obviously, bounds the number of true matches that can be found.

We then generated quantitative data with known distributional properties and adjoined the data to the files. These variations are described below and displayed in Figure 1 where we show the poor scenario (labeled "first poor") of our previous 1993 paper and the "second poor" scenario used in this paper. In the figure, the match weight, the logarithm of **R**, is plotted on the horizontal axis with the frequency, also expressed in logs, plotted on the vertical axis. Matches (or true links) appear as asterisks (*), while nonmatches (or true nonlinks) appear as small circles (o).

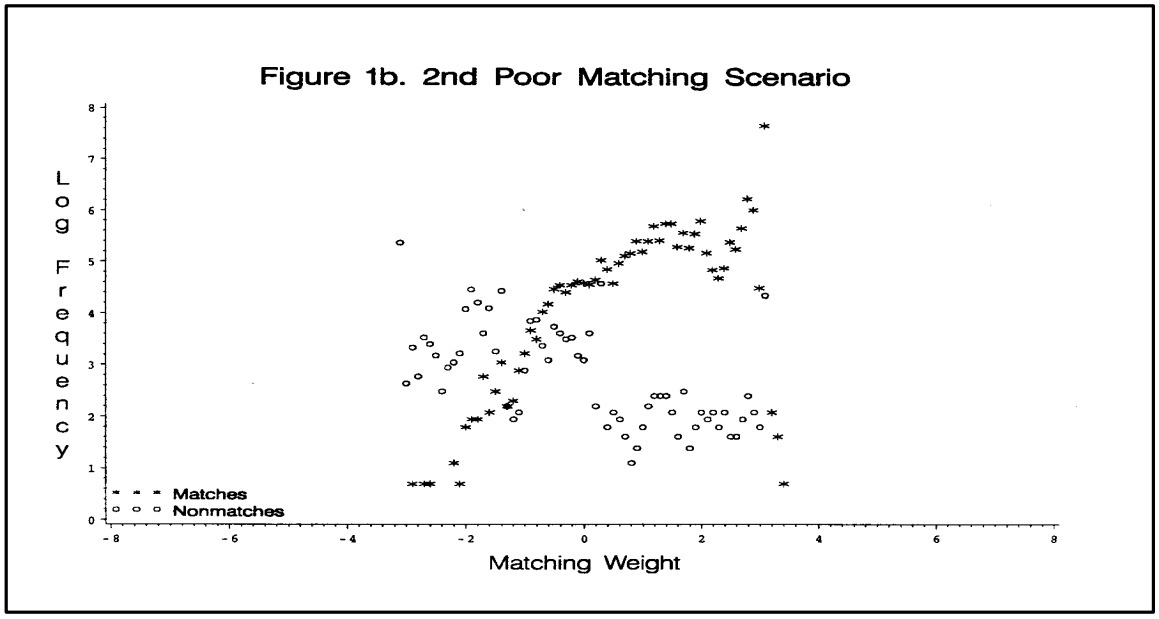
"First Poor" Scenario (Figure 1a)

The first poor matching scenario consisted of using last name, first name, one address variation, and age. Minor typographical errors were introduced independently into one fifth of the last names and one third of the first names in one of the files. Moderately severe typographical errors were made independently in one fourth of the addresses of the same file. Matching probabilities were chosen that deviated substantially from optimal. The intent was for the links to be made in a manner that a practitioner might choose after gaining only a little experience. The situation is analogous to that of using administrative lists of individuals where information used in matching is of poor quality. The true mismatch rate here was 10.1%.



"Second Poor" Scenario (Figure 1b)

The second poor matching scenario consisted of using last name, first name, and one address variation. Minor typographical errors were introduced independently into one third of the last names and one third of the first names in one of the files. Severe typographical errors were made in one fourth of the addresses in the same file. Matching probabilities were chosen that deviated substantially from optimal. The intent was to represent situations that often occur with lists of businesses in which the linker has little control over the quality of the lists. Name information -- a key identifying characteristic -- is often very difficult to compare effectively with business lists. The true mismatch rate was 14.6%.



Summary of Matching Scenarios

Clearly, depending on the scenario, our ability to distinguish between true links and true nonlinks differs significantly. With the first poor scenario, the overlap, shown visually between the log-frequency-versus-weight curves is substantial (Figure 1a); and, with the second poor scheme, the overlap of the log-frequency-versus-weight curves is almost total (Figure 1b). In the earlier work, we showed that our theoretical adjustment procedure worked well using the known true match rates in our data sets. For situations where the curves of true links and true nonlinks were reasonably well separated, we accurately estimated error rates via a procedure of Belin and Rubin (1995) and our procedure could be used in practice. In the poor matching scenario of that paper (first poor scenario of this paper), the Belin-Rubin procedure was unable to provide accurate estimates of error rates but our theoretical adjustment procedure still worked well. This indicated that we either had to find an enhancement to the Belin-Rubin procedures or to develop methods that used more of the available data. (That conclusion, incidentally, from our earlier work led, after some false starts, to the present approach.)

Quantitative Scenarios

Having specified the above linkage situations, we used SAS to generate ordinary least squares data under the model $\mathbf{Y} = 6\,\mathbf{X} + \varepsilon, \varepsilon$. The $\mathbf{X}$ values were chosen to be uniformly distributed between 1 and 101. The error terms ε are normal and homoscedastic with variances 13000, 36000, and 125000, respectively. The resulting regressions of $\mathbf{Y}$ on $\mathbf{X}$ have $\mathbf{R}^2$ values in the true matched population of 70%, 47%, and 20%, respectively. Matching with quantitative data is difficult because, for each record in one file, there are hundreds of records having quantitative values that are close to the record that is a true match. To make modeling and analysis even more difficult in the high file overlap scenario, we used all false matches and only 5% of the true matches; in the medium file overlap scenario, we used all false matches and only 25% of true matches. (Note: Here to heighten the visual effect, we have introduced another random sampling step, so the reader can "see" better in the figures the effect of bad matching. This sample depends on the match status of the case and is confined only to those cases that were matched, whether correctly or falsely.)

A crucial practical assumption for the work of this paper is that analysts are able to produce a reasonable model (guesstimate) for the relationships between the noncommon quantitative items. For the initial modeling in the empirical example of this paper, we use the subset of pairs for which matching weight is high and the error-rate is low. Thus, the number of false matches in the subset is kept to a minimum. Although neither the procedure of Belin and Rubin (1995) nor an alternative procedure of Winkler (1994), that requires an *ad hoc* intervention, could be used to estimate error rates, we believe it is possible for an experienced matcher to pick out a low-error-rate set of pairs even in the second poor scenario.

Simulation Results

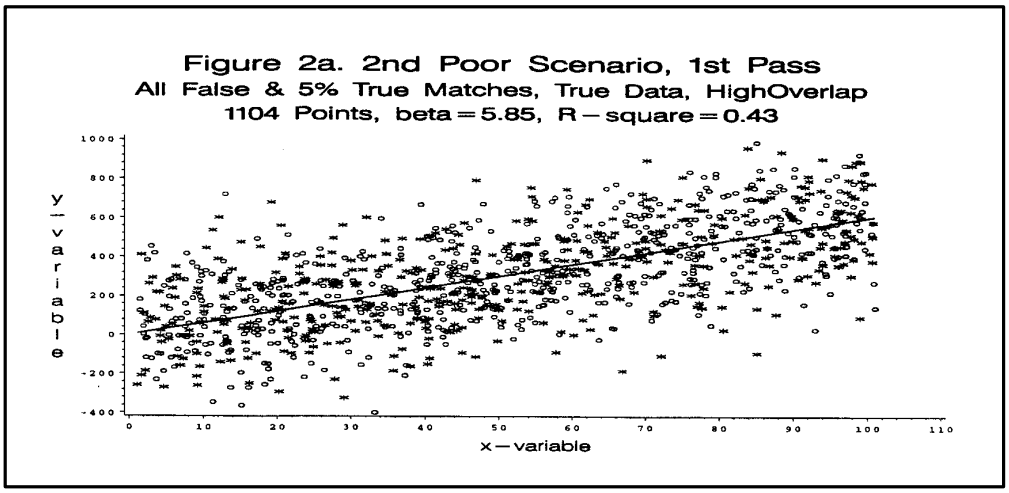
Most of this Section is devoted to presenting graphs and results of the overall process for the second poor scenario, where the $\mathbf{R}^2$ value is moderate, and the intersection between the two files is high. These results best illustrate the procedures of this paper. At the end of the Section (in subsection 4.8), we summarize results over all $\mathbf{R}^2$ situations and all overlaps. To make the modeling more difficult and show the power of the analytic linking methods, we use all false matches and a random sample of only 5% of the true matches. We only consider pairs having matching weight above a lower bound that we determine based on analytic considerations and experience. For the pairs of our analysis, the restriction causes the number of false matches to significantly exceed the number of true matches. (Again, this is done to heighten the visual effect of matching failures and to make the problem even more difficult.)

To illustrate the data situation and the modeling approach, we provide triples of plots. The first plot in the triple shows the true data situation as if each record in one file was linked with its true corresponding record in the other file. The quantitative data pairs correspond to the truth. In the second plot, we show the observed

data. A high proportion of the pairs is in error because they correspond to false matches. To get to the third plot in the triple, we model using a small number of pairs (approximately 100) and then replace outliers with pairs in which the observed **Y**-value is replaced with a predicted **Y**-value.

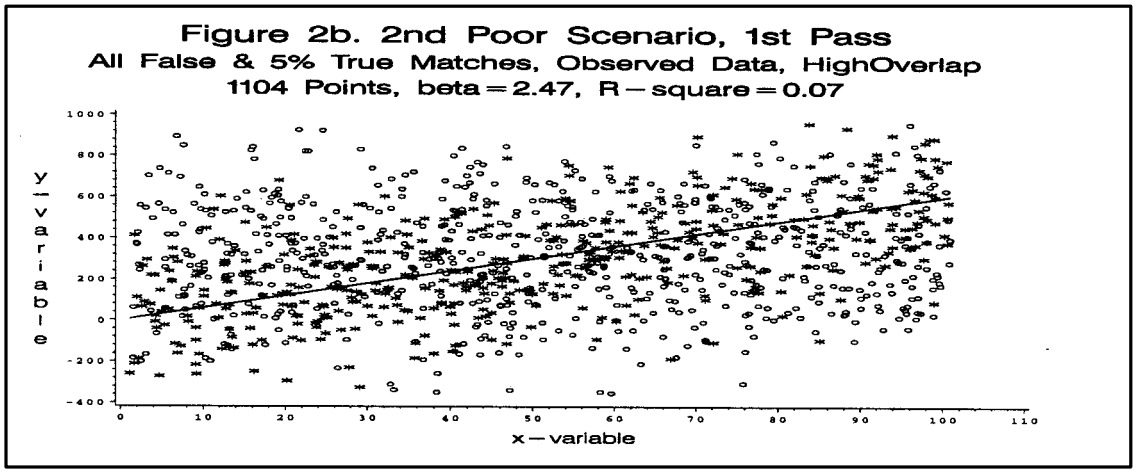
Initial True Regression Relationship

In Figure 2a, the actual true regression relationship and related scatterplot are shown, for one of our simulations, as they would appear if there were no matching errors. In this figure and the remaining ones, the true regression line is always given for reference. Finally, the true population slope or **beta** coefficient (at 5.85) and the R^2 value (at 43%) are provided for the data (sample of pairs) being displayed.



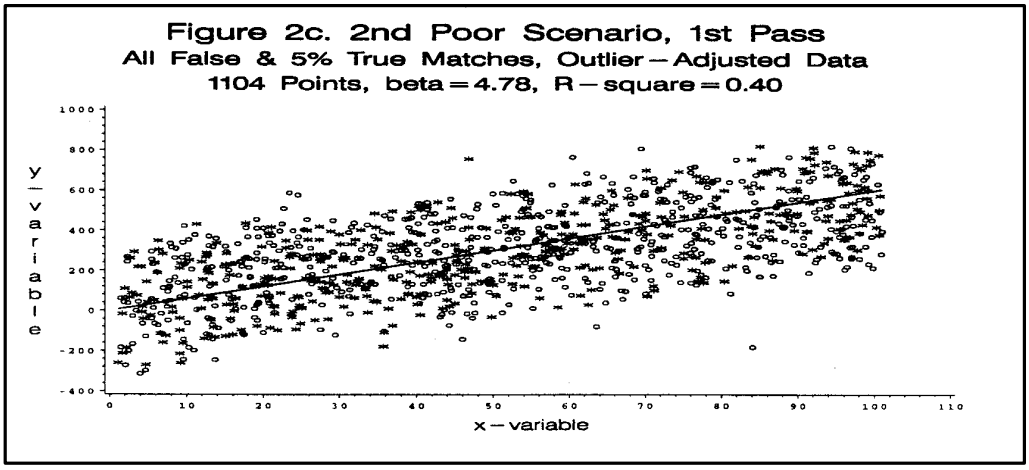
Regression after Initial RL Þ RA Step

In Figure 2b, we are looking at the regression on the actual observed links -- not what should have happened in a perfect world but what did happen in a very imperfect one. Unsurprisingly, we see only a weak regression relationship between **Y** and **X**. The observed slope or **beta** coefficient differs greatly from its true value (2.47 v. 5.85). The fit measure is similarly affected -- falling to 7% from 43%.



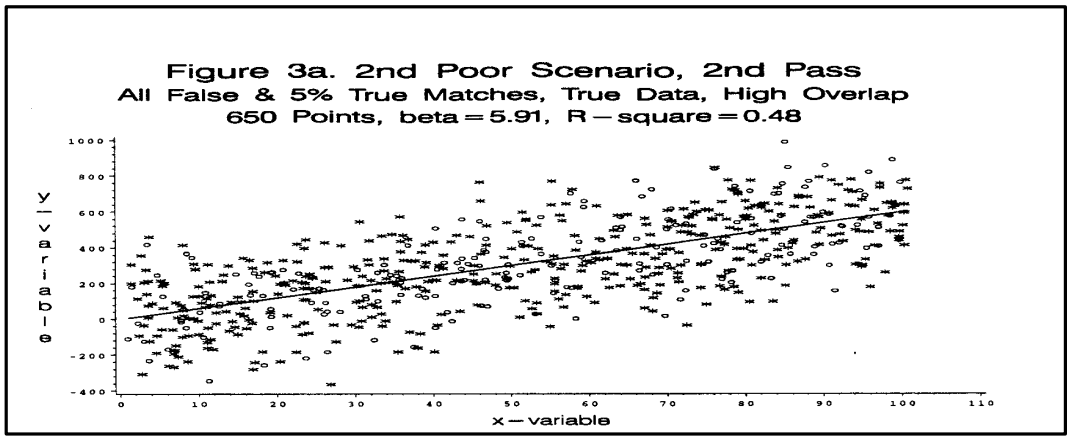
Regression After First Combined RL Þ RA Þ EI Þ RA Step

Figure 2c completes our display of the first cycle of the iterative process we are employing. Here we have edited the data in the plot displayed as follows. First, using just the 99 cases with a match weight of 3.00 or larger, an attempt was made to improve the poor results given in figure 2b. Using this provisional fit, predicted values were obtained for all the matched cases; then outliers with residuals of 460 or more were removed and the regression refit on the remaining pairs. This new equation, used in figure 2c, was essentially $\mathbf{Y} = 4.78\mathbf{X} + \mathbf{,}$ **,** with a variance of 40000. Using our earlier approach (Scheuren and Winkler, 1993), a further adjustment was made in the estimated **beta** coefficient from 4.78 to 5.4. If a pair of matched records yielded an outlier, then predicted values (not shown) using the equation $\mathbf{Y} = 5.4\mathbf{X}$ were imputed. If a pair does not yield an outlier, then the observed value was used as the predicted value.



Second True Reference Regression

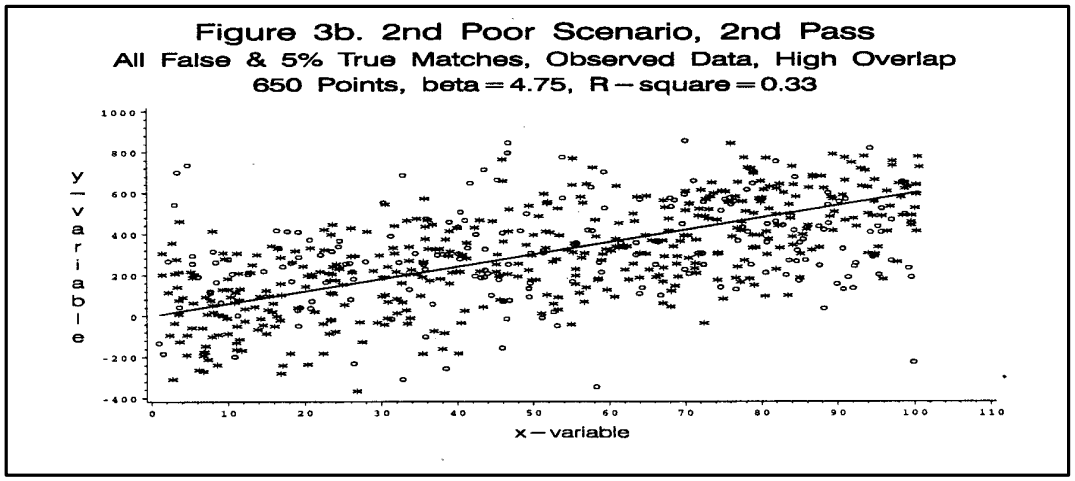
Figure 3a displays a scatterplot of **X** and **Y** as they would appear if they could be true matches based on a second **RL** step. Note here that we have a somewhat different set of linked pairs this time from earlier, because we have used the regression results to help in the linkage. In particular, the second **RL** step employed the predicted **Y** values as determined above; hence it had more information on which to base a linkage. This meant that a different group of linked records was available after the second **RL** step. Since a considerably better link was obtained, there



were fewer false matches; hence our sample of all false matches and 5% of the true matches dropped from 1104 in Figures 2a thru 2c to 650 for Figures 3a thru 3c. In this second iteration, the true slope or **beta** coefficient and the $\mathbf{R^2}$ values remained, though, virtually identical for the estimated slope (5.85 v. 5.91) and fit (43% v. 48%).

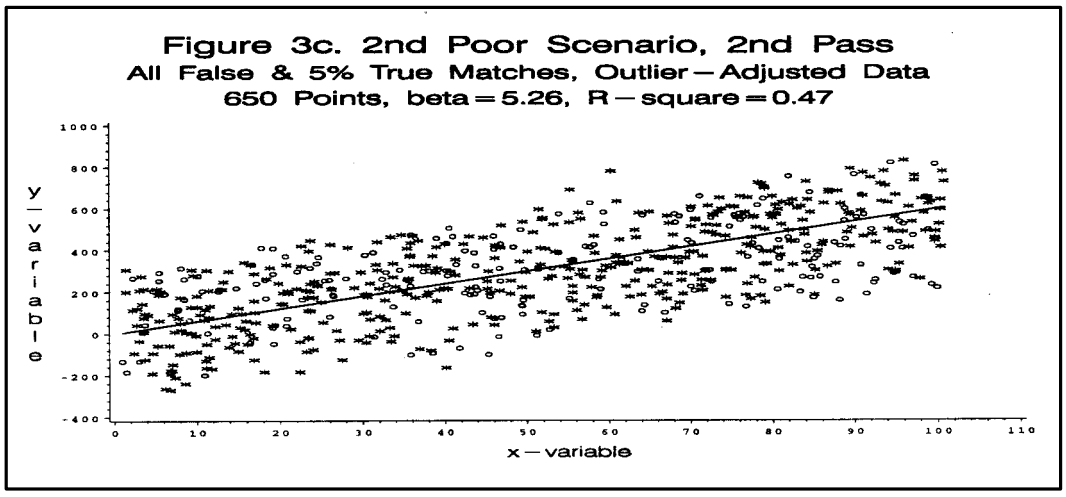
Regression After Second RL Þ RA Step

In Figure 3b, we see a considerable improvement in the relationship between **Y** and **X** using the actual observed links after the second **RL** step. The estimated slope has risen from 2.47 initially to 4.75 here. Still too small but much improved. The fit has been similarly affected, rising from 7% to 33%.



Regression After Second Combined RL Þ RA Þ EI Þ RA Step

Figure 3c completes the display of the second cycle of our iterative process. Here we have edited the data as follows. Using the fit (from subsection 4.5), another set of predicted values was obtained for all the matched cases (as in subsection 4.3). This new equation was essentially $\mathbf{Y} = 5.26\mathbf{X} +$ **, ,** with a variance of about 35000. If a pair of matched records yields an outlier, then predicted values using the equation $\mathbf{Y} = 5.3\mathbf{X}$ were imputed. If a pair does not yield an outlier, then the observed value was used as the predicted value.



Additional Iterations

While we did not show it in this paper, we did iterate through a third matching pass. The **beta** coefficient, after adjustment, did not change much. We do not conclude from this that asymptotic unbiasedness exists; rather that the method, as it has evolved so far, has a positive benefit and that this benefit may be quickly reached.

Further Results

Our further results are of two kinds. We looked first at what happened in the medium $\mathbf{R}^2$ scenario (i.e., $\mathbf{R}^2$ equal to .47) for the medium- and low- file intersection situations. We further looked at the cases when $\mathbf{R}^2$ was higher (at .70) or lower (at .20). For the medium $\mathbf{R}^2$ scenario and low intersection case the matching was somewhat easier. This occurs because there were significantly fewer false-match candidates and we could more easily separate true matches from false matches. For the high $\mathbf{R}^2$ scenarios, the modeling and matching were also more straightforward than there were for the medium $\mathbf{R}^2$ scenario. Hence, there were no new issues there either.

On the other hand, for the low $\mathbf{R}^2$ scenario, no matter what degree of file intersection existed, we were unable to distinguish true matches from false matches, even with the improved methods we are using. The reason for this, we believe, is that there are many outliers associated with the true matches. We can no longer assume, therefore, that a moderately higher percentage of the outliers in the regression model are due to false matches. In fact, with each true match that is associated with an outlier **Y**-value, there may be many false matches that have **Y**-values that are closer to the predicted **Y**-value than the true match.

Comments and Future Study

Overall Summary

In this paper, we have looked at a very restricted analysis setting: a simple regression of one quantitative dependent variable from one file matched to a single quantitative independent variable from another file. This standard analysis was, however, approached in a very nonstandard setting. The matching scenarios, in fact, were quite challenging. Indeed, just a few years ago, we might have said that the "second poor" matching scenario appeared hopeless.

On the other hand, as discussed below, there are many loose ends. Hence, the demonstration given here can be considered, quite rightly in our view, as a limited accomplishment. But make no mistake about it, we are doing something entirely new. In past record linkage applications, there was a clear separation between the identifying data and the analysis data. Here, we have used a regression analysis to improve the linkage and the improved linkage to improve the analysis and so on.

Earlier, in our 1993 paper, we advocated that there be a unified approach between the linkage and the analysis. At that point, though, we were only ready to propose that the linkage probabilities be used in the analysis to correct for the failures to complete the matching step satisfactorily. This paper is the first to propose a completely unified methodology and to demonstrate how it might be carried out.

Planned Application

We expect that the first applications of our new methods will be with large business data bases. In such situations, noncommon quantitative data are often moderately or highly correlated and the quantitative variables

(both predicted and observed) can have great distinguishing power for linkage, especially when combined with name information and geographic information, such as a postal (e.g., ZIP) code.

A second observation is also worth making about our results. The work done here points strongly to the need to improve some of the now routine practices for protecting public use files from reidentification. In fact, it turns out that in some settings -- even after quantitative data have been confidentiality protected (by conventional methods) and without any directly identifying variables present -- the methods in this paper can be successful in reidentifying a substantial fraction of records thought to be reasonably secure from this risk (as predicted in Scheuren, 1995). For examples, see Winkler, 1997.

Expected Extensions

What happens when our results are generalized to the multiple regression case? We are working on this now and results are starting to emerge which have given us insight into where further research is required. We speculate that the degree of underlying association $\mathbf{R}^2$ will continue to be the dominant element in whether a usable analysis is possible.

There is also the case of multivariate regression. This problem is harder and will be more of a challenge. Simple multivariate extensions of the univariate comparison of **Y** values in this paper have not worked as well as we would like. For this setting, perhaps, variants and extensions of Little and Rubin (1987, Chapters 6 and 8) will prove to be a good starting point.

"Limited Accomplishment"

Until now an analysis based on the second poor scenario would not have been even remotely sensible. For this reason alone we should be happy with our results. A closer examination, though, shows a number of places where the approach demonstrated is weaker than it needs to be or simply unfinished. For those who want theorems proven, this may be a particularly strong sentiment. For example, a convergence proof is among the important loose ends to be dealt with, even in the simple regression setting. A practical demonstration of our approach with more than two matched files also is necessary, albeit this appears to be more straightforward.

Guiding Practice

We have no ready advise for those who may attempt what we have done. Our own experience, at this point, is insufficient for us to offer ideas on how to guide practice, except the usual extra caution that goes with any new application. Maybe, after our own efforts and those of others have matured, we can offer more.

References

Alvey, W. and Jamerson, B. (eds.) (1997), *Record Linkage Techniques -- 1997* (Proceedings of An International Record Linkage Workshop and Exposition, March 20-21, 1997, in Arlington VA).

Belin, T. R., and Rubin, D. B. (1995). A Method for Calibrating False-Match Rates in Record Linkage, *Journal of the American Statistical Association*, 90, 694-707.

Fellegi, I. and Holt, T. (1976). A Systematic Approach to Automatic Edit and Imputation, *Journal of the of the American Statistical Association*, 71, 17-35.

Fellegi, I. and Sunter, A. (1969). A Theory of Record Linkage, *Journal of the of the American Statistical Association*, 64, 1183-1210.

Jabine, T.B. and Scheuren, F. (1986). Record Linkages for Statistical Purposes: Methodological Issues, *Journal of Official Statistics*, 2, 255-277.

Jaro, M. A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association*, 89, 414-420.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, J. Wiley: New York.

Newcombe, H. B.; Kennedy, J. M.; Axford, S. J.; and James, A. P. (1959). Automatic Linkage of Vital Records, *Science*, 130, 954-959.

Newcombe, H.; Fair, M.; and Lalonde, P. (1992). The Use of Names for Linking Personal Records, *Journal of the American Statistical Association*, 87, 1193-1208.

Oh, H. L. and Scheuren, F. (1975). Fiddling Around with Mismatches and Nonmatches, *Proceedings of the Section on Social Statistics, American Statistical Association.*

Scheuren, F. and Winkler, W. E. (1993). Regression Analysis of Data Files that are Computer Matched, *Survey Methodology*, 19, 39-58.

Scheuren, F. (1995). Review of Private Lives and Public Policies, *Journal of the American Statistical Association*, 90.

Scheuren, F. and Winkler, W. E. (1996). Recursive Merging and Analysis of Administrative Lists and Data, *Proceedings of the Section of Government Statistics, American Statistical Association*, 123-128.

Winkler, W. E. (1994). Advanced Methods of Record Linkage, *Proceedings of the Section of Survey Research Methods, American Statistical Association*, 467-472.

Winkler, W. E. (1995). Matching and Record Linkage, in B. G. Cox *et al*. (ed.), *Business Survey Methods*, New York: J. Wiley, 355-384.

Winkler, W. E. and Scheuren, F. (1995). Linking Data to Create Information, *Proceedings of Statistics Canada Symposium, 95*.

Winkler, W. E. and Scheuren, F. (1996). Recursive Analysis of Linked Data Files, *Proceedings of the 1996 Census Bureau.*

Winkler, W.E. (1997). Producing Public-Use Microdata That are Analytically Valid and Confidential. Paper presented at the 1997 Joint Statistical Meetings in Anaheim.

Chapter 5

Using Record Linkage to Thwart the Data Terrorist

Robert Burton, National Center for Education Statistics

Abstract

The traditional goal of record linkage is to maximize the number of correct matches and minimize the number of incorrect matches. Since this is also the goal of the "data terrorist" seeking to identify subjects from microdata files, it makes sense to use record linkage software to thwart the data terrorist.

There are, however, differences between the two situations. At the conceptual level, the relevant loss functions seem quite different. At the practical level, some modifications may be desirable in regard to blocking, assignment, and recoding when using record linkage methods to test the disclosure potential of microdata files. The speaker will discuss these differences and suggest promising ways to use record linkage methods to protect respondent confidentiality.

Chapter 5

m- and t-ARGUS: Software for Statistical Disclosure Control

Anco J. Hundepool and Leon C. R. J. Willenborg, Statistics Netherlands

Abstract

In recent years, Statistics Netherlands has developed a prototype version of a software package, ARGUS, to protect microdata files against statistical disclosure. The launch of the SDC-project within the 4th framework of the European Union enabled us to make a new start with the development of software for Statistical Disclosure Control (Willenborg, 1996). The prototype has served as a starting point for the development of m-ARGUS, a software package for the SDC of microdata. This SDC-project, however, also plans to develop t-ARGUS, software devoted to the SDC of tabular data. The development of these software packages also benefits from the research of other partners in this project. This paper gives an overview of the development of these software packages and an introduction to the basic ideas behind the implementation of Statistical Disclosure Control at Statistics Netherlands.

Introduction

The growing demands from researchers, policy makers and others for more and more detailed statistical information leads to a conflict. The statistical offices collect large amounts of data for statistical purposes. The respondents are only willing to provide the statistical offices with the required information if they can be certain that these statistical offices will treat their data with the utmost care. This implies that their confidentiality must be guaranteed. This imposes limitations on the amount of detail in the publications. Research has been carried out to establish pragmatic rules to determine which tables can be regarded safe with respect to the protection of the confidentiality of the respondents. The well-known dominance rule is often used.

On the other hand, statistical databases with individual records (microdata files) are valuable sources for research. To a certain extent the statistical offices are prepared to make these microfiles available to researchers, but only under the provision that the information in these databases is sufficiently protected against disclosure. At Statistics Netherlands a lot of research has been carried out to establish rules to determine whether a specific database is safe enough to make it available to researchers. In the next section, we will give an introduction to this research. Then, we will go into the development of μ-Argus and we will conclude with an overview of τ-Argus.

SDC for Microdata at Statistics Netherlands

Re-identification

The aim of statistical disclosure control (SDC) is to limit the risk that sensitive information of individual respondents can be disclosed from a data set (Willenborg and DeWaal, 1996). In case of a microdata set, i.e., a set of records containing information on individual respondents, such disclosure of sensitive

information about an individual respondent can occur after this respondent has been re-identified; that is, after it has been deduced which record corresponds to this particular individual. So, disclosure control should hamper re-identification of individual respondents.

Re-identification can take place when several values of so-called identifying variables, such as "Place of residence," "Sex," and "Occupation" are taken into consideration. The values of these identifying variables can be assumed to be known to friends and acquaintances of a respondent. When several values of these identifying variables are combined, a respondent may be re-identified. Consider for example the following record obtained from an unknown respondent:

"Place of residence = Urk," "Sex = Female" and "Occupation = Statistician."

Urk is a small fishing village in the Netherlands, in which it is unlikely for many statisticians to live, let alone female statisticians. So, when we find a statistician in Urk, a female one moreover, in the microdata set, then she is probably the only one. When this is indeed the case, anybody who happens to know this rare female statistician in Urk is able to disclose sensitive information from her record if it contains such information.

An important concept in the theory of re-identification is a *key*. A key is a combination of identifying variables. Keys can be applied to re-identify a respondent. Re-identification of a respondent can occur when this respondent is rare in the population with respect to a certain key value, i.e., a combination of values of identifying variables. Hence, rarity of respondents in the population with respect to certain key values should be avoided. When a respondent appears to be rare in the population with respect to a key value, then disclosure control measures should be taken to protect this respondent against re-identification (DeWaal and Willenborg, 1995a).

In practice, however, it is not a good idea to prevent only the occurrence of respondents in the data file who are rare in the population (with respect to a certain key). For this, several reasons can be given. Firstly, there is a practical reason: rarity in the population, in contrast to rarity in the data file, is hard to establish. There is generally no way to determine with certainty whether a person who is rare in the data file (with respect to a certain key) is also rare in the population. Secondly, an intruder may use another key than the key(s) considered by the data protector. For instance, the data protector may consider only keys consisting of at most three variables, while the intruder may use a key consisting of four variables. Therefore, it is better to avoid the occurrence of combinations of scores that are *rare* in the population in the data file instead of avoiding only population-uniques in the data file. To define what is meant by rare, the data protector has to choose a threshold value D_k, for each key value k, where the index k indicates that the threshold value may depend on the key k under consideration. A combination of scores, i.e., a key value, that occurs not more than D_k times in the population is considered *unsafe*; a key value that occurs more than D_k times in the population is considered *safe*. The unsafe combinations must be protected, while the safe ones may be published.

There is a practical problem when applying the above rule that the occurrence (in the data file) of combinations of scores that are rare in the population should be avoided. Namely, it is usually not known how often a particular combination of scores occurs in the population. In many cases, one only has the data file itself available to *estimate* the frequency of a combination of scores in the population. In practice, one therefore uses the estimated frequency of a key value k to determine whether or not this key value is safe or not in the population. When the *estimated* frequency of a key value, i.e., a combination of scores, is larger than the threshold value D_k, then this combination is considered *safe*. When the *estimated* frequency of a key value is less than or equal to the threshold value D_k, then this combination is considered *unsafe*. An example of such a key is "Place of residence," "Sex," and "Occupation."

SDC Techniques

Statistics Netherlands, so far, has used two SDC techniques to protect microdata sets, namely global recoding and local suppression. In case of global recoding, several categories of a variable are collapsed into a single one. In the above example, for instance, we can recode the variable "Occupation." For instance, the categories "Statistician" and "Mathematician" can be combined into a single category "Statistician or Mathematician." When the number of female statisticians in Urk plus the number of female mathematicians in Urk is sufficiently high, then the combination "Place of residence = Urk," "Sex = Female," and "Occupation = Statistician or Mathematician" is considered safe for release. Note that instead of recoding "Occupation," one could also recode "Place of residence" for instance.

The concept of MINimum Unsafe Combinations (MINUC) plays an important role in the selection of the variables and the categories for local suppression. A MINUC provides that suppressing any value in the combination yields a safe combination.

It is important to realize that global recoding is applied to the whole data set, not only to the unsafe part of the set. This is done to obtain a uniform categorization of each variable. Suppose, for instance, that we recode "Occupation" in the above way. Suppose furthermore that both the combinations "Place of residence = Amsterdam," "Sex = Female," and "Occupation = Statistician," and "Place of residence = Amsterdam," "Sex = Female," and "Occupation = Mathematician" are considered safe. To obtain a uniform categorization of "Occupation" we would, however, not publish these combinations, but only the combination "Place of residence = Amsterdam," "Sex = Female," and "Occupation = Statistician or Mathematician."

When local suppression is applied, one or more values in an unsafe combination are suppressed, i.e., replaced by a missing value. For instance, in the above example we can protect the unsafe combination "Place of residence = Urk," "Sex = Female" and "Occupation = Statistician" by suppressing the value of "Occupation," assuming that the number of females in Urk is sufficiently high. The resulting combination is then given by "Place of residence = Urk," "Sex = Female," and "Occupation = missing." Note that instead of suppressing the value of "Occupation," one could also suppress the value of another variable of the unsafe combination. For instance, when the number of female statisticians in the Netherlands is sufficiently high then one could suppress the value of "Place of residence" instead of the value of "Occupation" in the above example to protect this unsafe combination. A local suppression is only applied to a particular value. When, for instance, the value of "Occupation" is suppressed in a particular record, then this does not imply that the value of "Occupation" has to be suppressed in another record. The freedom that one has in selecting the values that are to be suppressed allows one to minimize the number of local suppressions. More on this subject can be found in De Waal and Willenborg (1995b).

Both global recoding and local suppression lead to a loss of information, because either less detailed information is provided or some information is not given at all. A balance between global recoding and local suppression has to be found in order to make the information loss due to SDC measures as low as possible. It is recommended to start by recoding some variables globally until the number of unsafe combinations that have to be protected by local suppression is sufficiently low. The remaining unsafe combinations have to be protected by suppressing some values.

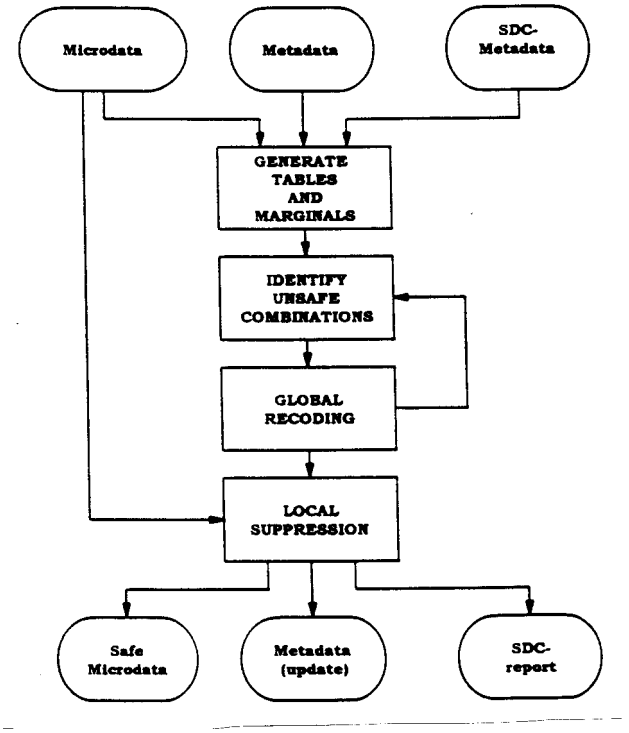
μ-ARGUS allows the user to specify the global recodings interactively. The user is provided by μ-ARGUS with information, helping him to select these global recodings. In case the user is not satisfied with a particular global recoding, it is easy to undo it. After the global recodings have been specified the values that have to be suppressed are determined automatically and optimally, i.e., the number of values that have to be suppressed is minimized. This latter aspect of μ-ARGUS, determining the necessary local suppressions

automatically and optimally, makes it possible to protect a microdata set against disclosure quickly.

The Development of m-ARGUS

As is explained above, a microdata file should be protected against disclosure in two steps. In the first step some variables are globally recoded. In the second step some values of variables should be locally suppressed. μ-ARGUS, currently under development, will be able to perform these tasks (see Figure 1). μ-ARGUS is a Windows 95 program developed with Borland C++.

Figure 1. -- m-ARGUS Functional Design



Metadata

To perform its task, μ-ARGUS should be provided with some extra meta information. At this moment, μ-ARGUS expects the data in a flat ASCII file, so the meta information should contain the regular meta data like the name, the position and the field width of the variables in the data file. Besides this the user needs to specify some additional (SDC-specific) metadata:

- the set of tables to be checked;
- the priority level for local suppression;
- an indication whether a variable has a hierarchical coding scheme -- this knowledge can be used for global recodings, as the truncation of the last digit is a sensible recoding operation for these coding schemes;
- a coding scheme for each variable; and
- a set of alternative codes (recoding schemes) for each key-variable.

The user is not required to specify the coding schemes for all the identifying variables. If the coding

scheme is not specified, μ-ARGUS will inspect the data file and establish the coding scheme itself from the occurring codes.

The level of identification is used to determine the combinations of the variables to be inspected. However, the user is free to determine the set of combinations to be checked, himself.

Generation of Tables and Marginals and Identification of the MINUCs

In order to identify the unsafe combinations and the MINUC's, the first step will be to generate the required tables and marginals. When the data files are available on the PC, the tables will be generated directly on the PC. However, in the case of very large files stored at an other computer (e.g., a UNIX-box), the part of μ-ARGUS that generates the tables can also be transferred to the UNIX-computer to generate the tables there. The ability to run μ-ARGUS on different platforms was the major reason for choosing C++ as our programming language.

When the tables have been generated, it is possible to identify the unsafe combinations. We are now ready to start the process of modifications to yield a safe file.

Global Recoding and Local Suppression

If the number of unsafe combinations is fairly large, the user is advised to first globally recode some variables interactively. A large number of unsafe combinations is an indication that some variables in the microdata set are too detailed in view of the future release of the data set. For instance, region is at the municipality level, whereas it is intended to release the data at a higher hierarchical level, say at the county or province level. To help the user decide which variables to recode and which codes to take into account, μ-ARGUS provides the user with the necessary auxiliary information. After these initial, interactive recodings, the user may decide to let μ-ARGUS eliminate the remaining unsafe combinations automatically. This automated option involves the solution of a complex optimization problem. This problem is being studied by Hurkens and Tiourine of the Technical University of Eindhoven, The Netherlands. Details can be found in Tiourine (1996). In that case, only those MINUCs can be eliminated automatically for which the SDC-metadata file contains alternative codes. The user should specify a stopping criterion, defining which fraction of the set of original MINUCs is allowed to remain, i.e., to be eliminated later by local suppression. The user can continue to experiment with recodings -- both interactive and automatic ones (by undoing them and searching for alternatives) -- until deciding which set of recodings to keep. Recodings that have been interactively established imply that the corresponding metadata (code descriptions, etc.) should be updated as well. If no MINUCs remain the job is finished and the global recodings can be performed on the microdata. However, in general, there are still MINUCs left which have to be eliminated by local suppression.

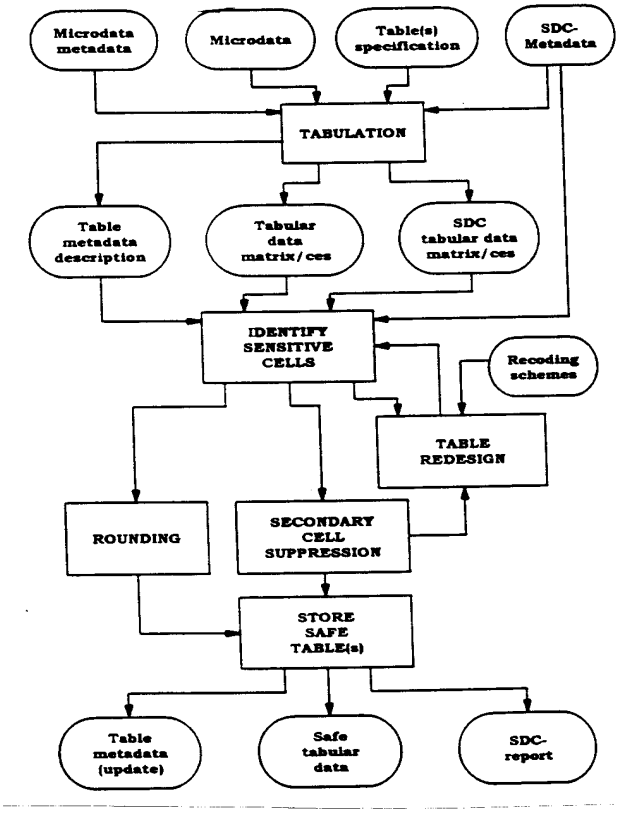
Final Steps

When the above-mentioned steps have been executed, the result is a safe microdata set. The only thing left is to write the safe file to disk and to generate a report and a modified metadata description. In the event that the original data file reside on another computer, μ-ARGUS will generate the necessary recoding information that will be used by a module of μ-ARGUS that runs on that other machine.

The Development of t-ARGUS

Besides the development of μ-ARGUS for microdata sets, the SDC-Project also plans development of τ-ARGUS. τ-ARGUS is aimed at the disclosure control of tabular data. (See Figure 2.) The theory of tabular disclosure control focuses on the "dominance rule." This rule states that a cell of a table is unsafe for publication if a few (n) major contributors to a cell are responsible for a certain percentage (p) of the total of that cell. The idea is that, in that case at least, the major contributors themselves can determine with great precision the contributions of the other contributors to that cell. Common choices for n and p are 3 or 70%, but τ-ARGUS will allow the users to specify their own choices. However, some modifications to this dominance rule exist.

Figure 2. -- t-ARGUS Functional Design



With this "dominance rule" as a starting point, it is easy to identify the sensitive cells, provided that the tabulation package cannot only calculate the cell totals, but also calculates the number of contributors and the individual contributions of the major contributors. Tabulation packages like ABACUS (made by Statistics Netherlands) and the Australian package SuperCross have that capability.

The problem, however, arises when the marginals of the table are published also. It is no longer enough to just suppress the sensitive cells, as they can be easily recalculated using the marginals. Even if it is not possible to exactly recalculate the suppressed cell, it is possible to calculate an interval which contains the suppressed cell. If the size of such an interval is rather small, then the suppressed cell can be estimated rather precisely. This is not acceptable either. Therefore, it is necessary to suppress additional information to ensure that the intervals are sufficiently large. Several solutions are available to protect the information of the sensitive cells:

- combining categories of the spanning variables (table redesign) -- more contributors to a cell tend to protect the information about the individual contributors;
- rounding the table, while preserving the additivity of the marginals; and
- suppressing additional (secondary) cells, to prevent the recalculation of the sensitive (primary) cells to a given approximation.

The calculation of the optimal set (with respect to the loss of information) of secondary cells is a complex OR-problem that is being studied by Fischetti. Details can be found in Fischetti (1996). τ-ARGUS will be built around this solution and take care of the whole process. For instance, in a typical τ-ARGUS session, the user will be presented with the table indicating the primary unsafe cells. The user can then choose the first step. He may decide to combine categories, like the global recoding of μ-ARGUS. The result will be an update of the table with presumably fewer unsafe cells. Eventually, the user will request that the system solve the remaining unsafe cells, by either rounding the table or finding secondary cells to protect the primary cells. The selection of the secondary cells is done so that the recalculation of the primary cells can only yield an interval. The size of these intervals must be larger than a specified minimum. When this has been done, the table will be stored and can be published.

The first version of τ-ARGUS will aim at the disclosure control of one individual table. A more complex situation arises when several tables must be protected consistently, generated from the same data set (linked tables). Then, there will be links between the tables that can be used by intruders to recalculate the sensitive cells. This is a topic of intensive research at this moment. The results from this research will be used to enhance future versions of τ-ARGUS, to take into account links between tables.

References

De Waal, A.G. and Willenborg, L.C.R.J. (1995a). A View on Statistical Disclosure Control for Microdata, *Survey Methodology*, 22, 1, 95-101, Voorburg: Statistics Netherlands.

De Waal, A.G. and Willenborg, L.C.R.J. (1995b). Global Recodings and Local Suppressions in Microdata Sets, Report, Voorburg: Statistics Netherlands.

Fischetti, M. and Salazar, J.J. (1996). Models and Algorithms for the Cell Suppression Problem, paper presented at the 3rd International Seminar on Statistical Confidentiality, Bled.

Tiourine, S. (1996). Set Covering Models for Statistical Disclosure Control in Microdata, paper presented at the 3rd International Seminar on Statistical Confidentiality, Bled.

Willenborg, L.C.R.J. (1996). Outline of the SDC Project, paper presented at the 3rd International Seminar on Statistical Confidentiality, Bled.

Willenborg, L.C.R.J. and de Waal, A.G. (1996). *Statistical Disclosure Control in Practice*, New York: Springer-Verlag.

Chapter 5

Investigating Auto Injury Treatment in a No-Fault State: An Analysis of Linked Crash and Auto Insurer Data

Lawrence H. Nitz and Karl E. Kim, University of Hawaii at Manoa

Abstract

Hawaii is a no-fault insurance State which provides for choice of a variety of alternative therapies for motor vehicle injury victims. The two most frequently used providers are medical doctors (MD) and doctors of chiropractic (DC). A large portion of this care is rendered in office visits, and can be readily identified from insurance payment records. The focus of this study is the distribution of these types of direct medical care across crash types and circumstances. Study data include police crash reports and 6,625 closed case files of a Hawaii auto insurer for the years 1990 and 1991. The files were linked with Automatch, a probabilistic record linkage program, using crash date, crash time, gender and birth date as match fields (Kim and Nitz, 1995; Matchware Technologies, 1996). The insurance payment file indicates the type of treatment received by persons injured in collisions. The study asks two questions about the choice of care among crash victims:

- *Who goes to a chiropractor?*
- *What is the relationship between occupant, vehicle and crash characteristics and the choice of care?*

Background

Hawaii has had a no-fault insurance system for over twenty-five years (HRS 431:10C). The program was initially introduced to assure the availability of automobile liability insurance to all residents without age or gender discrimination. Underwriting is limited to rate adjustments based on the driving record of the individual applicant (HRS 531:10C-111(c)). The program, since its inception, provided that each motor vehicle operator's own insurance carrier would provide coverage for personal injury protection (PIP) for all injuries to the operator or his or her passengers, without examination of the issue of fault, up to a certain value, the medical-rehabilitative limit, or "tort floor," (HRS 431:10C-103(10)(c). This was $15,000 in 1990.) Injury costs beyond this value could be recovered from a party deemed to be at fault through a tort action in the courts. Each vehicle operator's auto policy was also required to provide at least a minimum level of bodily injury (BI) protection for others who might be injured by the insured driver. This BI coverage could normally only be touched in the event that the injured party had reached the tort floor by claims against his or her own PIP coverage or had no PIP coverage. (A pedestrian, for example, would be able to make a BI claim directly.) Insurance carriers also offered additional coverage for uninsured and under insured motorists (UI and UIM). Hawaii drivers typically purchased these coverages to protect against catastrophic losses that might be incurred should they be in a collision with an un- or under-insured motorist. In the event that all available coverages had been exhausted, the injured party's medical insurer was held responsible for all remaining medical expenses. The medical insurer was deemed not responsible for any

auto-related injury cost prior to the exhaustion of auto policy benefits.

The State Insurance Commissioner is authorized to adust the *tort floor*, every year to set a level at which 90% of all injury claims will be covered by PIP without resort to a suit (HRS 431:10C-308). In 1990, the tort floor was $15,000. If total medical and rehabilitation expenses and loss of wages exceeded this value, the patient could file a tort suit. Tort suits could be filed without respoect to monetary values for permanent and serious disfigurement, permantent loss of use of a body part, or death. In addition, the Insurance Commissioner set an annual "medical rehabilitative limit," a value of medical and rehabilitative expenses which would be sufficient to permit filing of a tort suit. The medical rehabilitative limit for 1990 was $7,000, and for 1991 was $7,600. Until recent reform legislation passed in 1994, the auto injury patient's choice of medical care facility and the scheduling of therapies recommended was very broad (HRS 431:10C-103(10)(A)(i)).

> "All appropriate and reasonable expenses necessarily incurred for medical, hospital, surgical, professional, nursing, dental, optometric, ambulance, prosthetic services, products and accommodations furnished, and x-ray. The foregoing expenses may include any non-medical remedial care and treatment rendered in accordance with the teachings, faith or belief of any group which depends for healing upon spiritual means through prayer...."

In this context of open benefit provisions, the state has shown dramatic growth in the availability of chiropractic services, pain clinics, physical therapy facilities, and massage therapy practitioners.

The 1992 legislative session (1992 Hawaii Session Laws Act 123, Sec.7) put the auto injury treatment allowances on the same regimen as the disability-graded allowances for workers' compensation medical care and rehabilitative therapy (HRS 431:10C-308.5). The fact that there has been broad choice of type and amount of therapy for many years suggests that it might be useful to understand the relationships between objective features of the crash event and the actual choice of care. To make this analysis possible, it was necessary to link an auto insurer's payment file to the police motor vehicle accident report file. The next section outlines the data and procedures used to make this linkage.

Data

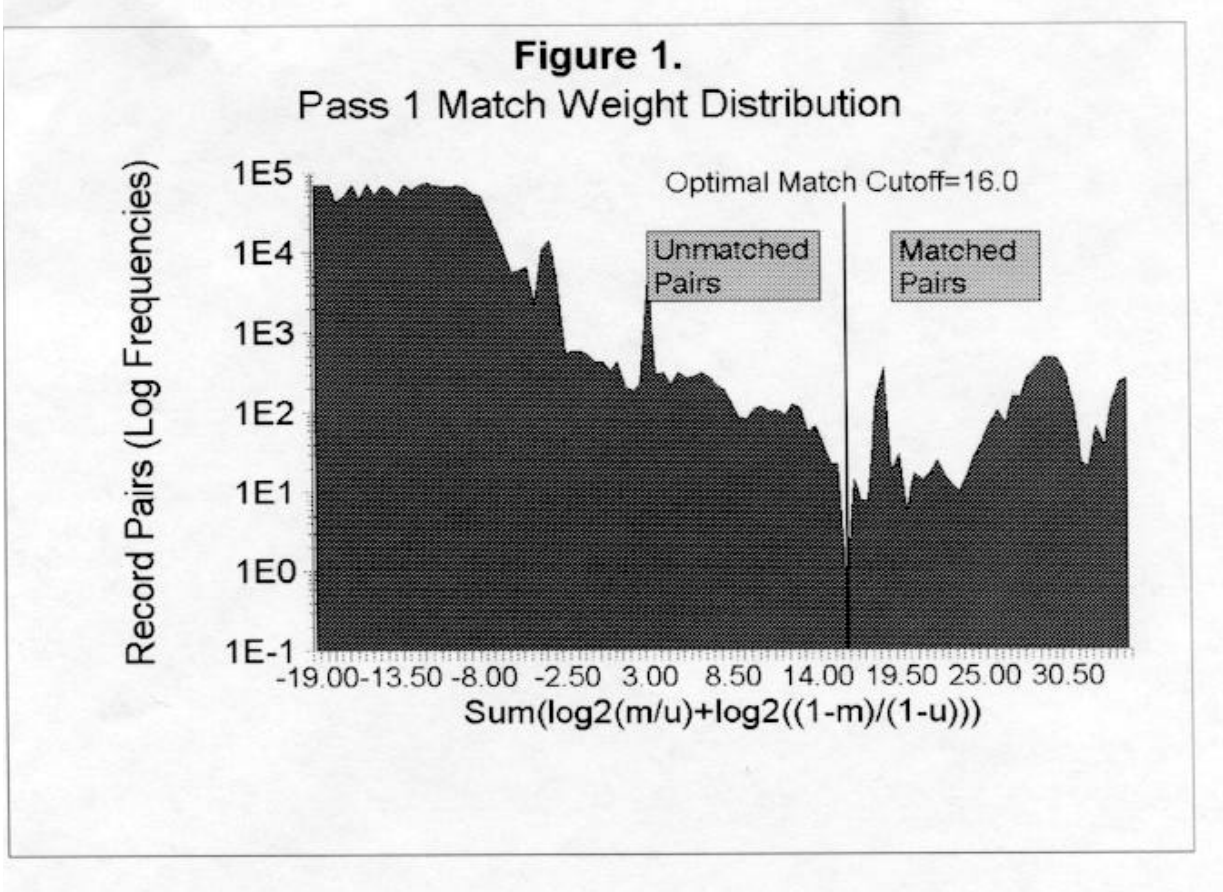
The police crash report file is maintained by the Hawaii State Department of Transportation. The four county police departments in Hawaii are required to report every motor vehicle collision on a public road which involves an injury or death or estimated damages of $1,000 or more (in 1990). The reporting form contains extensive description of the crash circumstances, features of the roadway and traffic environment, and driver characteristics. Where an injury has been reported, it also contains the police officer's description of the severity of injury on a five-level scale (K = killed, A = incapacitating injury, B = non-incapacitating injury, C = possible injury, and 0 = no injury). Drivers were also identified by their birth dates. Two years, 1990 and 1991, form the pool of reported motor vehicle collisions for this analysis.

The insurance file consisted of 6,625 closed case records of a Hawaii auto insurer for the years 1990 and 1991. The file contained the closing accounting records on these cases, showing the total of all sums paid out, and the elementary data on the characteristics of the crash and the injured party. All records represented claims actually paid; the maximum payment is the policy limit chosen by the insured in buying the policy. For this particular group of policies, the maximum policy limit the company offered was $300,000. Any additional coverage that might have been carried by way of an umbrella clause appeared in a different set of accounts. Injured persons were identified by birth date, gender, and date and time of the collision. The crash date, crash time, gender and driver age are common to both files, and provide a basis for linking the insurance payouts to details on the crash itself. Another insurance file contained details of about 58,000 transactions recording specific payments made in processing the claim. This file was initially processed by

extracting the payee field as a SAS character field, and parsing it for common character combinations of the license designations MD (Medical Doctor) and DC (Doctor of Chiropractic). The file was then summed by claimant; applicable office call charges for MD and DC services were used to create separate analysis variables. This summary file was then merged with the matched crash and insurance event record file.

The Matching Procedure

Matchware Technologies's Automatch 4.0 was used to match the crash and the closed case file in three steps (Matchware Technologies, 1996). (See Kim and Nitz, 1995, for a more extensive discussion of the Automatch application.) Pass 1 divided the file into homogeneous blocks based on age and sex, and matched on time of crash, date of crash, and birth date. The log frequency distribution of the Fellegi-Sunter match weights for Pass 1 is shown in Figure 1. For Pass 1, 2,565 record pairs were designated as matches, with match weights meeting or exceeding the cutoff value. The optimal cutoff for the first pass was 16.0, which marks a relatively clear division in the distribution, as indicated in Figure 1. (The count of nonmatches with match weights of -19 and lower was truncated at 80,000 pairs, and the observed frequencies (+.5) were logged for display purposes.)



Pass 2 blocked on date of crash and sex, and matched on time and birth date. An additional 1,001 record pairs were selected as matches with a criterion value of 12.0. Pass 3 was designed to pick up erroneous

recording of military time values. A field was created in the driver crash record which recoded afternoon and evening times into the 12-hour time scale. The cases were blocked on date and birth date and matched on sex and the 12-hour time value. Only seven additional cases were identified, using a criterion value of 7.0, to produce a total matched set of 3,573 cases. (The analyses to follow report the smaller numbers with complete information on crash characteristics of interest.)

ID values were extracted from the match output file and used to index the matching cases in the crash report, EMS run report, and insurance case files. These were then directly merged by common ID values in a SAS step.

Match Results

A comparison of the distributions of the police crash file and the matched file suggest parallel profiles for temporal and environmental crash characteristics. Of the insurance cases, 3,573 (54%) were matched to drivers and other principals identified by birth date in the crash file. There were no significant differences in the two distributions by intersection/mid-block location (Chi-square=.51, 1 df, p<.48, phi=.002), month (Chi-square 5.77, 11df, p<.89, phi=.008) or day (Chi-square= 4.09, 6 df, p<.66, phi=.007). The profiles for time distribution by hour, urban/rural location, and daytime and nighttime peak traffic periods showed significant, but low level differences (phi coefficients generally <.02). Gender, human factor, and police judgements of injury severity differed substantially across the two files, with the matched insurance file being more seriously injured (57% of insurance claimants denoted "not injured" versus 74% of the police report file), more female (46% female in the insurance file to 34% in the police report file), less likely to report driving errors (62% v. 55%), and less likely to report human factor problems (55% to 49%).

Findings

Who Goes to the Chiropractor?

Earlier work with matched cases in Hawaii suggests that the configuration of the crash event, in particular crash type, and driver behaviors (human factors, driving errors, and other fault indicators) are major determinants of injury outcomes (Kim et al., 1995; Kim and Nitz, 1996; Kim et al., 1994). The linked insurance file allows further examination of the role of these factors, along with standard demographic indicators, in the choice of medical and therapy office calls.

In this discussion, we will first present effects that distinguish crash victims who use three classes of therapy: only chiropractic services; only medical services (MD-only); and some combination of chiropractic and medical services. Next we will discuss the patterns of therapy choices for demographic groups, then we will examine care usage for specific crash circumstances.

Relatively few of the crash drivers -- 89 persons, about 7% of the cases with detailed crash data, used only chiropractic services. This is a somewhat unexpected finding, considering the popularity of chiropractic care for auto trauma cases. Forty-four percent of the group using only chiropractic services were male, as opposed to 55% for those using both chiropractic and physician services, and 50% for those using only physician services. The most frequent age group for chiropractic-only use was 21-34 year-olds, with a 53% use rate, as opposed to 45% for those using both chiropractic and physician care and 36% for MD care alone. Those in the 45-64 age group comprised about 10% of the chiropractic service users, compared with 19% of the MD-only users. Seventy-one percent of the chiropractic-only users had no police recorded driving errors, as opposed to about 58% for users of the chiropractic and physician combination users and the MD-only users.

The small number of chiropractic-only users suggests that the group might be combined with those who use both chiropractic and physician services. Over 1,400 cases used a combination of chiropractic and physician services, and 1,105 used only physician services. Grouping all chiropractic users together will permit examination of the question of who goes to a chiropractor in Hawaii.

Table 1.—Chiropractic and Physician Office Visits by Driver Gender

Sex of Occupant	Chiropractic Use		MD Only		Totals	
	N	%	N	%	N	%
Female	722	45.35	550	50.00	1,272	47.25
Male	870	54.65	550	50.00	1,420	52.75
All	1,592	100.00	1,100	100.00	2,692	100.00

Chi-square = 5.64, 1 df, $p < .018$

Table 1 presents the choice of therapy by gender. The MD-only users were evenly split between men and women, while 55% of the users of chiropractic services were men. The pattern of usage by age is shown in Table 2. The age profiles differ significantly ($p<.001$). The 21-34 year-old group constitutes 47% of the chiropractic users, while it accounts for only 37% of the MD-only users. At the higher end of the age scale, the 45-64 year-old group comprises 13% of the chiropractic users, but 20% of the MD-only users. The pattern appears to suggest rapidly declining use of chiropractic services and relatively slower decline in use of MD-only services as individuals age. Other age groups do not differ meaningfully in chiropractic and MD use.

What Is the Relationship Between Occupant, Vehicle, and Crash Characteristics and Choice of Care?

There are slight differences in police-reported seatbelt use for chiropractic and physician service users, with 97% of the chiropractic users reporting belt use, and 95% of MD-only users report having been belted during the crash, as shown in Table 3. The crash report belt use rate is higher than previous independently observed belt use rates in the 80% range. Hawaii has a primary enforcement law for seatbelt violations. The penalties for being unbelted may raise reported belt use rates.

Table 2.—Chiropractic and Physician Office Visits by Driver Age

Driver Age	Chiropractic Use		MD Only		Totals	
	N	%	N	%	N	%

Less than 15	27	1.70	17	1.55	44	1.64
15 - 18	84	5.30	102	9.32	186	6.94
18 - 21	143	9.02	95	8.68	238	8.88
21 - 34	748	47.16	401	36.65	1,149	42.87
35 - 45	302	19.04	215	19.65	517	19.29
45 - 64	212	13.37	214	19.56	426	15.90
65 +	70	4.41	50	4.57	120	4.48
All	1,592	100.00	1,100	100.00	2,692	100.00

Chi-square = 47.76, 6 df, p < .001

Table 3.—Chiropractic and Physician Office Visits by Seatbelt Use

Seatbelt Use	Chiropractic Use		MD Only		Totals	
	N	%	N	%	N	%
Not belted	48	3.17	52	4.98	100	3.90
Belted	1,468	96.83	993	95.02	2,461	96.10
All	1,516	100.00	1,045	100.00	2,561	100.00

Chi-square = 5.4, 1 df, p < .020

The distribution of users of chiropractic versus MD-only care differs in a number of ways across types of crashes. Table 4 illustrates the distribution of care choices across crash types commonly considered "at fault." (The “at fault” drivers are identified as those striking another car, and those involved in rollovers.) The fault profiles of the two usage groups differ significantly. The not-at-fault drivers comprise 58% of the chiropractic users and 64% of the MD-only users. These rates are consistent with a pattern of using services more frequently when another party is felt to be at fault.

Table 4.—Chiropractic and Physician Office Visits by Crash Fault

Fault	Chiropractic Use		MD Only		Totals	
	N	%	N	%	N	%
At fault	540	42.32	287	35.83	827	39.82
Not at fault	736	57.68	514	64.17	1,250	60.18

All	1,276	100.00	801	100.00	2,077	100.00

Chi-square =8.65, 1 df, p < .003

Table 5 illustrates the relationship between care choice and police reported injury severity. The profiles differ significantly (p< .001). Of the chiropractic service users, 64% are reported by police to be "no injury" cases, while only 46% of the MD-only group are reported without injury at the scene. Computing the fraction of each injury level which uses MD-only services indicates that the 34% of those reported as *no injury* use MD services only, the rest using a combination of chiropractic and physician services. Sixty-six percent of those with *incapacitating injuries* using MD-only services, while only 17% use a combination of both chiropractic and physician services.

Table 5.—Chiropractic and Physician Office Visits by Police Reported Injury Severity

Police Injury Severity	Chiropractic Use		MD Only		Totals	
	N	%	N	%	N	%
No injury	853	63.70	447	46.13	1,300	56.33
Possible injury	252	18.82	231	23.84	483	20.93
Non-incapacitating injury	211	15.76	243	25.08	454	19.67
Incapacitating injury	22	1.64	43	4.44	65	2.82
Fatality	1	0.07	5	0.52	6	0.26
All	1,339	100.00	969	100.00	2,308	100.00

Chi-square = 82.21, 4 df, p < .001

Crash type affects the distribution of care choices also, as shown in Table 6. Sixty-two percent of the chiropractic service users were involved in rear-end collisions, while only 54% of the MD-only users were involved in rear-end collisions. Head-on collisions and rollovers account for 3.5% of chiropractic users, compared to 8.1% of MD-only users. Broadside collisions, similarly, account for more MD-only than chiropractic usage (25% to 21%).

Table 6.—Chiropractic and Physician Office Visits by Crash Type

Crash Type	Chiropractic Use		MD Only		Totals	
	N	%	N	%	N	%
Broadside	269	20.53	215	25.29	484	22.41
Head on/rollover	46	3.51	69	8.12	115	5.32
Rear end	816	62.29	458	53.88	1,274	58.98
Sideswipe	179	13.66	108	12.71	287	13.29

All	1,310	100.00	850	100.00	2,160	100.00

Chi-square = 32.29, 8 df, p < .001

Human factors show several small effects on care choice (see Table 7). Chiropractic service users are slightly more likely to have committed misjudgements than MD-only users (12% to 9%), and about half as likely to have been in an alcohol or fatigue related crash as MD-only users (1.4% to 3.5% and 1.3% to 2.1%, respectively).

Table 7.—Chiropractic and Physician Office Visits by Human Factors

Human Factors	Chiropractic Use		MD Only		Totals	
	N	%	N	%	N	%
Inattention	331	20.92	233	21.47	564	21.15
Misjudgement	190	12.01	100	9.22	290	10.87
Fatigue	22	1.39	38	3.50	60	2.25
Alcohol	21	1.33	23	2.12	44	1.65
Other	70	4.42	61	5.62	131	4.91
None	948	59.92	630	58.06	1,578	59.17
All	1,582	100.00	1,085	100.00	2,667	100.00

Chi-square = 22.17, 5 df. P < .001

A variety of driving errors are known to lead to different types of crashes (Kim et al., 1995) Table 8 shows that chiropractic user crashes consisted of 3.5% involving excess speed, while MD-only users involved 4.9% crashes involving excess speed. A similar pattern is found for driving the wrong way: 11% of chiropractic users, and 13% of MD-only users were involved in these types of crashes. (Driving the wrong way includes driving in the wrong lane, crossing the centerline, improper passing, and related offenses.) Following too closely accounted for nearly twice as much chiropractic as MD-only use -- 9.5% of chiropractic users followed too closely, while only 5.3% of the MD-only users did so.

Table 8.—Chiropractic and Physician Office Visits by Driver Errors

Driver Errors	Chiropractic Use		MD Only		Totals	
	N	%	N	%	N	%
Excess speed	56	3.53	54	4.94	110	4.10
Disregard controls	20	1.26	28	2.56	48	1.79
Driving wrong way	174	10.96	146	13.36	320	11.94
Improper turn	8	0.50	2	0.18	10	0.37
Following too closely	145	9.13	58	5.31	203	7.57
Other	150	9.45	121	11.07	271	10.11
None	1,035	65.18	684	62.58	1,719	64.12

All	1,588	100.00	1,093	100.00	2,681	100.00

Chi-square = 29.08, 6 df, $p < .001$

Summary and Discussion

The set of police crash report records matched to insurance claim records is associated with higher levels of injury than the unmatched police reports. It also has a larger proportion of females than the overall police report population. Typical environmental variables, such as time of day and intersection vs. mid-block location, show no meaningful differences between the matched and unmatched cases. The principal distinguishing factor in the reported crashes which match to insurance records appears to be the severity of injury -- more severe injuries result in more frequent insurance claims.

This study provides new and useful information about the choice between traditional medicine, and approaches which rely to some degree on alternative forms of care, in this case, chiropractic care. Several conclusions emerge from the analysis:

- persons who are "not-at-fault" (usually the struck party) use MD and chiropractic services more frequently than those "at-fault;"
- the use of chiropractic services is substantially higher than the use of MD-only services among occupants with low severity police reported injuries;
- those who commit what might be seen as the most serious driving errors in the course of a collision (driving on the wrong side, ignoring traffic controls, speeding) are less likely to use chiropractic care than those who commit no errors or more minor errors (e.g., following too closely, inattention, misjudgment);
- when the driver has been impaired by alcohol, the use of chiropractic services is about half the level of use of MD-only services; and
- more chiropractic services are used by men than women, particularly in the ages of 21 to 34.

Our study has a number of limitations. The first stems from the nature of the auto insurance market in Hawaii: no single insurer holds a very large share of the total market, so the number of policies, and thus claims paid by each insurer are relatively small. The second is that the data spans only a portion of a time period in which three substantial changes have been made to Hawaii's motor vehicle insurance laws. The third is the restriction of the present analysis to choice of care, rather than total cost of care.

There is decidedly a need for additional research to address these limitations. Case files from additional insurers would increase the pool of claims, and allow exploration of whether company practices affect claim patterns. Extending the time period covered from 1990-91 through 1995-96 would span a major change in the way chiropractic charges were to be reimbursed under Hawaii motor vehicle insurance policies, providing a natural quasi-experiment. This would allow a test of the effect of subjecting chiropractic treatment to the workers' compensation schedule.

New research questions on choice of therapy could extend the results of this study by examining some questions not raised in the present study.

- That is the relationship between fault and the quantity and cost of chiropractic or other alternative care used?
- How do drivers' prior histories in terms of traffic violations or insurance claims affect the nature of the care they choose when injured in a collision?
- How would the patterns of choices and costs of care differ in a pure tort law state?
- There is clearly a need for more research on the role of crash and occupant characteristics in choice of therapies.

References

Hawaii Revised Statutes, Annotated. 1994 Replacement and 1996 Supplement. Michie & Co.

Hawaii Session Laws (1992). Act 123, Sec. 7.

Kim, K. E. and Nitz, L. H. (1995). Application of Automated Records Linkage Software in Traffic Records Analysis, *Transportation Research Record*, 1467, 40-55.

Kim, K. E. and Nitz, L. H. (1996). Understanding Causes of Injuries in Motor Vehicle Crashes, *Transportation Quarterly*, 50, 1, 105-113.

Kim, K. E.; Nitz, L. H.; Richardson, J.; and Li, L. (1994). Analyzing the Relationship between Crash Types and Injuries in Motor Vehicle Collisions in Hawaii, *Transportation Research Record*, 1467, 9-13.

Kim, K.E.; Nitz, L. H.; Richardson, J.; and Lei, L. (1995). Personal and Behavioral Predictors of Automobile Crash and Injury Severity, *Accident Analysis and Prevention*, 27, 4, 469-481.

Matchware Technologies, Inc. (1996). Automatch 4.0.

Chapter 5

Quantitative Evaluation of the Linkage Operations of the 1996 Census Reverse Record Check

Julie Bernier, Statistics Canada

Abstract

A probabilistic linkage of two files is performed using the theory derived by Fellegi and Sunter (1969). The decision on whether a unit from each file are linked is based on the linkage weight obtained. In effect the linkage weight, a one-dimensional variable, is divided into three ranges: one for which links are accepted, one for which they are rejected and the intermediate range, where a link is possible. Manual inspection of possible links is needed to decide which ones represent the same unit. At the end of the linking procedure, the accepted links and those possible links that were confirmed by the manual check are kept. Under certain conditions, the results of this check provide all the information needed for a quantitative evaluation of the quality of the links. In this article we present a brief description of the Reverse Record Check (RRC) and the role of probabilistic linkage in this project. We then offer a definition of the reliability of a link and describe a procedure for estimating the minimum reliability of the links kept, using a fitted logistic regression model based on the manual checking of the possible links. Finally, we present the results obtained for the RRC96, describing the number of links obtained and the reliability of those links.

Introduction

When a probabilistic linkage is performed between two files, any of several approaches may be used. Depending on the approach chosen, it may be that among the linkage weights obtained, there will be a limited number of different values. In this case, a number of links are associated with each weight, and we can proceed by sampling to estimate the proportion of true links for each possible weight value. The next step is to manually inspect the links sampled. It may also happen that the set of possible values of the linkage weight will be quite varied. This may result in the use of a great number of comparison rules, each making a different contribution to the total weight depending on whether or not there is a match between the fields compared. This variety of weights may also result from the use of comparison rules that assign frequency weights in the event of a match. The use of frequency weights means that where there is a match, a different contribution is made to the total weight depending on whether the content of the fields compared is more or less frequent in the population. For example, a larger contribution is made when there is a match on a relatively rare family name. In the case of a set of varied weights, the distribution of links on the basis of the linkage weights closely resembles a continuous distribution. The proportion of true links may then be estimated by grouping the weights by intervals or by using a logistic regression. The use of logistic regression was chosen as the method of estimating the proportion of true links in the linkage of the 1996 Reverse Record Check (RRC96) with the 1990 Revenue Canada files (RCT90), since in that linkage, a number of comparison rules were involved. Furthermore, for two of the fields compared, namely family name and the first three entries of the postal code, frequency weights were used.

The Reverse Record Check

The purpose of the reverse record check is to estimate the errors in coverage of the population and of private households in the Census. It also seeks to analyse the characteristics of persons who either were not enumerated or were enumerated more than once. The method used is as follows:

- Using a sample frame independent of the 1996 Census, a sample is drawn of persons who should have been enumerated in the Census.
- A file is created containing as much information as possible on these persons and their census families.
- If possible, the addresses of the selected persons (SP) and their family members (close relatives living under the same roof) are updated using administrative files.
- Retrieval operations are carried out by interviewers in order to contact the selected person and administer a questionnaire to him or her. The purpose of the questionnaire is to determine the addresses at which the person could have been enumerated.
- Search operations are carried out on the questionnaires and in the Census database in order to determine how many times the selected person was enumerated.

The Role of Probabilistic Linkage

Probabilistic linkage is used in the address updating procedure. In this procedure there are two principal stages. First, probabilistic linkage of the RRC96 with the Revenue Canada 1990 (RCT) file is carried out. The reason for choosing the year 1990 is that this database was created in early 1991 and the sample frame of the RRC is largely made up of the database of the 1991 Census and the files of the RRC91. When this linkage is successfully completed, we obtain the social insurance number (SIN) of the selected person or a member of that person's family. In the second stage, an exact linkage is made between the RRC96 and the 1991, 1992, 1993, and 1994 Revenue Canada files in order to obtain the most recent address available in those files. For this linkage, the SIN is used as an identifer. It is by means of these addresses that we can begin tracing the selected persons by the RRC.

During operations to link the RRC sample with the 1990 Revenue Canada files, we determined, for each of the eight region-by-sex groups, a threshold linkage weight beyond which all links were considered definite or possible and were retained for the next stage. Subsequently, we checked the weakest links in order to determine whether they were valid or false. This enabled us firstly to eliminate the false links before proceeding to subsequent operations and secondly to determine the reliability of the links retained. Two other approaches may be used. One can define a fairly low linkage weight beyond which all links are kept without being checked. This yields a greater number of links, some of which have little likelihood of being valid. There are two drawbacks to this approach. First, it means that the interviewers responsible for tracing selected persons are given more false leads. This can result in time loss during tracing and a greater probability of interviewing by error a person other than the one selected in the sample. Second, the update address is processed in the search operation. This too can needlessly increase the size of this operation. The other possible approach is to define a fairly high linkage weight beyond which all links are retained without being checked. They yields fewer links, but those obtained have a strong probability of being valid. The disadvantage of this method is that it increases the number of persons not traced. This type of nonresponse is more common in the case of persons living alone, and such persons are also the ones who have the greatest likelihood of not being enumerated. It is for this reason that we preferred the approach that requires

manual checking but serves to reduce this type of nonresponse without needlessly expanding the tracing and search stages.

Checking Procedure

In light of the amount of data to be processed, linkage is carried out separately in eight groups defined by the sex and the geographic region of the individual to be linked. The four geographic regions are: Eastern Canada, Quebec, Ontario, and lastly, Western Canada and the Northwest Territories. For each of the four regions it is necessary to define a grey area within which links are considered "possible" rather than being accepted automatically. This area extends from the lower boundary weight (LOW) to a weight determined iteratively (UPP) in the course of checking. The point LOW is determined through guesswork, by examining the links ranked by descending order of weight. The persons engaged in this task try to choose a point LOW such that manual checking will be done when links have a fairly high probability of being valid (approximately 75%). The checking begins with UPP chosen such that the grey area contains roughly 1.5% of the links retained for the region in question. To reduce the workload, some of these links are then checked automatically, in the following manner: when both spouses in a household have linked, if one of the two (C1) obtained a high linkage weight and if in addition that person's record at RCT is found to contain the SIN of the spouse (C2), and if that SIN is the same as the one found in the record that is linked with C2, then the link of C2 with RCT is considered reliable, even if it obtained a linkage weight within the grey area. All the links in the grey area that did not satisfy the foregoing criterion were checked manually. These checks were carried out using all available information on the household as a whole. After the entire grey area was checked, if the number of rejected links seemed high, UPP was changed so as to add from 1.5% to 2% more links to the grey area. These two steps (choosing UPP and checking) were repeated until the rejection rate for the links checked seemed lower than 10% for links with a linkage weight close to UPP.

Shown below, for each region, are the grey area boundaries, the percentage of links within those boundaries and the total percentage of links rejected in the grey area.

Table 1. -- Results of Checking

Region	LOW	UPP	Percentage of Links Checked	Percentage of Links in Grey Area Rejected
Eastern	222	244	1.5	2.1
Quebec	221	304	7.5	23.0
Ontario	274	309	2.0	1.2
Western	219	258	1.5	4.9

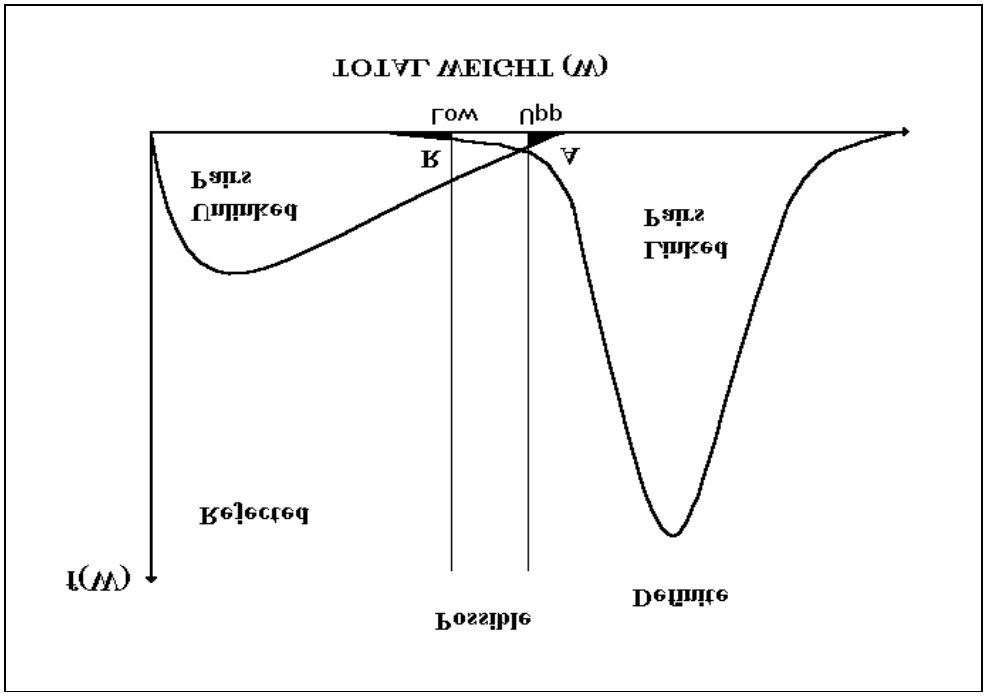
As stated in the introduction, these checks are useful in two ways. First, they serve to eliminate most of the false links. They therefore enhance the quality of the set of links obtained. Second, these checks enable us to form a data set that contains, for various links, both their status (valid or false) and their linkage weight. Using this data set, we were then able to assess the reliability of the accepted links.

Definition of the Reliability of a Link

The probabilistic linkage procedure consists in calculating, for each pair of records, a weight W based on whether the fields compared match or do not match and on the probability of matching these fields

given a linked pair or an unlinked pair. Generally, during the matching procedure, we try to determine a lower boundary and an upper boundary such that pairs with a weight lesser than LOW are rejected, those with a weight greater than UPP are accepted and those between these two boundaries are considered as possible links and eventually undergo another classification procedure. The following figure illustrates these concepts.

Figure 1.--Distributions of Pairs of Records by Linkage Weight



In linkage, two types of error are possible: accepting a pair that is not linked (A) or rejecting a linked pair (R). We are usually interested in the following probabilities:

> **P(accept a link | the pair is not linked) = P(W>UPP | the pair is not linked) and**
> **P(reject a link | the pair is linked) = P(W<LOW | the pair is linked)**,

which are called ***classification error probabilities***. We try, then, to choose LOW and UPP such that these two probabilities meet certain optimization criteria (see Fellegi and Sunter). Methods for estimating these probabilities may also be obtained by using samples of accepted links and rejected links that are checked manually (see Armstrong and Mayda, 1993 for a partial review of these methods). For the RRC96, we proceed differently. We determine a point LOW below which all links are rejected, but we do not define in advance a point UPP that would separate possible links from definite links. This point is instead determined during the manual check when it is felt that the links checked exhibit a high enough frequency to stop checking. Here we are instead interested in the following probabilities:

> **P(the link is valid | the link is accepted) = P(valid | W>UPP)** and
> **P(the link is valid | the link is rejected) = P(valid | W<LOW).**

These two probabilities will be called ***the reliability of accepted links and the reliability of rejected links.*** It should be noted that the term reliability applies here to a link and not to the procedure that leads to the

acceptance or rejection of this link. We therefore speak of the reliability of a rejected link as being the probability that this link is valid, which in fact amounts to a classification error. We could estimate these two probabilities respectively by the proportion of linked pairs among the accepted links and the proportion of linked pairs among the rejected links. These estimates would require manual checking of two samples drawn respectively from the accepted links and the rejected links. We ruled out this method for two reasons. First, the rejected links set was not retained. Second, for an estimate of a very low error rate to be acceptable, a very large sample is required, which means that the more successful the linkage procedure, the more costly the quantitative evaluation of the reliability of links using two samples. We therefore chose an alternative that allows us to use checking in the grey area rather than requiring checking of one or two additional samples.

Reliability Evaluation Procedure

We can speak generally of P(valid | W>UPP), which is the reliability of the links in the accepted links set, and of P(valid | W<LOW), the reliability of the links in the rejected links set; but we cannot speak more specifically of P(valid | W), the proportion of valid links in the subset consisting of pairs with linkage weight W. The proportion P(valid | W) may be defined as the reliability of a link of weight W. When we speak of quantitative evaluation, we may want to obtain a general estimate of the reliability of the accepted links and the rejected links, or we may want more specifically to estimate P(valid | W) for certain critical values of W. Since this probability increases with W, we have only:

P(valid | W=UPP) constitutes a lower boundary for the reliability of the accepted links.
P(valid | W=LOW) constitutes an upper boundary for the reliability of the rejected links.

In addition, no error is associated with the grey area, meaning that we consider that the manual check is a total success. Our quantitative evaluation therefore consists in estimating P(valid | W=UPP) and P(valid | W=LOW). To do this we use a logistic regression model, the parameters of which are estimated from the links in the grey area. This method is based on two assumptions. First, it must be assumed that the variable W is linked linearly to the logit function of the reliability to be estimated (logit(p)=log(p/1-p)). The logistic model is of the following form:

logit(p | W) = a + b W, where p is the probability that the link is valid.

This condition, which constitutes a test of goodness of fit for the model, is verified by a method described in the appendix. Second, the grey area must contain a sufficient number of unlinked pairs with various W values. When the number of unlinked pairs in the grey area is insufficient, the hypothesis $\beta=0$ cannot be rejected at a meaningful level. In the latter case, the proportion of valid links in the grey area is very high and can serve as the upper boundary for P(valid | LOW) and the lower boundary for P(valid | UPP). It should be noted that such a situation means that we have been too strict in choosing the cutoff point LOW in the linkage operations, and have therefore rejected many valid links and inappropriately used manual checking on a set of links with very high reliability. The procedure proposed is therefore the following:

- Check links in the grey area; each pair is considered linked or unlinked.
- Estimate parameters α and β of the logistic regression.
- Test the goodness of fit of the model and test the hypothesis $\beta=0$.
- If the results of the tests are satisfactory, estimate P(valid | W=UPP) by using logit(P(valid | W=UPP)) = α +β UPP and estimate P(valid | W=LOW) by using logit(P(valid | W=LOW)) = α +β LOW.

- If the results of the tests do not allow us to use the model, we merely estimate the proportion of valid links in the grey area.

Results

Shown below are the results obtained for the four regions. The estimates are made using both males and females, since introducing the sex variable into the logistic regression does not make a significant contribution.

Table 2. -- Estimate of Reliability

Region	Eastern	Quebec	Ontario	Western
Estimate of regression equation		logit(p)= -2.82 + 0.0165 W		logit(p)= -12.70 +0.0665 W
Estimate of reliability at W=LOW		69.6%		86.6%
Estimate of reliability at W=UPP	> 97.9%	90.0%	> 98.8%	98.8%
Estimated W for which reliability is 90%		304		224

For Eastern and Ontario regions, we didn't find enough unlinked pairs to do a logistic regression. This means that we could probably have set LOW lower in the linkage operations. For Quebec and Western regions, we estimated the reliability using the logistic regression model. It will be recalled that we check either 1.5% of the weakest links or several series of links until the estimated reliability at W=UPP seems to us to be greater than 90%. For Region 2, the estimate of 90% for reliability at UPP shows that we succeeded in choosing UPP such as to ensure good reliability of links while minimizing manual checking.

It should be recalled that:

- All links in the grey area were checked, and those that were false were rejected.

- The estimated reliability at point UPP is a lower boundary for the reliability of the accepted links.
- The overall reliability in the interval [LOW,UPP] is also a lower boundary for the reliability of the accepted links.

We therefore estimate that the accepted links have a reliability greater than 97.9% in the Eastern region, greater than 90% in Quebec, greater than 98.8% in Ontario, and greater than 98.8% in the Western region.

It should lastly be noted that often in linkage procedures, the approach used is one that seeks to retain as many links as possible. In such cases, the LOW and UPP boundaries are set much less strictly than was done for the RRC-RCT linkage. In that situation, using the method described here could prove to be ineffective or even discouraging, since the reliability calculated by means of logistic regression is a lower boundary for the reliability of the accepted links. In some cases, that boundary could be very low although

the overall rate of false links is acceptable. In such cases, it may be preferable to instead use a sample of the accepted links to estimate reliability generally.

Linkage Results and Conclusion

After choosing LOW and UPP and determining the links to be retained (either automatically or by manual checking), we obtain the following linkage rates for Canada's different geographic regions:

Table 3. -- Linkage Results by Region

Region	Eastern	Quebec	Ontario	Western
Selected person(SP) linked	57%	54%	58%	54%
SP not linked but other family member linked	36%	35%	31%	36%
No linkage	6%	10%	9%	9%
Linkage not attempted	1%	1%	2%	1%
Sample size	12,440	7,328	9,243	16,820

As may be seen, an update address is obtained for more than half of the selected persons, with an address reliability greater than 90%. As regards persons who are not linked, in many cases another member of the household is linked, so that we can nevertheless obtain a valid address for tracing in roughly an additional 35% of cases.

These results should enable us to obtain a satisfactory response rate for the RCC96.

To verify the linearity of the relationship between the logit of reliability and weights W, we grouped the weights into intervals and worked with the midpoints of these intervals. For the two regions where a model has been used, the model obtained in this way is very close to the one obtained by means of logistic regression. This confirms that the logistic model functions well for predicting the reliability of the links in the manual checking range. This model could also be used on a sample of links checked during the linkage procedure, so as to determine UPP and LOW points that result in both an acceptable level of reliability and a reasonable amount of manual checking, or even to choose to change the linkage rules if we suspect that it will not be possible to achieve these two objectives simultaneously.

References

Armstrong, J. B. and Mayda, J. E. (1993). Model-Based Estimation of Record Linkage Error Rates, *Survey Methodology*, 19, 2, 137-147.

Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, 64, 1183-1210.

Chapter 6

Linking Federal Estate Tax Records

Jenny B. Wahl, St. Olaf College

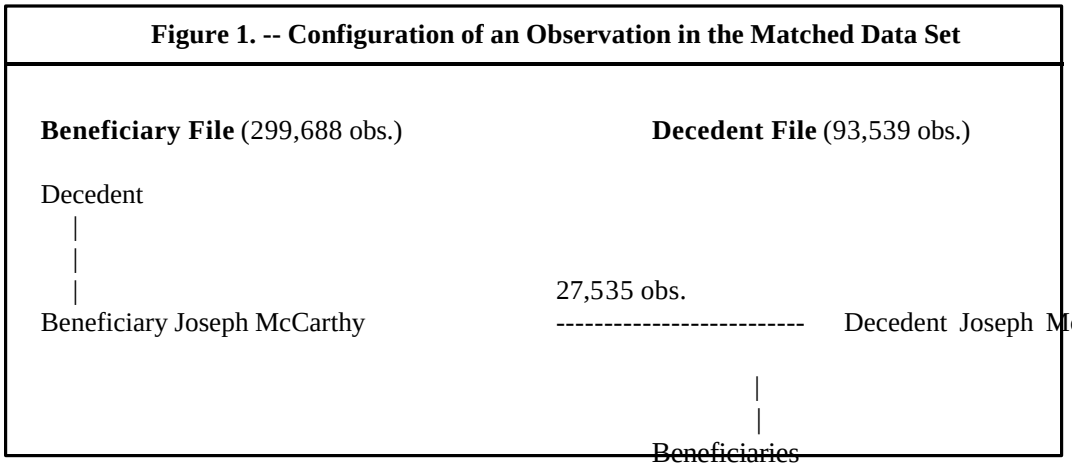
Abstract

This chapter focuses on the construction of a dataset that links together tax records and contemplates possible uses of these data. I first provide an overview of scholarly work regarding inherited wealth and establish the need for intergenerationally linked data. I then discuss techniques that I used to work with Federal estate tax returns filed in Wisconsin up to 1981 (which included 93,539 decedents and their 299,688 beneficiaries). By combining a standardizing/matching software package with a series of SAS programs, I linked these records to form a database containing 27,535 observations. Each observation has information on an individual who was reported on at least two estate tax returns: once as a decedent and at least once as a beneficiary. Of the 27,535 observations, 6,453 are matched pairs and the remaining 21,082 are likely pairs. I conclude by revealing certain problems associated with linking together tax records and by suggesting future research.

Introduction

The only sure things in life are death and taxes – and, unfortunately for some, death taxes. Fortunately for the rest of us, Federal estate tax data offer a rare opportunity to observe the total wealth, portfolios, and bequest behavior of certain individuals. Not only that, these data can be linked across generations, providing testing grounds for hypotheses about motives for intergenerational transfers, tradeoffs of family size and bequest amount, and the like. I have used all the estate tax records filed in Wisconsin from 1916 to 1981 to assemble just such a data set. These data consist of 27,535 observations; each observation has information on a single individual who was reported on two estate tax records: once as a decedent and once as a beneficiary [1]. Of this number, 6,453 are matched pairs and the remaining 21,082 are likely pairs. Figure 1 illustrates the configuration of each observation. In addition to the linked data, a residual set of 272,153 beneficiaries did not match to any decedent.

Figure 1. -- Configuration of an Observation in the Matched Data Set



What follows is, first, a brief overview of some of the questions and theories that scholars have put forth regarding wealth and intergenerational transfers. I then turn to a fuller description of the data, a discussion of the linking methodology, and a short mention of the empirical work that lies ahead.

The Importance of Inherited Wealth

For a variety of reasons, scholars have studied the transfer of wealth across generations. Some have focused on macroeconomic issues such as the influence of wealth transfers on the distribution of wealth (Menchik, 1979; Kotlikoff and Summers, 1981; Modigliani, 1988; and Tachibanaki, 1994), the degree to which intergenerational wealth transfers affect savings rates across countries (Darby, 1979; Hayashi, 1986; and Kotlikoff, 1988), and the interaction of cross-generational transfers and fiscal policy (Barro, 1974; and Aaron and Munnell, 1992). Others have concentrated on microeconomic questions such as the propensity of parents to compensate their less able children or, alternatively, to leave more money to their relatively capable offspring (Becker and Tomes, 1979; and Tomes, 1981).

In the process, researchers have speculated as to the appropriate model of behavior. Do individuals leave bequests because they care about their descendants or other heirs? Or do people design bequests strategically to induce potential heirs to offer attention and companionship? Or might the leaving of an estate simply be a mistake born of miscalculating one's own mortality? (Kotlikoff and Spivak, 1981; Bernheim, Schleifer, and Summers, 1985; Abel, 1985; Hurd, 1987; Modigliani, 1988; Lord and Rangazas, 1991; Altonji, Hayashi, and Kotlikoff, 1992; Gale and Scholz, 1994; Abel and Kotlikoff, 1994; Hurd, 1994; Arrondell, Perelman, and Pestieau, 1994; Yagi and Maki, 1994; and Tachibanaki and Takata, 1994.) Professors Martin David and Paul Menchik (1982) took yet a different tack. They used wealth data to estimate propensities to bequeath out of earnings. Although they did not propose any new theories, David and Menchik cast doubt on an old one: their results indicated that the life-cycle hypothesis cannot explain the bequest behavior of a set of Wisconsin decedents [2].

Others have posed additional interesting questions. Do people behave differently – choose alternative occupations or retire early, for example – if they receive or anticipate a bequest? What relationship do estate size and life insurance bear to a decedent's earnings? What connections exist among fertility, estate size, and earnings? Can one find evidence, for instance, of a tradeoff between the number of children and the wealth left to each one (Becker and Tomes, 1976; Behrman, Pollak, and Taubman, 1982; Wahl, 1986; and Wahl, 1991)? Do people tend to allocate estates equally among their children? Are people increasingly "spending the kids' inheritance," as the bumper stickers proclaim? What patterns in charitable giving have appeared over the years? Is age at death related to lifetime earnings? Many of these questions remain open. Answering them requires a sufficiently large, intergenerationally linked data set that contains comprehensive demographic and socioeconomic information.

The Original Estate Tax Data: Saved in the Nick of Time

Estate tax records contain a wealth of data on a nation's citizens. One can find not only detailed information on accumulated capital and portfolio holdings but also clues about family composition, residence and migration patterns, fertility, and mortality. By dint of much effort (and good computer software) one can even link records together to reconstitute families and their financial and demographic histories. I have drawn upon Federal estate tax records to do just this.

Let me offer a short history of the initial data collection effort. In 1916, the modern Federal estate tax came into being – well before modern computers, but recently enough that paper documents still existed in archives seven decades later. In efforts to clean house during the Reagan years, zealous politicians nearly caused an untimely end for the boxed estate tax returns that were scattered in warehouses around the coun-

try. Fortunately, the Statistics of Income (SOI) Division at the Internal Revenue Service marshaled its forces to preserve these important historical artifacts in computerized form. The result was two enormous files: one consisting of economic and demographic information on decedents, the other of information on beneficiaries (linked via record number to the original estate tax record).

Any attempt to match these two files required reducing their size. Because other researchers have used Wisconsin data to investigate wealth and estate issues (for example, David and Menchik, 1982), SOI extracted all the Wisconsin estate tax returns to use for a pilot project. The result was a decedent file with 93,711 observations and a beneficiary file with 300,269 observations. In the decedent file, 93,539 are unique individuals. For consistency's sake, omitting records from the decedent file meant purging the same records from the beneficiary file. The outcome was a file of 299,688 beneficiaries. Of this number, 188 seem to be duplicates on the same estate tax record – that is, beneficiaries with the same name and same relationship code to the decedent, but appearing twice on a given tax return. Such apparent duplicates may, however, represent different persons with the same name – cousins, for example. Alternatively, these may constitute separate bequests to a single individual – one direct and one in trust. Rather than investigate these observations before the match procedure, I simply marked them so that, if any appeared after the match, I could inspect them more carefully at that time.

Linking the Data: Overlapping Estate Tax Returns

Linking data from one set of records to another requires much information and, frequently, creative computer programming (Fellegi and Sunter, 1969). The AUTOMATCH software written by Matt Jaro provides a solid foundation (Jaro, 1997); variations on his programs coupled with SAS programming produced the linked estate tax records. The critical linkage was this: Joseph McCarthy, say, appears as a beneficiary on his father's estate tax return. In turn, the estate of Joseph McCarthy also files a tax return. The two are linked into a single observation, given consistency in social security numbers, sex, years of birth and death, and the like. Each observation then contains detailed information about the Joseph the decedent: his portfolio, age, marital status, and number of children, for instance. Information about Joseph the beneficiary appears as well: his relationship to his benefactor, receipt of a trust, and sometimes the size of his bequest.

The AUTOMATCH software contains several attractive features that help create good links between records. It standardizes individual names and creates NYSIIS and Soundex codes. (Because I had maiden names for many women, I ran the standardization/coding step twice.) These codes work well as blocking variables in the match process. The software also allows specification of values for missing variables; this helps distinguish between true mismatches and apparent mismatches caused by missing data. The match procedure itself allows multiple rounds so that I could block and match over different sets of variables. Table 1 shows the salient variables for each match round.

The matching process itself also has nice characteristics. I could request multiple matches -- important, because Joseph McCarthy may have inherited from more than one person. Each matching variable has a designation to control for miskeying in the original data. For example, I could allow for mismatched numbers in the social security number string and mismatched letters in the name character strings. These designations also allow matching around intervals, which proved essential for my year-of-birth variables because I had to construct them from rounded-year ages. Each matching variable also carries a set of probabilities to allow for type I and type II errors [3]. All together, these probabilities translate into a single weight associated with each match in each match round. I could choose two cutoff weights per round: one the lower bound for declared matches, the other the lower bound for potential matches. After each match round, I could perform an interactive clerical review on the potential matches and change their status to declared matches or residuals. Following the clerical review, the software outputs all residuals to the next match round.

Table 1. -- Matching Rounds

Match Pass	Blocking Variables	Matching Variables	Original Matches	Original Clerical	Final Matches	Final Clerical
1	SSN	surname first name maiden name suffix initial sex year of birth	4,805	119	4,884	0
2	SSN	surname/ maiden name first name suffix initial sex year of birth	4,906	43	4,928	0
3	surname NYSIIS first name Soundex sex	SSN surname first name maiden name initial suffix year of birth	5,514	30,651	5,514	30,651
4	surname/ maiden name NYSIIS first name Soundex sex	SSN surname/ maiden name first name initial suffix year of birth	5,515	30,652	5,515	30,652

The clerical review process is extremely time-consuming. Although I used it for the first two match rounds, thereafter I used SAS programs to decide whether to change the status of potentially matched pairs [4]. Simply put, I distilled a set of decision rules into SAS programs rather than using the same rules on an interactive, case-by-case basis. For example, suppose the initial matching process paired Joseph McCarthy from the decedent file to Joseph McCarthy from the beneficiary file. The beneficiary file includes a date of death for the Joe's benefactor. If this date of death was after the date of death of Joe the decedent, I called it a nonmatch.

Particular Features of Estate Tax Data

Any two data sets have quirks that make matching difficult. Let me point out a few issues associated with matching data on people observed at two different points in time, often several years apart.

Some problems pertained primarily to females. During the time period covered by my data, a woman often took her husband's social security number at marriage. Sorting and matching by SSN for women was therefore problematic if a woman got married after receiving a bequest. Women also sometimes changed their middle initials upon marriage to reflect their maiden names. I had to take care, then, with the probabilities placed on type I and type II errors when initials appeared as matching variables.

Yet women provided information – namely, maiden names -- that helped me refine the likelihood of matches as well. Suppose a decedent carried the maiden name Scheuren. Say that the decedent potentially matches to a beneficiary, whose benefactor carried the last name Scheuren. Provided that birth and death years were logical, I could declare this a match. By the same token, if a (potentially matched) decedent had the last name Winkler and the benefactor named on the beneficiary file had the maiden name Winkler, again this might be considered a match.

Males created certain problems as well, albeit less directly than females. I had hoped to use cities as matching variables. Yet this hope was dashed: Wisconsin men seemed to like passing their names on to their sons, people did not seem to move around much, and missing ages for beneficiaries frequently meant that I could not screen matches by birth year. As a result, I could not use locational variables to improve the matching process.

A last discovery: one should always assign unique record identification numbers to observations on each file. Initially, the beneficiary file contained identifiers that pointed back to the estate tax record, but it did not have unique identifiers. Because my original files were so large, I excluded some variables while performing the match. When I attempted to reattach data after the match, I could not be sure that the right data went to the right individual. I therefore had to retrace my steps, this time with unique identifiers for each original file.

What Lies Ahead

In the coming months, I will use these linked data to fulfill two objectives. One is to compare matched and unmatched beneficiaries and report any significant differences. The other is to generate a proxy for bequest amount. To proceed, I must convert dollar figures to constant-dollar amounts, control for changes in filing thresholds, and implement a logical cutoff process so as to separate nonmatches from impossibilities. That is, I do not want to call unmatched data a "nonmatch" if the individual could not possibly have entered the matched data set because he or she was born before 1916 or was still living after 1981. Eventually, I hope to extend matches forward and back to reconstitute multiple generations of families.

Acknowledgments

Many thanks to Fritz Scheuren, who made sure this work saw the light of day; Dan Skelly, who encouraged me to dirty my hands with these data; and Barry Johnson, who answered my questions and did most of the hard stuff.

Footnotes

[1] Individuals can appear as beneficiaries on more than one estate tax return. The pairs do not therefore represent unique persons.

[2] The life-cycle hypothesis, associated originally with Franco Modigliani, suggests that people tend to decumulate wealth after a certain age, as they begin to anticipate death. For a review, see Ando and Modigliani (1963) and Modigliani (1988).

[3] Type I errors occur when true matches are declared nonmatches; Type II errors occur when nonmatches are declared matches.

[4] Here is a time comparison: using the clerical review process on 3,827 potential pairs took me seven hours. Writing and running SAS programs with embedded decision rules took about one-half hour for the same data.

References

Aaron, H. and Munnell, A. (1992). Reassessing the Role for Wealth Transfer Taxes, *National Tax Journal,* 45: 119-44.

Abel, A. (1985). Precautionary Saving and Accidental Bequests, *American Economic Review,* 75: 777-91.

Abel, A. and Kotlikoff, L. (1994). Intergenerational Altruism and the Effectiveness of Fiscal Policy – New Tests Based on Cohort Data, *Savings and Bequests*, ed. T. Tachibanaki, Ann Arbor: University of Michigan Press, 167-96.

Altonji, J.; Hayashi, F.; and Kotlikoff, L. (1992). Is the Extended Family Altruistically Linked? New Tests Based on Micro Data, *American Economic Review,* 82: 1177-98.

Ando, A. and Modigliani , F. (1963). Lifecycle Hypothesis of Savings: Aggregate Implications and Tests, *American Economic Review,* 53.

Arrondell, L.; Perelman, S.; and Pestieau, P. (1994). The Effect of Bequest Motives on the Composition and Distribution of Assets in France, *Savings and Bequests*, ed. T. Tachibanaki, Ann Arbor: University of Michigan Press, 229-44.

Barro, R. (1974). Are Government Bonds Net Wealth? *Journal of Political Economy,* 82: 1095-1118.

Becker, G. and Tomes, N. (1976). Child Endowments and the Quantity and Quality of Children, *Journal of Political Economy,* 84: 143-62.

Becker, G. and Tomes, N. (1979). An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility, *Journal of Political Economy,* 87: 1153-89.

Behrman, J.; Pollak, R.; and Taubman, P. (1982). Parental Preferences and Provision for Progeny, *Journal of Political Economy,* 90: 52-73.

Bernheim, B. D.; Schleifer, A.; and Summers, L. (1985). The Strategic Bequest Motive, *Journal of Political Economy,* 93: 1045-76.

Darby, M. (1979). The Effects of Social Security on Income and the Capital Stock, Washington, DC: American Enterprise Institute.

David, M. and Menchik, P. (1982). Modeling Household Bequests, University of Wisconsin, working paper.

Fellegi, I. and Sunter, A. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association,* 64: 1183-1210.

Gale, W. and Scholz, J. K. (1994). Intergenerational Transfers and the Accumulation of Wealth, *Journal of Economic Perspectives,* 8: 145-60.

Hayashi, F. (1986). Why is Japan's Saving Rate So Apparently High, *NBER Macro Annual,* ed. S. Fisher, Cambridge: MIT Press.

Hurd, M. (1987). Savings of the Elderly and Desired Bequests, *American Economic Review,* 77: 298-312.

Hurd, M. (1994). Measuring the Bequest Motive: The Effect of Children on Saving by the Elderly in the U.S., *Savings and Bequests*, ed. T. Tachibanaki, Ann Arbor: University of Michigan Press, 111-36.

Jaro, M. (1997). Matchware Product Overview, *Record Linkage Techniques – 1997*, eds. W. Alvey and B. Jamerson, Washington, D.C.: Office of Management and Budget.

Kotlikoff, L. (1988). Intergenerational Transfers and Savings, *Journal of Economic Perspectives,* 2: 48-51.

Kotlikoff, L. and Spivak, A. (1981). The Family as an Incomplete Annuities Market, *Journal of Political Economy* , 89: 372-91.

Kotlikoff, L. and Summers, L. (1981). The Role of Intergenerational Transfers in Aggregate Capital Accumulation, *Journal of Political Economy,* 89: 706-32.

Lord, W. and Rangazas, P. (1991). Savings and Wealth in Models with Altruistic Bequests, *American Economic Review,* 81: 289-96.

Menchik, P. (1979). Intergenerational Transmission of Inequality: An Empirical Study of Wealth Mobility, *Economica* , 46: 749-62.

Modigliani, F. (1988). The Role of Intergenerational Transfers and Life Cycle Saving in the Accumulation of Wealth, *Journal of Economic Perspectives,* 2: 15-40.

Tachibanaki, T, ed. (1994). *Savings and Bequests*, Ann Arbor: University of Michigan Press.

Tachibanaki, T. and Takata, S. (1994). Bequest and Asset Distribution: Human Capital Investment and

Intergenerational Wealth Transfers, *Savings and Bequests*, ed. T. Tachibanaki, Ann Arbor: University of Michigan Press, 197-228.

Tomes, N. (1981). The Family, Inheritance, and Intergenerational Transmission of Inequality, *Journal of Political Economy*, 89: 928-58.

Wahl, J. (1991). American Fertility Decline in the Nineteenth Century: Tradeoff of Quantity and Quality? *Essays in Honor of Robert William Fogel*, eds. C. Goldin and H. Rockoff, Chicago: University of Chicago Press.

Wahl, J. (1986). New Results on the Decline in Household Fertility in the United States from 1750 to 1900, *Studies in Income and Wealth*, eds. R. Gallman and S. Engerman, Chicago: University of Chicago Press, 391-438.

Yagi, T. and Maki, H. (1994). Cost of Care and Bequests, *Savings and Bequests*, ed. T. Tachibanaki, Ann Arbor: University of Michigan Press, 39-62.

Chapter 6

Post-Matching Administrative Record Linkage Between Sole Proprietorship Tax Returns and the Standard Statistical Establishment List

Philip M. Steel and Carl A. Konschnik, Bureau of the Census

Abstract

In 1992 a match was performed between the IRS Form 1040, Schedule C file and the Standard Statistical Establishment List (SSEL). The match supplemented existing linkages already established between the two files. Though no matching operation has been performed on subsequent 1040 files, the links established on the 1992 data continue to be used in the processing of these files. We are now in a position to analyze the long term effectiveness of the procedure and how frequently it should be applied.

As a by product of the matching operation we obtained a measure of the fit between two records. We explore the possibility of utilizing this measure to link records or select records to be subjected to a matching procedure. The measure of fit derived from the 1992 processing can also be applied to test the validity of existing linkages derived from other procedures.

Introduction

This paper describes a matching process which improves the linkage between sole proprietorship income tax return records from the Internal Revenue Service (IRS) and their associated payroll records on the Census Bureau's Standard Statistical Establishment List (SSEL).

The matching process supplements the linkages made previously based on a common primary identifying number on the two types of records. This number is the Employer Identification Number (EIN), issued by IRS to businesses with employees, and used by them as a principal taxpayer identification number. Unfortunately this common identifier is omitted on roughly 30 percent of the annual income tax returns on which it should appear. In matching, our aim was to make the linkages more complete by using other information besides the EIN -- chiefly, name, city, state, ZIP code, payroll and kind-of-business activity code.

Context and Motivation for the Matching

Linking receipts and payroll records depends largely on associating the correct EIN with each annual income tax return. A sole proprietorship business, when filing the required annual Form 1040, Schedule C, (or, briefly, 1040-C) tax return with the IRS, uses the owner's social security number (SSN) as its taxpayer identification number. If the business has employees, it is required to have an EIN and use it for filing IRS Form 941, Employer's Quarterly Federal Tax Return. When it files its annual 1040-C tax return, the sole proprietorship business is asked to provide its EIN if it has one. This reported EIN is the principal link between the annual business income and quarterly payroll tax returns for sole proprietorship employers.

The IRS provides Form 941 payroll data to the Census Bureau weekly for updating the SSEL. These payroll records, along with data received monthly from the IRS Business Master File (BMF), serve to keep the SSEL current with name and address, employment, payroll, form of organization, and other key data for business. The primary identifier used for the BMF and the 941 files, is, of course, the EIN. By natural extension, for processing administrative records, the EIN is also the primary identifier on the SSEL. All employers -- corporation and partnership employers, as well as sole proprietorship employers, file their Form 941 under their EIN. Because partnership and corporation income tax returns are filed under an EIN, the linkage between receipts from annual tax returns and payroll records for these businesses is readily available. However, for sole proprietorships, if the EIN is missing or incorrect on the 1040-C, we obviously can't rely on the EIN to update the appropriate SSEL payroll record with 1040-C receipts.

Complete updating of receipts on the SSEL is important because the Census Bureau's economic censuses use the SSEL as a frame and use IRS tax return data from the 1040-C to tabulate receipts for single-establishment (singleunit), sole proprietorship businesses with payroll below prescribed cutoff levels. These cutoffs vary by kind of business. Singleunit businesses with payroll above the cutoffs and all multi-establishment (multiunit) businesses from the SSEL are mailed a census form. Tax return data from the 1040-Cs are also used to account for those who fail to respond to the mailing.

Incomplete linkage between 1040-C employers and the SSEL means that the file of 1040-C records from IRS for a census year such as 1992 (after removing 1040-C linked to the SSEL) still contains some employer as well as all nonemployer businesses. This causes two problems:

- tax-return receipts are not available on the SSEL for tabulating inscope EINs with nonzero 1992 payroll and missing receipts for the economic censuses; and

- the 1040-C file includes an unknown number of employers with unknown total receipts -- therefore, we cannot use it directly to tabulate census year receipts for nonemployer businesses.

Both problems are alleviated by improving the linkage between the file of 1040-C records and the SSEL. For the 1992 censuses, we obtained an EIN to SSN cross-reference (x-ref) file from IRS to aid in linking records. In addition to this, we used matching techniques to associate 1040-C records with their associated SSEL payroll record. In the following sections, we present the technical details of this matching work and discuss the impact it had on the 1992 census estimates.

Description of the Files for Matching

After updating the SSEL using the reported EIN on the 1040-C and the SSN to EIN cross-reference file from the BMF, we were left with EINs on the SSEL which were still missing receipts. A file of these EINs drawn from the SSEL, formed the primary file for the matching. The number of unlinked, potentially matchable EINs, on the file at this point was 419,494. The criteria for selecting these cases were that:

- the EIN be within U.S. boundaries;

- the Legal Form of Organization (LFO) be a sole proprietorship or form of organization unknown (as opposed to a partnership or a corporate form of organization);

- the EIN be taxable or have tax status unknown; and

- the EIN reported nonzero payroll for the 1992 census year.

The second file for matching consisted of 1040-C sole proprietorship tax return records. A 1040-C may have (in census processing) up to three schedules, each representing a separate business. The schedule, together with the name and address from the main 1040, form our 1040-C record. The SSN, together with one schedule number, formed the identifier for the second file. The original 1040-C file included those with EINs that linked to the SSEL, this file contained 16,540,844 schedules in all. Because this large file size exceeded the capacity of our software platform (a VAX minicomputer cluster), our matching was performed with this file split into 36 pieces.

Variables for Matching and their Comparability

Records on both files contain name, address, kind-of-business and payroll fields. Each of these fields has associated problems.

- **Name Fields**. -- The primary name field from the SSEL may be the name of a business, e.g., the American Bank Note Company, but in the case of sole proprietors this field is usually the proprietor's name, even where the LFO has not been determined. On the other hand, the 1040-C record has the Form 1040 name, which is a personal name, often including both husband and wife for joint filing. There are several other name fields available on the SSEL, such as the census name, physical location name, and mailing name. These were examined as candidates for matching fields, but appeared to contribute very little to establishing new linkages (during testing, the census name field update was incomplete, and may yet be shown to be useful for future matching). To summarize, on the EIN file we have a name field that may or may not contain a personal name; on the 1040-C file we have a name field that may contain compound names, with either of the components a candidate for matching. To deal with this, our name parser rejects records with identifiable business names from the EIN file and generates two records for compound names on both the 1040-C file and (in a few cases) the EIN file.

- **Address Fields**. -- Both files have address fields containing street address, city, state and ZIP code. However, the address on the SSEL is generally the business address and the 1040-C address is a personal address. Using a test file containing only known linkages between the EIN and 1040-C records, we found that the street addresses matched partially or better only 30 percent of the time (+/- 6%) based on a clerical review of a sample of 248 cases. This eventually led us to drop the street address as a matching variable. The same problem -- that the 1040-C address and SSEL address of known linkages can be different -- applies to city, state and ZIP code, but to a lesser degree. These variables were retained for matching.

- **Business Classification Codes**.--The EIN's business activity code from the SSEL is the Standard Industrial Classification (SIC) code, whereas the 1040-C record has a converted Primary Business Activity (PBA) code. The PBA code is an abbreviated but roughly comparable coding system. The SIC on the SSEL is generally coded from various sources. In contrast, the PBA is a self-reported code by the taxpayer. Previous studies indicate that we can expect the self-reported code to match the SIC code no more than about 67 percent of the time at the four-digit level and 75 percent at the two-digit level. See Konschnik et al., (1993) for more on the quality of self-coded PBAs.

- **Annual Payroll**.--We obtain a single annual payroll figure for EIN records from the SSEL. The 1040-C has two fields related to payroll -- wages and cost of labor. Technically, the wages field is supposed to correspond to payroll, and cost of labor to represent contracted labor where the em-

ployment taxes are born by the employer of the contracted laborers. Examination of the data shows this is not always the case. Although for most of the time, the SSEL payroll figure agrees with the wages figure from the 1040-C, the exact number sometimes appears in the cost of labor field or even split across both fields. Whether this is due to taxpayer reporting error, or keying error is an open question. Since legitimate (nonpayroll) data also appears in the cost of labor field, a statistical solution was required.

Software Used for the Matching

For the matching software, we used Winkler's mf3 matcher, with match specific modifications. We used both character-by-character comparisons and one of the native string comparators. For the numeric comparison on the payroll variable, we developed a new module, about which we will go into in some detail.

The EIN records from the SSEL were extracted and "prepped," forming a "stationary" file of 351,141 records. The 1040-C files were preprocessed and matched in 36 separated cuts of roughly 750,000 records (each).

Blocking Criteria

The blocking criteria, defined as the minimum characteristics necessary to consider a pair of records in the match, were the first six letters of the last name and the first letter of the first name. We originally explored the possibility of blocking by ZIP code but abandoned this when we realized the scope of the problem in business versus home addresses.

Matching Variables and Weights

The fit between any pair of records is determined by the sum of the weights of the match variables. We assign positive weights for agreement and negative weights for nonagreement. Below, in Table 1, is a list of the match variables, along with their positive and negative weights. A record from the 1040-C file is considered a match to a record from the SSEL file when the pair's match score (sum of weights) exceeds 15.15.

The match variables fall into three groups: Name, Location, and Business. This suggests the general strategy we employed for determining the weights. The role of the Name group was to further (beyond blocking) qualify pairs -- a failure on more than one of the name variables here should disqualify a record. The other two groups were weighted to balance one another -- a weak score on Location required a strong score on Business and vice versa, with Business given slight precedence over Location.

Table 1. — Match Variables by Weights

Group	Description	Positive Weight	Negative Weight
Name	last name	5.01	-8.11

	first name remainder	5.01	-7.82	
	middle initial	3.00	-8.06	
	middle name remainder	2.18	-0.01	
Location	city		3.04	-0.00
	state	0.00	-6.31	
	5 digit zip code	3.04	-0.00	
	first 3 digits of zip	3.04	-0.00	
	first 2 digits of zip	3.00	-2.00	
Business	entire SIC		3.06	-0.00
	first 2 digits of SIC	3.00	-3.00	

The annual payroll variable was handled somewhat differently when determining weights. The payroll variable looks at the ratio of 1040-C payroll (wages+cost of labor, combined in the prep phase) + 5,000 to SSEL payroll + 5,000. Calling this ratio R, weights were assigned based on the interval in which R fell (see Table 2). The factor of 5,000 keeps a small absolute difference of say 1,000 (possibly a rounding error) from making R too large or too small.

Table 2. — Weights for Payroll Variables

Range	Weight
0.00 < R <= 0.64	-7.00
0.64 < R <= 0.87	0.00
0.87 < R <= 0.93	4.50
0.93 < R <= 1.05	7.50
1.05 < R <= 1.13	4.50
1.13 < R <= 2.25	0.00
2.25 < R	-7.00

How the Model for Relating the Payroll Variables was Determined

We constructed two files to test competing models for the payroll comparison. A file of randomly joined payrolls from a known sole proprietors file and a sample file of 1040-Cs was created (random set). Both payrolls were taken from an EIN linked file of sole proprietors to create the second file (truth set). Next, we tested three models as shown below.

Model 1: $\frac{\text{wages} + X}{\text{SSEL payroll} + X}$

Model 2: $\frac{\text{wages} + \text{cost of labor} + X}{\text{SSEL payroll} + X}$

Model 3: $\frac{\text{cost of labor} + X}{\text{SSEL payroll} + X}$, if wages=0; $\frac{\text{wages} + X}{\text{SSEL payroll} + X}$, otherwise.

The discriminating power of the variable must contrast the behavior over the truth set against its behavior on the random set. The addition of a term to top and bottom of the ratio pulls the distributions toward 1. In fact, the distribution centralizes faster on the truth set than it does on the random set.

The criteria for selection was to select the model that produced the most ratios near 1 and the fewest ratios at the extremes on the truth set, and simultaneously produced the fewest ratios near 1 and the most at the extremes on the random set. Models 2 and 3 were clearly better than Model 1, Model 3 slightly better than Model 2. Model 3 was a later invention and did not make it into production. For the selected model, value of X = 5,000, and the four most critical conditions, we have:

P(strong agree|match) = 53.05

P(strong agree|nonmatch) = 0.4

P(disagree|match) = 12.3

P(disagree|nonmatch) = 88.8.

Match Results

The Parser

The parser behaved very differently on the two files. The 1040-C name field is highly structured, generally well keyed, and contains no legitimate business names. The SSEL name field may have a sole proprietor's personal or business name, or it may have the name of a corporation or partnership -- this latter group a contribution from the unknown LFO. The personal names include more abbreviations and are less structured.

Looking at the results of the parser on the 1040-C file (excluding schedules linked to the SSEL), we see that 23,670 of 14,894,578 (0.16 percent) schedules failed to parse and were not included in the match. Almost all failures were due to complicated name structures. About 10.6 million records with duplicate identifiers were created from joint returns -- i.e., roughly 10 duplicates for every 14 schedules. With duplicates, the prepped 1040-C file had 25,429,164 records.

From a test of confirmed sole proprietors (of about 19,000 records) we know that the parser succeeds about 97.4 percent of the time. The unparsed SSEL file, which included records with LFO unknown, had 419,494 records -- 332,441 of which parsed. Using the known rate we can deduce that the unparsed file contained about 341,315 sole proprietors (virtually none of the non-sole-proprietors parse). Hence, we have an estimated 8,874 sole proprietor establishments whose name failed to parse, and, consequently, were not included in the match. Roughly 25 percent of failures were due to unrecognized name patterns, the remainder were recognized as business names. We can infer from this that sole proprietors use a business name on the SSEL rather than their personal name about 1.9 percent of the time. This is computed by

(8,874)(.75)/341,315. In the matter of duplication, in contrast to the 1040-C file, only about 5.6 percent of the parsable names on the SSEL file generate duplicates.

Unduplication

There were several varieties of duplicates among the files produced by the match. In all cases, the pair with the highest match score was designated to be the match. In the event of a tie, the first pair was taken or both discarded depending on the type of duplication. Unduplication proceeded first by EIN then by SSN/Schedule Number.

Over half the duplication was caused by duplicates created in the name parse. In effect, the matcher picks two best candidate for these records. Ties frequently occurred where husband and wife appeared in the name field of both records. For duplicates fitting this pattern, both candidates having the same EIN and the same SSN, the pair with the highest match strength. In event of a tie, the first record was taken.

When pairs were presented with the same EIN and different SSNs, the highest match strength was taken. In the event of a tie, no match was made for that EIN. Family businesses seemed to be the main cause for ties. The file of ties has been retained for further study.

After the EIN side was resolved, the file was resorted to look for instances of the same Schedule C attempting to match more than one EIN record. This occurred almost exactly 1 percent of the time. Again, if the matcher rated one pair higher than all others, this pair was designated a match. Otherwise, although rarely, the first instance of the tied match strength was taken. An examination of the file of duplicates and winners revealed the following common pattern. Husband and wife had distinct businesses on the SSEL, each under their individual names. In theory each business should correspond to a distinct schedule, but for some reason one of the businesses did not fit any of the schedules. If the 1040-C had only one schedule the same error would occur.

Table 3 gives the unduplication by type and the final number of designated matches.

Table 3. — Unduplication of Matches

Unduplication Type	No. of Records
Match pairs before unduplication	156,836
EIN unduplication (records dropped):	
Same EIN, same SSN, same schedule no.	5,211
Same EIN, same SSN, diff. sched. no.	100
Same EIN, diff. SSN, same match strength	506
SSN Unduplication (records dropped):	
Same SSN, diff. EIN, same schedule	156
Match pairs after unduplication	148,679

Match Error Rates

Modeling the Error

Our approach to error estimation is to study a population, similar to the match population, where the link between pairs has already been established. From a 1 in 50 master sample of the original 1040-C file, we identified 18,595 records that reported an EIN on the 1040-C, are known matches to a record on the SSEL, and meet the following conditions:

- the EIN was valid;
- the EIN was reported on only 1 schedule;
- the EIN record had a sole proprietorship LFO compatible with a 1040-C filing;
- the pair passed a mild payroll/receipts edit;
- the name field on the EIN record was parsable; and
- the EIN had positive 1992 payroll on the SSEL.

The count of the 1 in 50 sample that parsed, and including the 18,595 links based on a reported EIN, was 559,514 records. This set of record was used to model two situations: first, where there existed a 1040-C that ought to be linked to the SSEL record ("matchable"); and second, where no record should be linked to an SSEL record ("unmatchable"). By including or excluding the 18,595 linked 1040-C records, and always retaining the linked SSEL records, we modeled both conditions.

False Match Rate for "Matchable" Records

A match was performed with the 559,514 parsed 1040-C records against approximately 361,000 parsed SSEL records. The 1040-C file contained 18,595 linked records, the remaining 540,919 were used to represent the 25,429,164 parsed 1040-C records, giving each a weight of 47. The results of the match on the known links are shown in Table 4, below.

Table 4. — Matches of Linked Records

Condition	No. of Records
True matches	16,364
Type A false matches	100
Type B false matches	1
False nonmatches	2,130
Total	18,595

The type A false matches involved a correct linkage between EIN and SSN, but with the incorrect schedule number. This can only happen within the sample of 559,514, since the sample was based on SSN and every (prior) linked SSN had all schedules present for the match. Thus, type A false matches represent only themselves. The type B false match involved an incorrect linkage between EIN and SSN, and the one occurrence represents approximately 47 others. Since the event is so rare, we calculated an upper bound and used it in subsequent calculations.

The apparent rate is so low, 1 in 540,919, that we are required to estimate from the binomial probabilities -- the normal approximation does not apply and the Poisson approximation is poorest near the mean, where our estimation occurs. We observe that for any $p>4.6/540{,}919$ the probability of getting only 1 occurrence is less than

$$(540{,}919)\ (.0000085)\ (1 - .0000085)^{540{,}919} \approx .05;$$

i.e., if the number of type B false matches were greater than 216 (i.e., 4.6 x 47) across all 25.5 million records, we would have less than a 5 percent chance of getting 1 occurrence in our sample. The question then arises how to distribute the additional 215 estimated false matches between converted true matches and converted false nonmatches. We assume a range on the match score of a false match from 15.15 to 20.51, where 20.51 was the highest false match score observed during all testing. This range contains 1456 of the true matches -- those which are in a range where it is fairly believable that they can be supplanted by a false match. Assuming an even distribution between these and the false nonmatch set, the additional false matches should be allocated by the proportion 1456:2130 or 87:128; i.e., we will take 87 from the true match count and 128 from the false nonmatch count. In Table 5, we estimate the results if we were to run against the whole 1040-C file.

Table 5. — Estimated Matches for 1040C File

Condition	Frequency	Percent of File
True match	16,277	87.5
Type A false match	100	.5
Type B false match	216	1.2
False nonmatch	2,002	10.8

False Match Rate for "Unmatchable" Records

We reran the match excluding the 1040-Cs belonging to the truth set. Four matches were produced, linking a SSEL truth record to a 1040-C when the true 1040-C was suppressed. We can only have (type B) false matches and true nonmatches on this set. The highest match strength among the 4 was 20.5. Again resorting to the binomial calculation, we estimate the upper bound for the number of false matches in a hypothetical run of the whole file to be 390. That is a false match rate of 2.1 percent on "unmatchable" records.

Error Estimation

We are now in a position to recover the composition of the original SSEL file. Let x be the number of "matchable" records and y be the number of "unmatchable" records. Then, using upper bounds and adding the type A and B false match rates, .5% + 1.2% = 1.7%, we must solve:

$$(.875+.017)x + .021y = 148{,}679$$
$$x + y = 332{,}443;$$

i.e., $x = 162{,}684$, and $y = 169{,}759$. From this we computed the upper estimate of the false match rate to be 4.3% as indicated below.

$$(.017 \times 162{,}684 + .021 \times 169{,}759)/148{,}679 = .043.$$

This is an upper bound only. A similar calculation on the point estimate gives an error rate of less than 2

percent.

Conclusion

What was the impact of our matching efforts in the context of the overall task of linking SSEL payroll records of sole proprietors with their 1040-C tax return? For the 1992 tax year, about 1.37 million sole proprietors had their 1040-C tax return linked to their payroll records on the SSEL. The breakdown by source of linkage is given in Table 6.

Table 6. — SSEL-IRS Linkages by Source

Source	Linked Cases
An EIN reported on 1040-C	71%
Use of the EIN-SSN x-ref file	15%
Use of the x-ref file & matching	4%
Matching on name, address, etc.	10%

After all attempts at linking, we still have about 200,000 inscope sole proprietors on the SSEL for which we could not post receipts. Of these, we estimate that about 170,000 had no 1040-C on the file we used. We believe this may be due to non-filing of the 1040-Cs because of extensions or other late filings. We estimate that about 6,000 were linked but failed to have receipts posted because they failed a payroll to receipts edit. About 9,000 failed to match due to parse failures. We have estimated (above) that there are about 15,000 false nonmatches -- i.e., a 1040-C was on the file but the matching program failed to link to it.

The linkage for the 1992 censuses were much more complete than for the 1987 censuses, since in 1987 we used only the reported EIN on the 1040-C for linking purposes. The EIN-SSN x-ref file from IRS provided substantial additional links, and has been improved in subsequent years. These were nearly equaled by our matching. Moreover the additional links continue to contribute to the completeness of the SSEL in subsequent years. A check in 1997 showed 115,000 of the 148,000 still reside on the SSEL, though some portion of these may have reported an EIN or otherwise been linked during the intervening years. This is consistent with a 10-12 percent death process for small sole proprietors. Overall, the matching operations were quite efficient, and added significantly to the quality of the 1992 censuses.

Reference

Konschnik, C.; Black, J.; Moore, R.; and Steel, P. (1993). An Evaluation of Taxpayer-Assigned Principal Business Activity (PBA) Codes on the 1987 Internal Revenue Service (IRS) Form 1040, Schedule C, *Proceedings of the International Conference on Establishment Surveys,* American Statistical Association, 745-750.

Bibliography

Cochran, W.G. (1977). *Sampling Techniques*, Third edition, New York: J. Wiley.

Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*.

Winkler, W. E. (1992). Comparative Analysis of Record Linkage Decision Rules, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 829-833.

Winkler, W. E. (1995). Matching and Record Linkage, in B. G. Cox (Ed.), *Business Survey Methods*, New York: J. Wiley, 355-384 (also in this volume).

Note: A version of this paper appears in the 1994 ASA Proceedings as:
Administrative Record Matching for the 1992 Economic Censuses

Chapter 6

Approximate String Comparison and its Effect on an Advanced Record Linkage System

Edward H. Porter and William E. Winkler, Bureau of the Census

Abstract

Record linkage, sometimes referred to as information retrieval (Frakes and Baeza-Yates, 1992) is needed for the creation, unduplication, and maintenance of name and address lists. This paper describes string comparators and their effect in a production matching system. Because many lists have typographical errors in more than 20 percent of first names and also in last names, effective methods for dealing with typographical error can greatly improve matching efficacy. The enhanced methods of approximate string comparison deals with typographical variations and scanning errors. The values returned by the string comparator are used in a statistical model for adjusting parameters that are automatically estimated by an expectation-maximization algorithm for latent class, log linear models of the type arising in the Fellegi-Sunter model of record linkage (1969). Overall matching efficacy is further improved by linear assignment algorithm that forces 1-1 matching.

Introduction

Modern record linkage represents a collection of methods from three different disciplines: computer science, statistics, and operations research. While the foundations are from statistics, beginning with the seminal work of Newcombe (Newcombe et al., 1959, also Newcombe, 1988) and Fellegi and Sunter (1969) the means of implementing the methods have primarily involved computer science. Record linkage begins with highly evolved software for parsing and standardizing names and addresses that are used in the matching. Name standardization identifies components such as first names, last names (surnames), titles, and middle initials. Address standardization locates components such as house numbers, street names, PO Boxes, apartment numbers, and rural routes. With good standardization, effective comparison of corresponding components of information and the advanced methods described in this paper become possible.

Because pairs of strings often exhibit typographical variation (e.g., Smith versus Smoth), the record linkage needs effective string comparator functions that deal with typographical variations. While approximate string comparison has been a subject of research in computer science for many years (see survey article by Hall and Dowling, 1980), some of the most effective ideas in the record linkage context were introduced by Jaro (1989) (see also Winkler, 1985, 1990). Budzinsky (1991), in an extensive review of twenty string comparison methods, concluded that the original Jaro method, the extended method due to Winkler (1990) and a widely used computer science method called bigrams worked well. This paper describes two new enhancements to the string comparators used at the Census Bureau. The first, due to McLaughlin (1993), adds logic for dealing with scanning errors (e.g., "I" versus "1") and certain common keypunch errors (e.g., "V" versus "B"). The second due to Lynch and Winkler (1994) makes adjustments for pairs of long strings having a high proportion of characters in common. We also describe the method of computing bigrams and present results comparing them with the other string comparators of this paper.

Our record linkage system uses the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) to estimate optimal matching parameters. We use a linear sum assignment procedure (lsap) to force 1-1 matching. Jaro (1989) introduced the lsap as a highly effective means of eliminating many pairs that ordinarily might be clerically reviewed. With a household data source containing multiple individuals in a household, it effectively keeps the four pairs associated with father-father, mother-mother, son-son, and daughter-daughter pairs while eliminating the remaining twelve pairs associated with the household.

The next section describes the string comparator. In the third section, we provide a summary of the parameters that are obtained via the EM algorithm. The results of section four provide empirical examples of how matching efficacy is improved for three, small pairs of high quality lists. The final section consists of a summary and conclusion.

Approximate String Comparison

Dealing with typographical error can be vitally important in a record linkage context. If comparisons of pairs of strings are only done in an exact character-by-character manner, then many matches may be lost. An extreme example is the Post Enumeration Survey (PES) (Winkler and Thibaudeau, 1991; also Jaro, 1989) in which, among true matches, almost 20 percent of last names and 25 percent of first names disagreed character-by-character. If matching had been performed on a character-by-character basis, then more than 30 percent of matches would have been missed by computer algorithms that were intended to delineate matches automatically. In such a situation, required manual review and (possibly) matching error would have greatly increased.

Jaro (1989) introduced a string comparator that accounts for insertions, deletions, and transpositions. In a small study, Winkler (1985) showed that the Jaro comparator worked better than some others from computer science. In a large study, Budzinsky (1991) concluded that the comparators due to Jaro and Winkler (1990) were the best among twenty in the computer science literature. The basic Jaro algorithm is: compute the string lengths, find the number of common characters in the two strings, and find the number of transpositions. The definition of common is that the agreeing character must be within ½ the length of the shorter string. The definition of transposition is that the character from one string is out of order with the corresponding common character from the other string. The string comparator value (rescaled for consistency with the practice in computer science) is:

$$\text{jaro}(s1,s2) = 1/3(\ \#\text{common}/\text{str\_len1} + \#\text{common}/\text{str\_len2} + 0.5\ \#\text{transpositions}/\#\text{common}), \quad (1)$$

where s1 and s2 are the strings with lengths str_len1 and str_len2, respectively.

The new string comparator algorithm begins with the basic Jaro algorithm and then proceeds to three additional loops corresponding to the enhancements. Each enhancement makes use of information that is obtained from the loops prior to it.

The first enhancement due to McLaughlin (1993) assigns value 0.3 to each disagreeing but similar character. Each exact agreement gets value 1.0 and all exact agreements are located prior to searching for similar characters. Similar characters might occur because of scanning errors ("1" versus "l") or keypunch ("V" versus "B"). The number of common characters (#common) in equation (1) gets increased by 0.3 for each similar character, is denoted by #similar, and #similar is substituted for #common in the first two components of equation (1).

The second enhancement due to Winkler (1990) gives increased value to agreement on the beginning characters of a string. It was based on ideas from a very large empirical study by Pollock and Zamora (1984) for the Chemical Abstracts Service. The study showed that the fewest errors typically occur at the beginning of a string and the error rates by character position increase monotonically as the position moves to the right. The enhancement basically consisted of adjusting the string comparator value upward by a fixed amount if the first four characters agreed; by lesser amounts if the first three, two, or one characters agreed. The string comparator examined by Budzinsky (1991) consisted of the Jaro comparator with only the Winkler enhancement.

The final enhancement due to Lynch and Winkler (1994) adjusts the string comparator value if the strings are longer than six characters and more than half the characters beyond the first four agree. The final enhancement was based on detailed comparisons between versions of the comparator. The comparisons involved tens of thousands of pairs of last names, first names, and street names that did not agree on a character-by-character basis but were associated with truly matching records.

A common string comparison methodology is comparing the bigrams that two strings have in common. A bigram is two consecutive letters with a string. Hence the word "bigram" contains the bigrams "bi," "ig," "gr," "ra," and "am." The bigram function also returns a value between 0 and 1. The return value is the total number of bigrams that are in common divided by the average number of bigrams in the two strings. Bigrams are known to be a very effective, simply programmed means of dealing with minor typographical errors. They are widely used by computer scientists working in information retrieval (Frakes and Baeza-Yates, 1992).

Table 1 illustrates the effect of the new enhanced comparators on last names, first names, and street names, respectively. To make the value returned by bigram weighting function more comparable to the other string comparators, we make a functional adjustment. If x is the value returned by the bigram weighting function, we use $f(x) = x^{0.2435}$ if x is greater than 0.8 and 0.0 otherwise. If each string in a pair is less than four characters, then the Jaro and Winkler comparators return the value zero. The Jaro and Winkler comparator values are produced by the loop from the main production software (e.g., Winkler and Thibaudeau 1991) which is only entered if the two strings do not agree character-by-character. The return value of zero is justified because if each of the strings has three or less characters, then they necessarily disagree on at least one.

In record linkage situations, the string comparator value is used in adjusting the matching weight associated with the comparison downward from the agreement weight toward the disagreement weight. Using crude statistical modeling techniques, Winkler (1990) developed downweighting functions for last names, first names, street names, and some numerical comparisons that generalized the original down-weighting function introduced by Jaro.

Table 1. -- Comparison of String Comparators Using Last Names, First Names, and Street Names

Two Strings	Jaro	String Comparator Values			

			Wink	McLa	Lynch	Bigram
SHACKLEFORD	SHACKELFORD	0.970	0.982	0.982	0.989	0.925
DUNNINGHAM	CUNNIGHAM	0.896	0.896	0.896	0.931	0.917
NICHLESON	NICHULSON	0.926	0.956	0.969	0.977	0.906
JONES	JOHNSON	0.790	0.832	0.860	0.874	0.000
MASSEY	MASSIE	0.889	0.933	0.953	0.953	0.845
ABROMS	ABRAMS	0.889	0.922	0.946	0.952	0.906
HARDIN	MARTINEZ	0.000	0.000	0.000	0.000	0.000
ITMAN	SMITH	0.000	0.000	0.000	0.000	0.000
JERALDINE	GERALDINE	0.926	0.926	0.948	0.966	0.972
MARHTA	MARTHA	0.944	0.961	0.961	0.971	0.845
MICHELLE	MICHAEL	0.869	0.921	0.938	0.944	0.845
JULIES	JULIUS	0.889	0.933	0.953	0.953	0.906
TANYA	TONYA	0.867	0.880	0.916	0.933	0.883
DWAYNE	DUANE	0.822	0.840	0.873	0.896	0.000
SEAN	SUSAN	0.783	0.805	0.845	0.845	0.800
JON	JOHN	0.917	0.933	0.933	0.933	0.847
JON	JAN 0.000	0.000	0.860	0.860	0.000	
BROOKHAVEN	BRROKHAVEN	0.933	0.947	0.947	0.964	0.975
BROOK HALLOW	BROOK HLLW	0.944	0.967	0.967	0.977	0.906
DECATUR	DECATIR	0.905	0.943	0.960	0.965	0.921
FITZRUREITER	FITZENREITER	0.856	0.913	0.923	0.945	0.932
HIGBEE	HIGHEE	0.889	0.922	0.922	0.932	0.906
HIGBEE	HIGVEE	0.889	0.922	0.946	0.952	0.906
LACURA	LOCURA	0.889	0.900	0.930	0.947	0.845
IOWA	IONA	0.833	0.867	0.867	0.867	0.906
1ST	IST	0.000	0.000	0.844	0.844	0.947

Data and Matching Weights -- Parameters

In this section, we describe the fields and the associated matching weights that are used in the record linkage decision rule. We do not give details of the EM algorithm or the assignment algorithm because they have been given elsewhere (Winkler, 1994).

The fields used in the creation of mailing list during the 1995 test census are first name, last name (surname), sex, month of birth, day of birth, year of birth, race, and Hispanic origin. The census file is linked with an update file. These update files have been either IRS, driver's license, or school records. Only fields whose housing unit identifier agreed are compared in the first pass. The housing unit identifiers were calculated by the Census Bureau's geography division's address standardization software. It consists of a State Code, County Code, TIGER Line ID (e.g., a city block), Side ID (right or left), house number, and apartment number. In the 1995 test census of Oakland, California 95.0 percent of the records file were geocoded with housing unit identifier. Also, 94.7 percemt of the IRS file records for the corresponding area were geocoded with housing unit identifier. The names were standardized at a 95.2 percent rate in the test census file and 99.0 percent rate in the IRS file.

Each parameter was assigned an agreement and disagreement weight. (See Table 2.) Certain parameters such as first name are assigned a higher agreement weight. Since matching was done within a household, surname carried had less distinguishing power than first name. After initial trial runs and research of the

output, the expectation-maximization software (EM) was run to produce the parameters for the test.

Table 2. -- Parameters Used in Matching for the 1995 Test Census of Oakland, California

Parameter	Agreement Weight	Disagreement Weight
first	4.3385	-2.7119
last(surname)	2.4189	-2.5915
sex	0.7365	-3.1163
month	2.6252	-3.8535
day	3.5206	-2.9652
year	1.7715	-4.1745
Hispanic	0.2291	-0.3029
race	0.5499	-0.5996

String comparators were only used with first names and surnames. For example, if the first names were Martha and Marhta. The matching weight would be computed as follows:

	Jaro	Wink	McLa	Lynch
Comparator Value	0.944	0.961	0.961	0.971
Matching Weight	3.943	4.063	4.063	4.134 .

The piecewise linear function that uses the value returned by the different string comparators to adjust the matching agreement weight downward is detailed in Winkler (1990).

Results

Results are presented in two parts. In each part, the different string comparators are substituted in the string comparison subroutine of an overall matching system. The matching weights returned by the EM algorithm are held constant. Two different versions of a linear sum assignment procedure are used. For the description of the lsap, see Winkler (1994). The main empirical data consists of three pairs of files having known matching status. In the first part, we show how much the string comparators can improve the matching results. The second part provides an overall comparison of matching methods that utilize various combinations of the new and old string comparators and the new and old assignment algorithms.

Exact Matching Versus String Comparator Enhanced Matching

In Table 3, we illustrate how much string comparators improve matching in comparison with exact matching. After ordering pairs by decreasing matching weight in the first and third of the empirical data files, we plot the proportion of false matches against the total number of pairs. We see that, if matching is adjusted for bigrams and the string comparators, then error rates error rates are much lower than those obtained when exact matching is used. Since exact matching is not competitive, remaining results are only presented when string comparators are used.

Table 3. -- Matching Results at Different Error Rates: First Pair of Files with 4,539 and 4,859 Records 38,795 Pairs Agreeing on Block and First Character of Last Name

Link Error Rate	Link Match/Nonm	Clerical Match/Nonm
0.002		
base	3172/ 6	242/64
s_c	3176/ 6	236/64
as	3176/ 6	234/64
os_l	3174/ 6	242/64
bigram	3224/ 7	174/63
0.005		
base	3363/17	51/53
s_c	3357/17	55/53
as	3357/17	53/53
os_l	3364/17	52/53
bigram	3327/17	71/53
0.010		
base	3401/34	13/36
s_c	3396/34	16/36
as	3396/34	14/36
os_l	3402/34	14/36
bigram	3376/34	22/36
0.020		
base	3414/70	0/ 0
s_c	3411/70	0/ 0
as	3410/70	0/ 0
os_l	3416/70	0/ 0
bigram	3398/70	0/ 0

Overall Comparison of Matching Methods

The baseline matching is done under 3-class, latent class models under the conditional independence assumption. The 3-class models are essentially the same ones used in Winkler (1994). In Table 4, results are reported for error rates of 0.002, 0.005, 0.01, and 0.02, respectively. *Link*, *Nonlink*, and *Clerical* (*or Possible Link*) are the computer designations, respectively. *Match* and *Nonmatch* are the true statuses, respectively. The baseline results (designated by *base*) are produced using the existing lsap algorithm and the previous string comparator (see e.g., Winkler, 1990) but use the newer, 3-class EM procedures for parameter estimation (Winkler, 1994). The results with the new string comparator (designated *s_c*) are produced with the existing string comparator replaced by the new one. The results with the new assignment algorithm (designated *as*) use both the new string comparator and the new assignment algorithm. For comparison, results produced using the previous string comparator but with the new assignment algorithm (designated by *os_l*) are also given. Finally, results using the bigram adjustments are denoted by *bigram.*

Table 4. -- Matching Results at Different Error Rates: Second Pair of Files with 5,022 and 5,212 Records 37,327 Pairs Agreeing on Block and First Character of Last Name

Link Error Rate	Link Match/Nonm	Clerical Match/Nonm
0.002		
base	3475/ 7	63/65
s_c	3414/ 7	127/65
as	3414/ 7	127/65
os_l	3477/ 7	63/65
bigram	3090/ 7	461/66
0.005		
base	3503/18	35/54
s_c	3493/18	48/54
as	3493/18	48/54
os_l	3505/18	36/54
bigram	3509/18	42/55
0.010		
base	3525/36	13/36
s_c	3526/36	15/36
as	3526/36	15/36
os_l	3527/36	14/36
bigram	3543/36	8/73
0.020		
base	3538/72	0/ 0
s_c	3541/72	0/ 0
as	3541/72	0/ 0
os_l	3541/72	0/ 0
bigram	3551/73	0/ 0

As Table 5 shows, matching efficacy improves if more pairs can be designated as links and nonlinks at fixed error rate levels. In Tables 3-5, computer-designated links and clerical pairs are subdivided into (true)

matches and nonmatches. Only the subset of pairs produced via 1–1 assignments are considered. In producing the tables, pairs are sorted by decreasing weights. The weights vary according to the different model assumptions and string comparators used. The number of pairs above different thresholds at different link error rates (0.002, 0.005, 0.01, and 0.02) are presented. False match error rates above 2 percent are not considered because the sets of pairs above the cutoff threshold contain virtually all of the true matches from the entire set of pairs when error rates rise to slightly less than 2 percent. In each line, the proportion of nonmatches (among the sum of all pairs in the Link and Clerical columns) is 2 percent.

Table 5. -- Matching Results at Different Error Rates: Third Pair of Files with 15,048 and 12,072 Records 116,305 Pairs Agreeing on Block and First Character of Last Name

Link Error Rate	Link Match/Nonm	Clerical Match/Nonm
0.002		
base	9696/19	155/182
s_c	9434/19	407/182
as	9436/19	406/182
os_l	9692/19	157/182
bigram	9515/19	335/182
0.005		
base	9792/49	59/152
s_c	9781/49	60/152
as	9783/49	57/152
os_l	9791/49	58/152
bigram	9784/49	66/152
0.010		
base	9833/99	18/102
s_c	9822/99	19/102
as	9823/99	17/102
os_l	9831/99	18/102
bigram	9823/99	27/102
0.020		
base	9851/201	0/ 0
s_c	9841/201	0/ 0
as	9842/201	0/ 0
os_l	9849/201	0/ 0
bigram	9850/201	0/ 0

The results generally show that the different string comparators improve matching efficacy. In all of the best situations, error levels are very low. The new string comparator produces worse results than the previous one (see e.g., Winkler, 1990) and the new assignment algorithm (when combined with the new string comparator) performs slightly worse (between 0.1 and 0.01 percent) than the existing string comparator and lsap algorithm. In all situations (new or old string comparator), the new assignment algorithm slightly improves matching efficacy.

To test the effect of the Winkler variant of the Jaro string comparator and bigrams on more recent files, we use 1995 test census files from Oakland, California. (See Table 6.) The match rates were as follows. In

the first matching pass, we only used pairs of records that agreed on housing unit ID. Those that were not matched were processed in a second pass. Blocking during the second pass was on house number and first character of the first name. The results generally show that either string comparator produces good results. The variant of the Jaro string comparator yields a slightly smaller clerical review region.

Table 6. -- First Pass -- Housing Unit Identifier Match: Matching Results of a Pair of Files with 226,713 and 153,644 Records, Respectively

	Jaro String Comparator		Bigram	
	Links	Clerical	Links	Clerical
	78814	5091	78652	5888
Estimated false match rate	0.1%	30%	0.1%	35%

Second Pass -- House Number and First Character of First Name: Matching Results of a Pair of Files with 132,100 and 64,121 Records, Respectively

	Links	Clerical
	16893	7207
Estimated false match rate	0.3%	40%

Summary and Conclusion

Application of new string comparator functions can improve matching efficacy in the files having large amounts of typographical error. Since many of the files typically have high typographical error rates, the string comparators can yield increased accuracy and reduced costs in matching of administrative lists and census.

References

Budzinsky, C. D. (1991). Automated Spelling Correction, Statistics Canada.

Dempster, A.P.; Laird, N.M.; and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, series B, 39, 1-38.

Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, 64, 1183-1210.

Frakes, W. B. and Baeza-Yates, R. (eds.) (1992). *Information Retrieval: Data Structures and Algorithms*, Upper Saddle River, NJ: Prentice-Hall PTR.

Hall, P.A.V. and Dowling, G.R. (1980). Approximate String Comparison, *Computing Surveys*, 12, 381- 402.

Jaro, M. A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of

Tampa, Florida, *Journal of the American Statistical Association*, 89, 414-420.

Lynch, M. P. and Winkler, W. E. (1994). *Improved String Comparator*, Technical Report, Statistical Research Division, Washington, DC: U.S. Bureau of the Census.

McLaughlin, G. (1993). Private Communication of C-String-Comparison Routine.

Newcombe, H. B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press.

Newcombe, H. B.; Kennedy, J. M.; Axford, S. J.; and James, A. P. (1959). Automatic Linkage of Vital Records, *Science*, 130, 954-959.

Pollock, J. and Zamora, A. (1984). Automatic Spelling Correction in Scientific and Scholarly Text, *Communications of the ACM*, 27, 358-368.

Winkler, W. E. (1985). Preprocessing of Lists and String Comparison, *Record Linkage Techniques --1985*, W. Alvey and B. Kilss, (eds.), U.S. Internal Revenue Service, Publication 1299, 181-187.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354-359.

Winkler, W. E. (1994). *Advanced Methods for Record Linkage*, Technical Report, Statistical Research Division, Washington, DC: U.S. Bureau of the Census.

Winkler, W. E. and Thibaudeau, Y. (1991*). An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Decennial Census*, Statistical Research Division Report 91/09, Washington, DC: U.S. Bureau of the Census.

*This paper reports general results of research by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the U.S. Bureau the Census.

Chapter 7

A Comparison of Direct Match and Probabilistic Linkage in the Death Clearance of the Canadian Cancer Registry

Tony LaBillois, Marek Wysocki, and Frank J. Grabowiecki,
Statistics Canada

Abstract

The Canadian Cancer Registry (CCR) is a longitudinal person-oriented database containing all the information on cancer patients and their tumours registered in Canada since 1992. The information at the national level is provided by the Provincial and Territorial Cancer Registries (PTCRs). An important aspect of the CCR is the Death Clearance Module (DCM). It is a system that is designed to use the death records from the Canadian Mortality Data Base to confirm the deaths of the CCR patients that occurred during a pre-specified period. After extensive pre-processing, the DCM uses a direct match approach to death confirm the CCR patients that had a death registration number on their record and it performs a probabilistic record linkage between the remaining CCR patients and death records. For one province, death registration numbers are not provided with the cancer patient records. All these records go directly to the probabilistic linkage. For the rest of the country, a good proportion of the cancer patients reported as dead by the PTCRs have such a number that can be used to match directly the two databases. After an overview of the CCR and its DCM, this presentation will compare the situation where the direct match is used in conjunction with the probabilistic linkage to death confirm cancer patients versus the case where the probabilistic record linkage is used alone.

Introduction

In combining two sources of data, it is sometimes possible to match directly the records that represent the same units if these two sources have one common unique identifier. Nevertheless, it is often not possible to find all the common units using only this approach, either because the two sources do not have a common unique identifier, or because, even when used, it is not complete for all the records on the files.

The case of the Death Clearance (DC) of the Canadian Cancer Registry (CCR) is an example of the latter. The purpose of this task is to associate cancer patient records with death certificate records to identify the individuals that are present on both files. The CCR already contains the death registration identifier for some patients, but not for all that may indeed be deceased. Consequently, the most reasonable process involves matching directly all the CCR patient records that have this information, and then using probabilistic record linkage in an attempt to couple the remaining records that could not directly match. It is our belief that this maximises the rate of association between the two files while reducing the processing cost and time. In this situation, one could also use probabilistic linkage, alone, to perform the same task. The intention of this study is to compare these two approaches.

Firstly, this paper provides an overview of the CCR with emphasis on the Death Clearance module.

Secondly, the characteristics of the populations used in the study are described. Next, the paper explains the comparisons between the two approaches (process, results and interpretations); and finally, it presents the conclusions of this study.

Overview of the Canadian Cancer Registry

The Canadian Cancer Registry at Statistics Canada is a dynamic database of all Canadian residents diagnosed with cancer [1] from 1992 onwards. It replaced the National Cancer Incidence Reporting System (NCIRS) as Statistics Canada's vehicle for collecting information about cancer across the country. Data are fed into the CCR by the 11 Provincial and Territorial Cancer Registries (PTCRs) that are principally responsible for the degree of coverage and the quality of the data. Unlike the NCIRS that targeted and described the number of cancers diagnosed annually, the CCR is a patient-based system that records the kind and number of primary cancers diagnosed for each person over a number of years until death. Consequently, in addition to cancer incidence, information is now available about the characteristics of patients with multiple tumours, as well as about the nature and frequency of these tumours. Very importantly, since patients' records remain active on the CCR until confirmation of their death, survival rates for various forms of cancer can now be calculated.

The CCR comprises three modules: *core, internal linkage* and *death clearance*. The *core* module builds and maintains the registry. It accepts and validates PTCR data submissions, and subsequently posts, updates or deletes information on the CCR data base. The *internal linkage* module assures that the CCR is truly a person-based file, with only one patient record for each patient diagnosed with cancer from 1992 onwards. As a consequence, it also guarantees that there is only one tumour record for each, unique, primary tumour. The *internal linkage* identifies and eliminates any duplicate patient records that may have been loaded onto the database as a result of name changes, subsequent diagnoses, or relocations to other communities or provinces/territories. Finally, *death clearance* essentially completes the information on cancer patients by furnishing the official date and cause of their death. It involves direct matching and probabilistic linking cancer patient records to death registrations at the national level.

The Death Clearance Module

Death clearance is conducted on the CCR in order to meet a certain number of objectives (Grabowiecki, 1997). Among them, it will :

- permit the calculation of survival rates for patients diagnosed with cancer;
- facilitate epidemiological studies using cause-of-death; and
- help file management of the CCR and PTCRs.

The death clearance module confirms the death of patients registered on the CCR by matching/linking [2] their patient records to death registrations on the Canadian Mortality Data Base (CMDB), or to official sources of mortality information other than the CMDB. These other sources include foreign death certificates and other legal documents attesting to, or declaring death (*they are added to the CMDB file before processing*).

The first major input to this module is the CCR database that is built of patient and tumour records. For every person described on the CCR, there is only one patient record, but as many tumour records as there are distinct, primary cancers diagnosed for that person. Patient records contain nominal, demographic and mortality information about the person, while tumour records principally describe the characteristics of the cancer and its diagnosis. CCR death clearance uses data from the patient record augmented with some

fields from the tumour record (the tumour record describing the patient's *most recently diagnosed tumour* when there is more than one). More details on the variables involved are available in Grabowiecki (1997) and Statistics Canada (1994).

The second main input is the Canadian Mortality Data Base. This file is created by Statistics Canada's Health Division from the annual National Vital Statistics File of Death Registrations, also produced by Statistics Canada. Rather than going directly to the Vital Statistics File, death clearance uses the CMDB as the principal information source about all deaths in Canada, because of improvements that make it a better tool for record linkage. A separate record exists on the CMDB for every unique reported surname on each Vital Statistics record -- viz.: the deceased's surname, birth/maiden name, and each component of a hyphenated surname (e.g., Gérin-Lajoie, Gérin, and Lajoie). All of the above surnames and the Surname of the Father of the Deceased have been transformed into NYSIIS [3] codes. For details on the CMDB data fields needed for death clearing the CCR, consult Grabowiecki (1997) and Statistics Canada (1997).

Death clearance can be performed at any time on the CCR. However, the most efficient and effective moment for performing death clearance is just after the completion of the Internal Record Linkage module, that identifies and removes any duplicate patient records on the CCR data base.

The death clearance process has been divided into five steps.

- **Pre-Processing**

  In this phase the input data files for death clearance are verified and prepared for the subsequent processing steps. The specific years of CMDB data available to this death clearance cycle are entered into the system. Based upon these years, the cancer patient population from the CCR, and mortality records from the CMDB are selected.

- **Direct Match (DM)**

  The unique *key* to all the death registrations on the CMDB is a combination of three data fields:

  - Year of Death
  - Province (/Territory/Country) of Death
  - Death Registration Number.

  These three fields are also found on the CCR patient record. PTCRs can obtain this information by doing their own death clearance, using local provincial/territorial files of death registrations. Patient records having responses for all three *key* fields first pass through a direct match with the CMDB in an attempt to find mortality records with identical common identifiers. If none is found, they next pass through the probabilistic *record linkage* phase, along with those patient records missing one or more of the *key* match fields. For the records that do match, five data items common to both the patient and CMDB records are compared (Sex, Day of Death, Month of Death, Year of Birth, Month of Birth). On both the CCR patient records and matched CMDB records, the responses must be non-missing and identical. If they are not, both the patient and mortality records are free to participate in the record linkage, where they may link together. Matched pairs that pass the comparison successfully are considered to represent the same person; they then will move on to the *post-processing* phase.

- **Probabilistic Linkage (PL)**

  In order to maximise the possibility of successfully linking to the CMDB file, the file of unmatched CCR patient records is *exploded* by creating, for every person, a separate patient record for each

unique Surname, each part of a hyphenated Surname, and the Birth/Maiden Name -- a process similar to the one used to create the CMDB, described in above. NYSIIS codes are generated for all names.

The two files are then passed through the Generalised Record Linkage System (GRLS), and over 20 important fields are compared using a set of 22 rules. Based on the degree of similarity found in the comparisons, weights are assigned, and the CCR-CMDB record pairs with weights above the pre-established threshold are considered to be linked. When patient records link to more than one mortality record, the pair with the highest weight is taken and the other(s) rejected. Similarly, if two or more patients link to the same CMDB record, the pair with the highest weight is selected.

The threshold weight has been set at such a level that the probability of the linked pairs describing the same person is reasonably high; consequently, manual review is not necessary in the linkage phase. At the same time, the threshold has not been positioned too high, in order to avoid discarding too many valid links, and thus reducing the effectiveness of the record linkage process.

The death information of linked CMDB records is posted onto the CCR patient records, overlaying any previously reported data in these fields. The linked pairs and unlinked CCR patient records join the matched pairs in proceeding to the *post processing* phase of death clearance.

- **Post-Processing**

  Essentially, this phase updates the CCR data base with the results from the *match* and *linkage* phases. Also, the results are communicated to the PTCRs for their review, and for input into their own data bases. Before being updated, copies are made of the patient records from the database. This makes it possible to restore them to their pre-death confirmed state should the matches/linkages be judged to be incorrect later by the PTCRs.

- **Refusal Processing**

  Refusals are PTCR decisions, taken after their review of the feedback reports and files generated in the *post processing* phase, that specific matches and linkages are incorrect -- i.e., that the persons described on the CCR patient records are not the same persons to whose death registrations they matched or linked. In this step, the affected patient records have their confirmation of death reversed, and are restored to their pre-death clearance state.

A description of the entire DC Module is available in Grabowiecki (1997) and the detailed specifications of the Direct Match and Probabilistic Linkage can be found in Wysocki and LaBillois (1997).

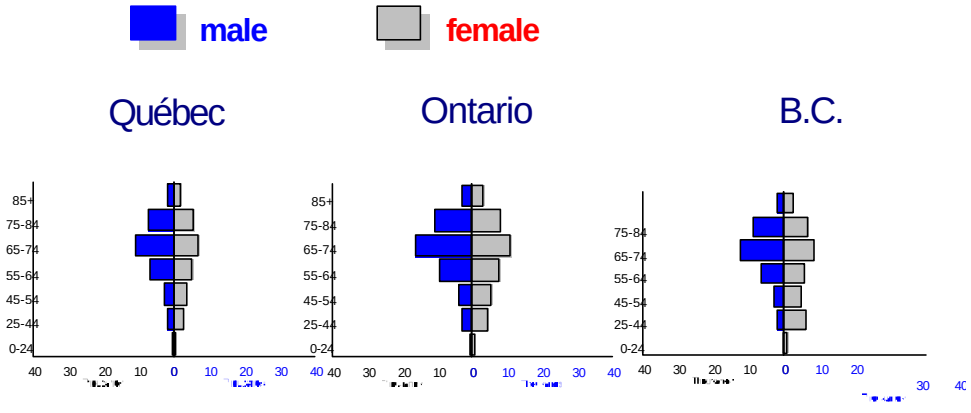
Characteristics of the Target Populations for this Study

To perform our comparisons, a subset of the CCR population was selected that could best illustrate the effect of direct match versus probabilistic linkage. Three provinces were chosen: British Columbia, Ontario and Québec. They were picked because they contain, within Canada, the largest populations of cancer patients, and the size of their respective populations is in the same order of magnitude. Québec was specifically taken because its provincial cancer registry does not do death clearance. Consequently, the patient files sent to the CCR by this registry never contain complete death information. Therefore, no cancer patient record from Quebec can obtain a confirmation of death by means of the Direct Match process; all Québec records participate in the Probabilistic Linkage. All other provinces do their own DC, and a significant number of their records on the CCR stand a good chance of being confirmed as dead as a result of the

Direct Match.

Due to the availability of data from the CCR and the CMDB at this time, we used reference years of diagnosis 1992 and 1993. The distribution by age and sex of the cancer patients in the three provinces is shown in Figure 1, below. It appears that there are only minor differences in the populations of cancer patients between these three provinces. Consequently, such differences are not expected to cause differences in the results of the death clearances.

Figure 1. -- Demographic Characteristics of the Populations of Cancer Patients



It is also important to note that the data coming from different provinces are gathered by different PTCRs. Even though there is little difference between them, in terms of coding practices, definitions and timeliness, certain variations still exist. In particular, the data sources used by the PTCRs to build their registries vary considerably among them (Gaudette et al., 1997). These considerations are taken into account in the interpretation of the results.

Direct Match and Probabilistic Linkage Vs. Only Probabilistic Linkage (Within the Same Province)

Process

This comparison is done by running the complete DC Module on the CCR data from British Columbia and Ontario. Both the DM and the PL are used to identify pairs for death confirmation. In the second run, any death information contained on the CCR records from these provinces is ignored. The system thus channels all the records directly to the PL. Québec data are not usable for this comparison because of the absence of complete death information on their CCR records. By comparing the two sets of pairs obtained in each approach for death confirmation, it is possible to measure different phenomena:

- overall percentage of accepted pairs (death confirmations) for each approach;
- percentage of pairs that are common to both approaches;

- percentage of pairs that were present in the regular DC process (DM & PL) but not in the PL only;
- percentage of pairs that were not present in the regular DC process (DM & PL) but were found in the PL only; and
- computer time and cost for each approach.

These measures help to evaluate the usefulness of the Direct Match in the DC process and contrarily, the impact of not having the CCR death information previously supplied by PTCRs.

Results and Observations

The results of this process are summarised in Figure 2, below. When both a DM and PL were performed, the majority of the pairs formed (approximately 95%) came from the DM. This was the case for both of the provinces involved in this part of the study. This result emphasises the importance of high quality death information in effectively matching records on these two files. There can be no direct match unless all of the death fields are identical on the two files, and these account for all but 5% of the total of pairs created in the DM and PL process.

Figure 2. -- Comparison of Ontario and British Columbia Using Both Methods

DC Population		DM and PL				PL	
		Matched	Linked	Total	%	Total	%
Ont.	84,926	22,648	1,183	23,831	28.1	23,670	27.9
B.C.	33,103	8,058	360	8,418	25.4	8,367	25.3
Total	118,029	30,706	1,543	32,249	27.3	32,037	27.1

It is evident that in terms of the number of pairs obtained in the end, one can expect little difference between the two methods of death clearance. Additionally, the particular pairs obtained (which specific patients are confirmed) will also be very similar. In this regard, there was less than a 1% difference in the two methods. Those differences that did exist tended to reflect favourably on the DM-PL method. Both methods found the same 32,035 pairs. On a net basis, the DM-PL method found 214 more pairs than did the PL only method. In percentage terms, this represented a negligible amount (again, less than 1%). Of those 214 pairs, roughly 94% were found in the direct match portion of the run; the others were found in the linkage. There were two pairs identified by the linkage-only method and not by its counterpart.

In regard to the actual cost of running the programs under the two different methods, the total for the DM-PL approach was 54% of the total cost incurred in running the PL alone. There is a certain small amount of instability in these numbers since the cost was dependent in part on the level of activity on the mainframe computer at the time that the programs were run. However, the percentage difference in the two costs is substantial even when this is considered. The relatively high cost of the linkage-only approach is due to the fact that the usual preprocessing steps must still be done but, at the same time, the number of records that are compared in the probabilistic linkage is considerably higher than the number used in the DM-PL approach (since many patient records, and their associated death records, will have been accounted for in the DM).

A Province With Only Probabilistic Linkage Vs. Provinces With Direct Match and Probabilistic Linkage

Process

For this part, the complete death clearance system is used to process the data of the three selected provinces. It will automatically produce death confirmation pairs by using the Direct Match and the Probabilistic Linkage for British Columbia and Ontario. Simultaneously, it will only apply the Probabilistic Linkage for Québec, because the Québec cancer registry does not report the necessary identifiers for the Direct Match to the CCR. In comparing the death confirmation results obtained for each of the three provinces, it is possible to observe different phenomena. The first is the overall percentage of accepted pairs (death confirmations) for each province, and the possible contrast between Québec and the two others. Another aspect to consider is the comparison of the percentage of death confirmation in Québec versus those obtained with PL only for British Columbia and Ontario in the previous Section. It is also interesting to evaluate the impact of not having the CCR death information previously supplied by PTCRs.

Results and Observations

The results obtained from the above process are summarised in Figure 3.

Figure 3. -- Ontario and British Columbia vs. Quebec, Where Only PL Was Possible

DC Population		DM and PL				PL	
		Matched	Linked	Total	%	Total	%
Qué.	57,252	--	--	--	--	18618	32.5
Ont.	84,926	22,648	1,183	23,831	28.1	--	--
B.C.	33,103	8,058	360	8,418	25.4	--	--

The percentage of pairs found from among the Québec data is rather higher than the corresponding percentages for the other provinces. In addition, all the Québec patient records which contained some death information were successfully linked to a mortality record during probabilistic linkage. This was not the case for all of the Ontario and BC records which contained death information; that is, there were some patients reported as deceased by Ontario and BC which neither matched or linked to a CMDB record. Overall, 32.5% of the Québec records that were in scope were successfully linked to the death file, while 28.1% of the Ontario records and 25.4% of the BC records were matched or linked. As previously noted, the data from Québec does not contain complete death information; it does, however, contain some records where the patient was reported as deceased by this province. It is probable that these were hospital deaths and so it is in turn very unlikely that the corresponding patients are being mistakenly reported as deceased. In essence, these patients can be anticipated to be good candidates to be successfully linked to a death record.

More generally, some cancer patients in Québec receive treatment entirely outside of hospitals and such patients may not then be reported to the CCR. The data from Québec might, therefore, contain a greater proportion of more serious cancers than do the data from the other provinces used in the study. This offers a possible explanation for the higher percentages of cancer patients confirmed in Québec compared to Ontario and B.C.

Finally, we have seen that the differences between the outcomes observed for the Ontario-BC data, using the match and linkage, and the linkage only, in terms of the total number of pairs found, were relatively minor. Again, a greater percentage of pairs were found in Québec than in the other provinces, and possibly because of the reasons outlined above.

Conclusions

Death Clearance of the CCR using PL only can be conducted with equal effectiveness as the DM-PL approach because of the reporting of high-quality personal and cancer data by the PTCRs. The advantages of the DM-PL method include lower operating costs to perform death clearance (increased efficiency), and greater certainty with the results (minimum manual review of cancer-mortality record pairs by PTCRs).

Footnotes

[1] The cancers that are reported to the CCR include all primary, non-benign tumours (with the exception of squamous and basal cell skin cancers, having morphology codes 805 to 808 or 809 to 811, respectively), as well as primary, benign tumours of the brain and central nervous system. In the International Classification of Diseases System – 9th Revision (ICD-9), the following codes are included: for benign tumours, 225.0 to 225.9; for *in situ* / intraepithelial / noninfiltrating / noninvasive carcinomas, 230.0 to 234.9; for uncertain and borderline malignancies, 235.0 to 239.9; and finally, for primary site malignancies, 140.0 to 195.8, 199.0,199.1, and 200.0 to 208.9. Similarly, according to the International Classification of Diseases for Oncology – 2nd Edition (ICD-O-2), the target population of cancers includes: all *in situ*, uncertain / borderline, and primary site malignancies (*behaviour codes* 1, 2, or 3), as well as benign tumours (*behaviour code* 0) with topography codes in the range C70.0 to C72.9 (brain and central nervous system).

[2] Matching entails finding a unique, assigned, identification number on two or more records, thus identifying them as belonging to the same person; whereas linkage concludes that two or more records probably refer to the same person because of the number of similar, personal characteristics found on them.

[3] NYSIIS (New York State Identification and Intelligence System) assigns the same codes to names that are phonetically similar. It is used to group like-sounding names and thus take into account, during record linkage, variations (and errors) in spelling -- e.g., Burke and Bourque, Jensen and Jonson, Smith and Smythe.

References

Gaudette, L.; LaBillois, T.; Gao, R.-N.; and Whittaker, H. (1997). Quality Assurance of the Canadian Cancer Registry, *Symposium 96, Nonsampling Errors*, Proceedings, Ottawa, Statistics Canada.

Grabowiecki, F. (1997). *Canadian Cancer Registry, Death Clearance Module Overview,* Statistics Canada (internal document).

Statistics Canada (1994). *Canadian Cancer Registry Data Dictionary*, Health Statistics Division.

Statistics Canada (1997). *Canadian Mortality Data Base Data Dictionary*, Health Statistics Division, (preliminary version).

Wysocki, M. and LaBillois, T. (1997). Death Clearance Record Linkage Specifications, Household Survey Methods Division (internal document).

Note: For further information, contact: Tony LaBillois, Senior Methodologist, Household Survey Methods Division, Statistics Canada, 16-L, R.H. Coats Building, Ottawa, Ontario K1A 0T6, e-mail: labiton@statcan.ca; Marek Wysocki, Methodologist, Household Survey Methods Division, Statistics Canada, 16-L, R.H. Coats Building, Ottawa, Ontario K1A 0T6, e-mail: wysomar@statcan.ca; Frank Grabowiecki, Project Manager, Health Division, Statistics Canada, 18-H, R.H. Coats Building, Ottawa, Ontario K1A 0T6, e-mail: grabfra@statcan.ca .

Chapter 7

Improvements in Record Linkage Processes for the Bureau of Labor Statistics' Business Establishment List

Kenneth Robertson, Larry Huff, Gordon Mikkelson, Timothy Pivetz and Alice Winkler, Bureau of Labor Statistics

Abstract

The Bureau of Labor Statistics has historically maintained a Universe Database file that contains quarterly employment and wage information for all covered employees under the Unemployment Insurance Tax system. It is used as a sampling frame for establishment surveys, and also as a research database. Each quarter approximately seven million records are collected by the States and processed for inclusion on the file. There are many data items of interest associated with this database, such as an establishment's industry, county, employment information, and total wages. Historically, this database has contained five quarters of data. These data have been linked across the five quarters by both administrative codes and through a weighted match process. Recently, a project has been undertaken to expand this database so that it will include multiple years of data. Once several years of data have been linked, the database will expand as new data are obtained. This will create a new "longitudinal" establishment information database, which will be of prime interest to economic researchers of establishment creation, growth, decline, and destruction.

As one step in the creation of this new resource, research was initiated to refine the existing record linkage process. This paper will provide details of the processes used to link these data. First, we will briefly cover the processes in place on the current system. Then we will provide details of the refinements made to these processes to improve the administrative code match. These processes link nearly 95 percent of the file records. The remaining records are processed via a revised weighted match process. Information on the current state of the revised weighted match will be provided, as well as the details of work still in progress in this area.

Introduction

In preparation for the building of a new longitudinally-linked establishment database, the Bureau of Labor Statistics decided to review its current system for linking business establishments across time. Because the new database will be used to produce statistics on business births and deaths and job creation and destruction, we had to ensure that the linkage procedures used in building the database would yield the most accurate results possible. Since the current linkage system was built for different purposes than the new system, there were areas where we could potentially improve the process. This paper provides an explanation of the current linkage procedures, details of the work completed to date, and areas of research that need to be explored in the future.

Background

Quarterly Unemployment Insurance Address File

The Bureau of Labor Statistics oversees the Covered Employment and Wages, or ES-202 program, that provides a quarterly census of information on employers covered under the State Unemployment Insurance (UI) laws. These data are compiled into a data file, the Quarterly Unemployment Insurance (QUI) Address File.

The QUI file includes the following information for each active employer subject to UI coverage during the reported quarter: State UI Account Number, Establishment Reporting Unit Number (RUN), federal Employer Identification Number (EIN), four-digit Standard Industrial Classification (SIC) code, county/township codes, monthly employment during the quarter, total quarterly wages, and the establishment's name(s), address and telephone number. Known predecessor and successor relationships are also identified by UI Account Number and Establishment Reporting Unit Number (UI/RUN). These numbers are used as administrative codes for matching records from one quarter to the next. The State code, EIN, and UI/RUN allow establishments to be uniquely identified. Imputed employment and wage data are assigned specific codes to distinguish them from reported data. Codes are placed on the records to identify the type of address (i.e., physical location, mailing address, corporate headquarters, address on UI tax file, or "unknown").

The Universe Database

The State QUI files are loaded to a database, the Universe Database (UDB), for access by users for survey sampling and research purposes. The UDB is composed primarily of data elements drawn from the QUI files. In addition, there are a few system-assigned and derived data elements, as well as information on SIC code changes merged from other sources. An important system-assigned field is the UDB Number, a unique number identifying continuous business establishments.

UDB Record Linkage

When considering the linkage of these records, the reader should understand that we are linking files which have the same structure across time. These files are linked to a new iteration of themselves each quarter. This linkage allows us to identify business establishments which may have gone out of business; establishments which remain in business for both periods; and, new establishments. The quality of the administrative codes are very good, so we expect that we correctly link most records which should be linked. We follow the administrative code match with a probability-based match. This procedure is followed to identify the small percentage of links which are missing the appropriate administrative codes.

Each quarter prior to loading the QUI files to the UDB, a matching procedure is performed to link businesses. By default, all units that do not link are identified as either new establishments or closed establishments. In order to have accurate data on business births and deaths, it is critical that the matching system accurately link establishments. The intent of the original linkage system was to minimize the number of invalid matches. Unfortunately, however, this causes some good matches to be missed. Because statistics on business births and deaths were not being produced from these linked data and only a small percentage of the total number of records was affected, this situation was acceptable.

The match system was composed of four main components. The first component identified the most obvious continuous establishments -- those with the same State code-UI/RUN combination. These are

establishments that from one quarter to the next did not change their UI reporting -- no change of ownership, reorganization, etc. The second component matched units that States submitted with codes identifying predecessor/successor relationships. Given that State personnel have access to the information needed to determine these relationships, they are assumed to be correct.

The third component matched units based upon certain shared characteristics. Prespecified weights were assigned based on data element values that the units had in common. This weighted match routine processed the data in three steps (or blocks). All three blocks limited potential matches to those units coded in the same 4 digit SIC code and county. (The New England States also use township codes.) The first block included all units that also matched on a key constructed from the Trade Name field. This "Name Search Key" was composed of the first seven consonants of the Trade Name. The second block included all units that also matched on the first 15 positions of the Street Address field. The third block included all units that also contained identical phone numbers. Two matched units were considered a valid match when they exceeded a cutoff weight. A limitation stemming from this three-block structure is that units that had a valid relationship but had different 4 digit SIC or county codes are missed by the linkage system.

The fourth component of the matching routine attempted to capture changes that occurred within a quarter. It first linked units that had State-identified predecessor relationships already coded. It next performed a within-quarter weighted match to capture relationships not previously identified by the States. A significant restriction placed upon both parts of the within-quarter matching was that the potential predecessors had to contain zero employment in the third month of the quarter while the potential successor units had to contain zero employment in the first month of the quarter. Because of inconsistencies in reporting by some employers, valid relationships could exist that did not meet this criteria, and were not matched.

Reasons for Modifying the UDB Record Linkage Process

The UDB record linkage process effectively linked over 96 percent of all the records received each quarter. Nevertheless, because its methodology was designed to limit the number of false matches, the original linkage system may not have been the most effective at identifying all valid relationships that existed between the remaining four percent of establishments. The result was a potential under-counting of continuous businesses and over-counting of business births and deaths. It is for that reason that the research described in this paper was undertaken.

Furthermore, experience with the previous matching process had highlighted specific areas of the process that needed improvement or enhancement. Although these areas affect only the four percent of the records mentioned above, **the net effect on the number of births and deaths identified could be significant.**

New Approach

The matching process consists of the two major procedures described below -- an administrative code match and a probability-based weighted match.

Administrative Code Match

Imputed Records

The first step in our new linkage process is to identify the imputed records (i.e., non-reporting records that are assumed to remain in business), and flag the corresponding record in the preceding quarter of the match. We then temporarily remove the imputed records from the current quarter file. Rather than assume that these units are delinquent, we attempt to identify the units that actually may have been reported under new ownership. At the end of all of the other match processes, we identify the unmatched flagged records on the past quarter file. These records have their matching imputed record restored to the current quarter file, and the link is made between them.

Within-Quarter Matches

Establishments that experience a reporting change within a quarter are generally assigned either a predecessor code or successor code pointing to another record within the same quarter. We determined that these within-quarter links were legitimate, so we included a process to find them.

Remove Breakouts and Consolidation

After the within-quarter matches were identified, we examined situations where multi-establishment reporters changed the way they reported. States encourage these reporters to supply data for each worksite. When a reporter changes from reporting all worksites on one report to supplying multiple reports, there is a possibility of failing to capture this as a non-economic event. If we were just counting records, it would look like we have a lot more establishments in the current quarter than we did in the past quarter. The reverse situation is also possible.

We were interested in identifying these links in order to exclude them in the counts as business openings or closings. The limited number of situations found were sent to a data editing routine, where the employment values were checked for reasonableness. If the match failed the edits, it was not counted as a breakout or a consolidation, and not included in counts of establishments increasing or decreasing in employment. Those cases failing the edits were still linked. However, since there is some type of economic change occurring along with the reporting change, the units failing the edits are included in counts of establishments increasing or decreasing in employment.

All Other Administrative Code Matches

The files were then linked by UI/RUN. These administrative codes linked most of the records. Additional links were identified using Predecessor and Successor codes. In general these administrative code match processes link over 96 percent of the current quarter file, depending on economic conditions.

Probability-Based Match

The probability-based weighted match process involves only the unmatched records from the administrative code match process. In this process we generally expect to match less than one-half of one percent of the current quarter records. This can also be expressed as linking less than ten percent of the current quarter residuals. While this is not a large portion of the overall records, it is still an important part of the overall process. The more accurate we can make the overall linkage process, the more useful the database will be in identifying economic occurrences.

Theoretical Basis for Weighted Matching

The weighted match process is accomplished using the software packages AutoStan and Automatch, from Matchware Technologies Incorporated. The first is a software package used to standardize names and addresses for linking. The second package uses a record linkage methodology based on the work of Ivan P. Felligi and Alan B. Sunter. Automatch uses the frequency of occurrence of selected variable values to calculate the probability that a variable's values agree at random within a given block. The probability that the variable's values agree given that the record is a match can also be calculated by the software. These match and nonmatch probabilities form the basis of the weight assigned to the variable in the match process. The sum of these variable weights are assigned as the overall weight for a given record pair. The distribution of these summed weights, along with a manual review of selected cases, allows us to determine an appropriate region where we find mostly matches. The lower bound of this region is set as the match cutoff value. We expect that above this cutoff will be mostly good matches, and that below this cutoff will be mostly bad matches.

These theoretical constructs are the foundation of probability-based record linkage. However, the nature of the data, in combination with software, hardware, and resource limitations, sometimes requires that additional steps be taken to fine-tune this process. Fortunately, Automatch provides some capabilities in this direction. The weights assigned to a matched or nonmatched variable can be overwritten or augmented as needed. This allows the user to augment the weight of important variables, as well as to penalize certain combinations of variable values, so that a record pair will not match.

Weighted Matches

Blockings

While the UDB Record Linkage system only utilized three basic blocks (trade name, address, and phone number), the new system, using Automatch, provides the option to use as many blockings as needed to match records. Based on empirical studies using California data, we constructed 21 blocks for the new system. All blocks match on two to four data elements. Within these 21 blocks, there are three groups which block on certain data elements. The first group contains blocks that include either exact name or exact street address. The second group blocks on phone number, and the third group blocks on various other data elements, such as ZIP code and EIN.

Adjustments to Blockings

After the first few runs of Automatch, we adjusted the blockings and their probability weights to enhance their matching potential. One weight adjustment we made was to records with similar street addresses. If the street addresses contained different suite numbers, we reduced the weight. Similarly, we reduced the weight if primary names contained different unit numbers. If one data element was unknown or blank, we increased the weight because these data elements did not necessarily disagree. However, if both data elements were unknown or blank, we deducted weight because there was a greater possibility that they would disagree. Finally, we deducted weight if both records were part of a multi-establishment employer.

We also made adjustments based on the address types. Some accountants submit data for many companies. Therefore, more than one record could have the same accountant's address and telephone number. If two records contained the same physical location address, they were considered a good match and we gave them more weight. If one record contained a physical location address and the other record

contained an unknown or tax address, it is possible that it would be a good match, so we gave it slightly more weight.

Subjective Results and Cutoffs

Although these records contained some common data elements, frequently it was difficult for us to decide whether the records were good matches. We subjectively identified matches as being "good," "bad," or "questionable." We reviewed these data to determine the quality of each matched pair. Then we set the cutoff weights for each of the 21 blocks, approximately in the middle of the questionable records.

Results

California data files were linked forward from the first quarter of 1994 (1/94) through the first quarter of 1995 (1/95). We evaluated the matches resulting from the final two quarters (4/94 to 1/95). These results are shown in Tables 1 though 4. Additionally, two quarters of data were matched for three other States -- West Virginia, Georgia, and Florida -- using preliminary match parameters developed for California. The results were evaluated by the four analysts using the same rules used in evaluating the California results. Although the results are not tabulated, they are approximately equivalent to those obtained for California. This finding is significant since there are insufficient resources to manually review cases which fall close to the match cutoff parameter. It is, therefore, important that we find match cutoff parameters for each block which produce satisfactory results in all States.

Number of Units Matched

Table 1 provides a summary of the matches in California which were obtained from the current matching procedures and the new procedures as tested. Both procedures produce the same number of matches on administrative number identifiers (79.58% of the file matched on UI/RUN). The first improvement in the matching process appears among those "delinquent" reporters which are assumed to remain in business. In the new procedures being tested, these imputed records are not generated until after all other matching processes are completed. The rationale for this change is that these non-reporting records may represent administrative business changes such as a change in ownership and may be reporting with a new UI/RUN. These units were matched with new UI/RUNs in 342 cases in California (0.04 percent of the file). These cases represent Type II errors (erroneous matches) for the previous matching process. The remaining 154,143 delinquent reporters were later matched to a new imputed record, as in current procedures.

The second improvement in the matching process appears in the within-quarter administrative match. These within-quarter matches represent units which have undergone some administrative change such as a change in ownership in a quarter and appear twice in the quarter with different UI/RUNs. It has become apparent during study of these files over several years that these units do not always cease reporting in one month during the quarter and begin reporting as a new entity in the next month of the quarter. The previous match procedures restricted these within-quarter matches to those reporters which report in a very precise manner. The new procedures allow for some reporting discrepancies in the monthly employment in matching these cases. The new procedures obtained approximately 1,110 (0.12%) additional matches for these situations.

Table 1. -- Results: Match Comparison for CA 95 Qtr. 1

	Current	New

Match Type	Method		Method	
	Count	%	Count	%
UI/RUN to UI/RUN	739,442	79.58	739,442	79.58
Correct "Delinquent" Matches	154,143	16.59	154,143	16.59
Incorrect "Delinquent" Matches (Type II Errors)	342	0.04	0	0
Pred/Succ. Codes/Non-Economic Reporting Changes /Within-Quarter Administrative Matches	821	0.09	1,978	0.21
Weighted	686	0.07	1,513	0.16
Births	33,723	3.63	32,081	3.45
Total (Records = 929,157)	100.0		100.0	

Note that the new system identifies 1,642 more links than the old system.

Finally, the third improvement in the matching process is in the weighted matching for all units in both quarters which do not match during any of the administrative matching procedures. The new procedures make use of many additional block structures which make possible incremental increases in the number of matches without significantly increasing the number of Type II errors. This is accomplished by tailoring the match cutoff parameter for each block so that most of the good matches fall above the match cutoff parameter without including a large number of Type II errors. The good matches falling below the parameter in one block are captured as matches in other blocks without picking up significant numbers of additional errors. The number of weighted matches went from 686 to 1,513 for an increase of 827 (0.09%). The total number of additional matches from the new procedures over the current procedures is 1,642. This reduces the number of business births (and business deaths) by 1,642 per quarter.

Although the results of the new linkage procedures do not appear dramatically different from the results of the current linkage procedures, the marginal improvements are significant in terms of the uses of the linkages. As stated earlier, one of the principal uses of the linked data files is to estimate the number and characteristics of business births and deaths and to track business births over time to determine when they increase or decrease their employment and how long they continue in business. It is easy to see that even though a large portion of the units match through the administrative codes, it is the remainder of the units which are considered business births and business deaths. Marginal improvements in matching these other units can have a relatively large impact on the number of business births and deaths and the ability to track them over time.

Quality of Matched Units

Tables 2 and 3 compare the quality of the weighted matches resulting from each procedure. There are two conclusions of interest from these tables. First, there are many more good matches resulting from the

new procedures and fewer Type II matching errors. Also, there are approximately 150 to 300 good or questionable weighted matches obtained from the current matching procedures which are not being identified during the new weighted matching procedures. There are two possible explanations. The first is that we are missing these matches with the new procedures and we must find methods which will identify the good matches. The second is that although we are not identifying these matches during the weighted match, they may be identified in the enhanced administrative matching procedures which would preclude them from the weighted matching process. The truth may lie somewhere between these possibilities and will be one focus of our future research efforts.

Table 2. -- A Comparison of Weighted Match Counts and Quality

	Current Method		New Method	
Match Quality	Count	%	Count	%
Good	262	49.1	1,317	87.0
Questionable	198	37.1	173	11.4
Bad	74	13.9	23	1.5
Total	534		1,513	

Note that all weighted match results are based on a manual review of linked records, and are based on the subjective opinions of several reviewers.

Table 3 continues the comparison of the quality of matches obtained from the current and new match procedures. It is obvious from this table that, although, the new match procedures apparently miss some good and questionable matches at or near the cutoff parameters for a match, the new procedures identify many additional good matches which are missed by the current weighted match procedures. This is accomplished by the new procedures while picking up fewer questionable and bad matches than the current procedures.

Table 3. -- Weighted Matches

Match Quality	Current Method Only		New Method Only		Both Methods	
	Count	%	Count	%	Count	%
Good	156	38.4	1,211	87.4	76	91.6
Questionable	178	43.8	153	11.0	7	8.4
Bad	72	17.7	21	1.5	0	0
Total	406		1,385		83	

Finally, Table 4 provides an analysis of the overall quality of the weighted matches obtained from the new procedures. Those units above the match cutoff parameter are identified as matches while the units below the match cutoff parameter are not identified as matches. There are at least 23 Type II errors while there are at least 51 Type I errors. This rough balance in these error Types seems a reasonable one for the purposes for which we are matching the files. Since there are only 142 good or questionable matches which fall below the match cutoff parameter, it seems that a substantial portion of the weighted matches identified only by the current weighted match procedures are identified during the enhanced new administrative match procedures.

Table 4. -- New Weighted Match Distribution and Quality

Match Quality	Group			
	Above Cutoff	%	Below Cutoff	%
Good	1,317	12.2	51	0.5
Questionable	173	1.6	91	0.8
Bad	23	0.2	9,098	84.6
Total	10,753			

Nonmatches are only counted within the twenty-one designated blocks, and with a match weight greater than or equal to zero.

Future Areas of Research

The results shown in Tables 1 through 4 are based on the research completed to date. As we are now aware from this preliminary effort, the matching procedures used here can be improved and there are more areas of study which may yield further improvement. In addition, there is additional testing which will be necessary to complete an initial assessment of the quality of the matching process.

- Since the files which make up the UDB are the product of each of the State Employment Security Agencies, it is important that the new match procedures be tested on data files from each of the States. This is the only way to insure that anomalies in any of the State files will not adversely affect the match results. The short time available for completing each of the quarterly matches and the size of the files does not allow for a manual review of the quarterly results. This initial review of the matching process using the final parameter values will provide some measure of the quality of matches obtained. It may also be advantageous to tailor the match cutoff parameters independently for each State.

- It is apparent from our initial analysis that additional analysis of the results of the current and new match procedures is necessary to determine how many good matches are being missed by the new procedures and how many of these are being identified by the new enhanced administrative match procedures. Once it is determined how many of these matches are being missed by the new procedures and their characteristics, the new match procedures must be modified to identify these matches.

- Intra-quarter weighted matching procedures should be tested to determine if such a procedure should be added to the new match procedures and its impact on overall results.

- Once the new procedures are enacted, an ongoing review of selected States may be recommended to insure that the match results do not deteriorate over time.

Acknowledgments

The authors would like to acknowledge the contributions of Larry Lie and James Spletzer of the Bureau of Labor Statistics and Catherine Armington of Westat, Inc.

Disclaimer

Any opinions expressed in this paper are those of the authors and are not to be considered the policy of the Bureau of Labor Statistics.

Chapter 7

Technical Issues Related to the Probabilistic Linkage of Population-Based Crash and Injury Data

Sandra Johnson, National Highway Traffic Safety Administration

Abstract

NHTSA's Crash Outcome Data Evaluation System (CODES) project demonstrated the feasibility of using probabilistic linkage technology to link large volumes of frequently inaccurate state data for highway safety analyses. Hawaii, Maine, Missouri, New York, Pennsylvania, Utah, and Wisconsin were funded by NHTSA to generate population-based medical and financial outcome information from the scene to final disposition for persons involved in a motor vehicle crash. This presentation will focus on the technical issues related to the linkage of population-based person-specific state crash and injury data.

Data Sources and Access

Data for the CODES project included records for the same person and crash event located in multiple different files collected by different providers in different health care settings and insurance organizations at different points in time. Each data file had a different owner, was created for a specific use, and was not initially designed to be linked to other files. Crash data were more likely to be in the public domain. Injury data were protected to preserve patient confidentiality. Each data source added incremental information about the crash and the persons involved.

Six of the seven states linked person-specific crash data statewide to EMS and hospital data. The EMS data facilitated linkage of the crash to the hospital data because they included information about the scene (pick-up) location and the hospital destination. The seventh state was able to link directly to the hospital data without the EMS data because date of birth and zip code of residence were collected on the crash record for all injured persons. Other data files, such as vehicle registration, driver licensing, census, roadway/ infrastructure, emergency department, nursing home, death certificate, trauma/spinal/head registries, insurance claims and provider specific data, were incorporated into the linkage when available and appropriate to meet the state's analytical needs.

Importance of Collaboration

Collaboration among the owners and users of the state data was necessary to facilitate access to the data. A CODES Advisory Committee was convened within each state to resolve issues related to data availability, patient confidentiality, and release of the linked data. The committee included the data owners such as the Departments of Public Safety, Health, Office of EMS, Vital Statistics, private and public insurers of health care and vehicles among others. Users included the owners, researchers, governmental entities, and others interested in injury control, improving medical care, reducing health care costs, and

improving highway safety.

File and Field Preparation

File preparation usually began with the creation of a person-specific crash file to match the person-specific injury data. Some of the data files only had one record per person; others, such as the EMS and hospital data files had more than one, reflecting the multiple agencies providing EMS care and the multiple hospital admissions for the same injury problem respectively. In some instances, all of the records were included in the linkage; at other times, the extra records were stored in a separate file for reference and analysis.

Except for Wisconsin which benefitted from state data which were extensively edited routinely, all of the states spent time, sometimes months, preparing their data for linkage. In most states, the hospital data required the least amount of editing. Preparation included converting the coding conventions for town/county codes, facility/provider, address, gender, and date in one file to match similar codes in the file being linked. Newborns were separated from unknown age. Date of birth and age discrepancies were resolved. Out of sequence times were corrected and minutes were added when only hour was documented. New variables were created to designate blocks of time, service areas for police, EMS and the hospital, probable admit date and others. Ancillary linkages to other data files were performed to beef up the discriminating power of the existing variables. Name and date of birth were the most common data added to the original files to improve the linkage.

Blocking and Linking Data Elements

Persons and events were identified using a combination of indirect identifiers and, in some linkages, unique personal identifiers, such as name, when they existed. Each of the CODES states used different data elements to block and link their files. Which variables were used for blocking and which for linkage depended upon both the reliability and availability of the data within the state, the linkage phase, and the files being linked. Most states used location, date, times, provider service area, and hospital destination to discriminate among the events. Age, date of birth, gender, and description of the injury were used most often to discriminate among persons. Hawaii, Missouri, New York, and Utah had access to name or initials for some of the linkages.

Linkage Results

Conditions of uncertainty govern the linkage of crash and injury state data. It is not certain which records should link. In the ideal world, records should exist for every crash and should designate an injury when one occurs; injury records should exist documenting the treatment for that injury and the crash as the cause; and the crash and injury records should be collected and computerized statewide. Linkage of a crash with an injury record should confirm and generate medical information about the injury. No linkage should confirm the absence of an injury. But that is the ideal world. In the real world, the crash record may not indicate an injury even though an injury occurred; the matching injury record may not indicate a crash or even be accessible; so it is difficult to know which records should link.

Linkage rates varied according to the type of data being linked. In each of the CODES states, about 10% of the person-specific police crash reports linked to an EMS record and slightly less than 1.8% linked to a hospital inpatient record, a reflection of the low rate of EMS transport and hospitalization for crash injuries. The linkage rates also varied by police designated severity level (KABCO). Linkage to the fatal injury records was not always 100%, but varied according to whether deaths at the scene were transported either by EMS or a non-medical provider. For the non-fatal injuries, linkage rates were higher for the more severe cases which by definition were likely to require treatment and thus to generate a medical record. About 76-87% of the drivers with incapacitating injuries linked to at least one injury or claims record (except

for Wisconsin, which had limited access to outpatient data and Pennsylvania which used 6 levels to designate severity). Linkage rates for persons with possible injuries varied widely among the seven states. Because of extensive insurance data resources, about two-thirds of the possible injuries linked in Hawaii and New York compared to a third or less in the other states. Many more records indicating "no injuries" matched in New York and Utah, again because of access to extensive computerized outpatient data for the minor injuries. Included in this group of not injured were people who appeared uninjured at the scene but who hours or days after the crash sought treatment for delayed symptoms, such as whiplash. Overall, the CODES states without access to the insurance data linked between 7-13% of the person-specific crash reports for crashes involving a car/light truck/van to at least one injury record compared to 35-55% for Hawaii and New York, the states with extensive outpatient data. Wisconsin linked 2% of its drivers to the hospital inpatient state data and this rate matched that for the seven states as a group.

Linkage of the records for the motorcycle riders was much higher than the car/light truck/van group, a reflection of the high injury rate for cyclists involved in police reported crashes. As expected the linkage rates were lower for the lower severities. Except for Pennsylvania and Wisconsin, more than 45 per cent of the person-specific motorcycle crash records linked to at least one injury record.

Validation of the Linkages

Causes of false negatives and false positives vary with each linkage because each injury data file is unique. Since it is unknown which records should link, validation of the linkage results is difficult. The absence of a record in the crash file prevents linkage to an injury record; the absence of a cause of injury code in the injury record risks a denominator inflated with non-motor vehicle crashes. The states assigned a high priority to preventing cases which should not match from matching and conservatively set the weight defining a match to a higher positive score. At the same time, they were careful not to set the weight defining a nonmatch too low so that fewer pairs would require manual review. The false positive rate ranged from 3.0 - 8.8 percent for the seven states and was viewed as not significant since the linked data included thousands of records estimated to represent at least half of all persons involved in motor vehicle crashes in the seven CODES states.

False positives were measured by identifying a random sample of crash and/or injury records and reviewing those that linked to verify that a motor vehicle crash was the cause of injury. Maine, Pennsylvania, and Wisconsin read the actual paper crash, EMS, and hospital records to validate the linkage. Missouri compared agreement on key linkage variables such as injury county, last initial, date of event, trafficway/trauma indicators, date of birth, or sex. Wisconsin determined that the false positive rate for the Medicaid linkage varied from that for hospitalizations generally since Medicaid cases were more likely to be found in urban areas.

False negatives were considered less serious than a false positive so the states adjusted the cut-off weight defining a nonmatch to give priority to minimizing the total matched pairs requiring manual review. A false negative represents an injury record with a motor vehicle crash designated as the cause which did not link to a crash report or a crash record with a designated severe injury (i.e., fatal, incapacitating) for which no match was found. The rates for false negatives varied from 4-30 percent depending on the linkage pass and the files being linked. The higher rates occurred when the power of the linkage variables to discriminate among the crashes and the persons involved was problematical. False negatives were measured by first identifying the records which should match. These included crash reports indicating ambulance transport, EMS records indicating motor vehicle crash as the cause of injury or hospital records listing an E code indicating a motor vehicle crash. These records were then compared to the linked records to identify those that did not link. False negatives were also identified by randomly selecting a group of crash reports and manually reviewing the paper records to identify those which did not link.

Crash and injury records failed to match when one or the other was never submitted, the linking criteria were too restrictive, key data linkage variables were in error or missing, the case selection criteria, such as the E-code, were in error or missing, the crash-related hospitalization occurred after several hours or days had passed, the crash or the treatment occurred out-of-state, etc. Lack of date of birth on the crash report for passengers was a major obstacle to linkage for all of the states except Wisconsin which included this information for all injured passengers. (As the result of the linkage process, Maine targeted the importance of including this data element on the crash report.) Among the total false negatives identified by Wisconsin, 12 percent occurred because the admission was not the initial admission for the crash and 10 percent occurred because key linkage variables were missing. Another 7.5 percent occurred because the linking criteria were too strict. About 7 percent were missing a crash report because the crash occurred out of state or the patient had been transferred from another institution. Twelve percent of the false negatives were admitted as inpatients initially for other reasons than the crash. It was not possible to determine the false negative rates when the key data linkage variables or E-code were in error, when out of state injuries were treated in Wisconsin Hospitals and when the crash record was not received at DOT.

In spite of the failure of some records to match, the estimates of matching among those that could be identified as "should match" was encouraging. Missouri estimated linkage rates of 65 percent of the hospital discharge, 75 percent of the EMS records, and 88 percent of the head and spinal cord injury registry records when motor vehicle crash as the cause of injury was designated on the record. Comparison of Missouri's linked and unlinked records suggested that actual linkage rates were even higher, as unlinked records contained records not likely to be motor vehicle related injuries (such as gunshot, laceration, punctures, and stabs). The linked records showed higher rates of fractures and soft tissue injuries, which are typical of motor vehicle crashes. Seventy-nine percent of the fractures were linked, as were 78 percent of soft tissue injuries.

The comparison of linked and unlinked records does not suggest that significant numbers of important types of records are not being linked, though perhaps some less severely injured patients may be missed. Because ambulance linkage was used as an important intermediate link for the hospital discharge file, some individuals not injured severely enough to require an ambulance may have been missed, but they would also be less likely to require hospitalization. Any effect of this would be to erroneously raise slightly the estimate of average charges for hospitalized patients.

Significance of the False Positive and False Negative Rates

Although the rates for the false negatives and false positives were not significant for the belt and helmet analyses, they may be significant for other analyses using different outcome measures and smaller population units. For example, analyses of rural/urban patterns may be sensitive to missing data from specific geographic areas. Analyses of EMS effectiveness may be sensitive to missing data from specific EMS ambulance services or age groups. Another concern focuses on the definition of an injury link. Defining an injury to include linkage to any claim record that indicated medical treatment or payment increases the probability of including uninjured persons who go to the doctor for physical exams to rule out an injury. But this group also includes persons who are saved from a more serious injury by using a safety device, so although they inflate the number of total injuries, they are important to highway safety. When minor injuries are defined as injuries only if their existence is verified by linkage, then by definition the unlinked cases become non-injuries relative to the data sources used in the linkage. States using data sources covering the physician's office through to tertiary care will have more linkages and thus more "injuries." Estimates of the percentage injured, transported, admitted as inpatients, and the total charges will vary accordingly.

The Linkage Methodology is Robust and the Linked

Data Are Useful

Seven states with different routinely collected data that varied in quality and completeness were able to generate from the linkage process comparable results that could be combined to calculate effectiveness rates. The states also demonstrated the usefulness of the linked data. They developed state-specific applications to identify populations at risk and factors that increased the risk of high severity and health care costs. They used the linked data to identify issues related to roadway safety and EMS, to support safety legislation, to evaluate the quality of their state data and for other state specific purposes.

Chapter 7

Record Linkage of Progress Towards Meeting the New Jersey High School Proficiency Testing Requirements

Eva Miller, Department of Education, New Jersey

Abstract

The New Jersey Department of Education has undertaken a records linkage procedure to follow the progress of New Jersey's Public school students in meeting the state standardized graduation test--the High School Proficiency Test (HSPT). The HSPT is a test of higher order thinking skills mandated by state legislation in 1988 as a graduation requirement which measures "those basic skills all students must possess to function politically, economically, and socially in a democratic society." The HSPT is first administered in the fall of the student's eleventh grade. If the student is not successful in any of the three test sections -- reading, mathematics, writing -- he/she has additional opportunities, each semester, to retake those test sections for which the requirement is still unmet. In terms of public accountability of educational achievement, it is very important to define a population clearly and then to assess the quality of public education in two ways -- the ability of the educational program to meet the challenge of the graduation test at the first opportunity (predominantly an evaluation of the curriculum); and the ability of the school system, essentially through the effectiveness of its interventions or remediations, to help the population meet the graduation requirement over the time remaining within a routine progression to graduation.

New Jersey uses a unique student identifier (not social security number) and has designed a complete mechanism for following the students through the use of test answer folders, computerized internal consistency checks, and queries to the school districts. The system has been carefully designed to protect confidentiality while tracking student progress in the many situations of moving from school to school or even in and out of the public school system, changes in grade levels and changes in educational programs (such as mainstreaming, special education, and limited English proficient programs).

Preserving confidentiality, linking completely to maintain the accuracy and completeness of the official records, definitions and analysis will be discussed.

Introduction

The New Jersey Department of Education has undertaken a record linkage procedure involving use of computers in the deterministic matching of student records to follow the progress of New Jersey's public school students in meeting the state standardized graduation test -- the High School Proficiency Test (HSPT). The HSPT is a test of higher order thinking skills mandated by state legislation in 1988 as a graduation requirement which measures "those basic skills all students must possess to function politically, economically, and socially in a democratic society." The HSPT is first administered in the fall of the students' eleventh grade. If the student is not successful in any of the three test sections -- reading, mathematics, writing -- he/she has additional opportunities, each semester, to retake test section(s) not yet

passed.

On first glance it would seem that New Jersey Department of Education's records linkage task is an easy and straightforward one. Since in October 1995, 62,336 eleventh grade students were enrolled in regular educational programs in New Jersey's public schools and 51,601 (or 82.8%) of these students met the HSPT testing requirement on their first testing opportunity (also includes eleventh grade students who may have met the requirement in one or more test sections while categorized by their local educators as "retained tenth grade" students), only 10,730 students need to be followed forward for three more semesters until graduation! Since some of these students (probably half again) will meet the requirement upon each testing opportunity, the number diminishes and the task should be trivial ... right? We have high speed computers and the public wanting this information thinks we just have to push a few buttons!

The problem is complicated, however, especially by flows of migration (students entering or leaving New Jersey's public schools) and mobility (students transferring from one public school to another), and gets increasingly more subject to error as time from the original eleventh grade enrollment passes. From the perspective of the policy maker in the Department of Education whose intent it is to produce a report of test performance rates which are comparable over schools, districts, and socio-demographic aggregations, the problem is further complicated by the fact that grade designation is a decision determined by local educators and rules may vary from school district to school district. Changes in a student's educational status with respect to Limited English Proficiency programs and/or Special Education programs also complicate tracking.

In terms of public accountability of educational achievement, it is very important to define a population clearly and then to assess the quality of public education in two ways:

- the ability of the educational program to meet the challenge of the graduation test at the first opportunity (predominantly an evaluation of the curriculum); and

- the ability of the school system, essentially through the effectiveness of its interventions or remediations, to help the population meet the graduation requirement over the time remaining within a routine progression to graduation.

Before the New Jersey Department of Education developed the cohort tracking system, information on HSPT test performance was reported specific to each test administration. This cross-sectional method of analysis was dependent on which students attended school during the test administration, and even more dependent on local determination of students' grade level attainments than in a longitudinal study. Using the cross-sectional reports, it was very difficult, if not impossible, to meaningfully interpret reports which were for predominantly retested student populations (i.e., what did the fall grade 12 test results report really mean?).

Methodology

The cohort tracking project is a joint effort involving the New Jersey Department of Education, National Computer Systems (NCS), and New Jersey educators in public high schools. The department is responsible for articulation of the purpose of the project and establishing procedures to be used -- including such activities as statistical design and decision-making rules, maintaining confidentiality of individual performance information, and assuring appropriate use and interpretation of reported information. NCS is responsible for development and support of a customized computer system, its specifications and documentation. The system is written in COBOL and provides features necessary for generation of the identifier; sorting and matching; data query regarding mismatches, nonmatches the uniqueness of the identifier, and assurances of the one-to-one correspondence of identifier to student. The department and

NCS share responsibility in maximizing the efficiency and effectiveness of the system and in trying to reduce the burden of paper work involved in record keeping. minimizing queries back to local educators, utilizing the computer effectively in checking information for internal consistency, developing and maintaining quality control procedures of interim reports to the local educators and public reports, and maximizing yield of accurate information. The local educator maintains primary responsibility related to the validity of the information by: assuring the accuracy of identifier information about individual students, reviewing reports sent to them to assure the accuracy and completeness of information about their enrolled and tested student population; and the responsibility to ascertain that every enrolled student is listed on the school's roster once and only once!

At its inception in October 1995, the cohort tracking project was intended to follow a defined population of eleventh grade students forward to their anticipated graduation (the static cohort). Local educators objected to this methodology because they could only educate students who were currently enrolled. To address this very important concern, the dynamic cohort was defined (see Figure 1). In effect, the dynamic cohort represents statistical adjustment of the original static cohort at each test administration to allow students who have left the reference group (school, district, or statewide) without meeting the graduation testing requirement to be removed from observation and adds those students who entered the reference group after fall of the eleventh grade and have not already met the testing requirement. Statistics produced for either the static cohort (prospective perspective) or the dynamic cohort (retrospective perspective) were not true rates, but rather were indices since after the first test administration (on the last day of testing in the fall of the eleventh grade) these populations are no longer groups of students served within a school, district, or state at a specific moment in time.

The mobility index -- simply the sum of the number of students entering and the number of students leaving the reference group since the last day of testing in fall of the eleventh grade, divided by the reference group at the initial time point -- was designed to help the user interested in evaluating educational progress as assessed by the HSPT (educator, parent, student, citizen and/or policy maker) decide which set of statistics, static or dynamic, would be more appropriate with respect to a particular reference group (school, district, or state). The higher the mobility level, the greater the difference between the set of statistics, and the more likely reliance should be made of the dynamic statistics.

In developing the system, the department had a need for a cost-effective, accurate, and timely system. The department needed exact matches and, therefore, could not rely on probability matching or phonetic schemes such as NYSIIS. A system with a number of opportunities for the local educator to review and correct the information was developed. A mismatch (or Type II error) was considered to have far more serious consequences in this tracking application than a nonmatch (or Type I error) because an educator might be notified that a student met requirements in one or more testing sections when that has not yet occurred (and the student might have been denied an opportunity to participate in a test administration based on a mismatch). The nonmatch is especially of concern to the local educator, because the most likely scenario here is that the student is listed in the file more than once, and none of these (usually incomplete) student records are likely to show all of the student's successes, therefore, the student was in the denominator population multiple times and had little or no chance of entering the numerator of successful students. In working with various lists and HSPT ID discrepancy reports, local educators have had heightened awarenesses of the "Quality in ... Quality out" rule mentioned by Martha Fair (Fair and Whitridge, 1997).

Figure 1. -- Definitions of Static and Dynamic Cohort

	Static Cohort	Dynamic Adjusted Cohort (after each admin)
	Enrollment Fall Junior / (*N*)	Adjusted after Each Administration
State Level	No Adjustments $\frac{\Sigma_{n_{p_i}}}{N}$	$\frac{\Sigma_{n_{p_i}}}{N+\Sigma_{y_{1_i}}-\Sigma_{y_{4_i}}}$
District Level:	No Adjustments $\frac{\Sigma_{n_{p_i}}}{N}$	$\frac{\Sigma_{n_{p_i}}}{N+\Sigma_{y_{2_i}}-\Sigma_{y_{6_i}}+\Sigma_{y_{1_i}}-\Sigma_{y_{4_i}}}$
School Level:	No Adjustments $\frac{\Sigma_{n_{p_i}}}{N}$	$\frac{\Sigma_{n_{p_i}}}{N+\Sigma_{y_{1_i}}-\Sigma_{y_{4_i}}+\Sigma_{y_{2_i}}-\Sigma_{y_{5_i}}+\Sigma_{y_{3_i}}-\Sigma_{y_{6_i}}}$

N = fall enrollment

y = in and out

n_{p_i} = # pass

pass

i = test administration, where
1 = 1st test administration for cohort
2 = 2nd test administration for cohort

Varies by administration (i)

y_{1_i} = in from out of state or private school

y_{2_i} = in from within state(public school), out of district

y_{3_i} = in from within state (public school), within district

y_{4_i} = migrate out of state public schools

y_{5_i} = migrate out of school, in district

y_{6_i} = migrate out of district, out of school, within state public schools

(out of state means out of N.J. Public schools)

Y's are only for those who have yet to pass test

Statistical notation developed by Gerald E. DeMauro, Director, Bureau of Assessment, NJDOE

Considerations Regarding the Identifier

This records linkage application is a relational database dependent on a unique identifier -- the HSPT identification number (HSPT ID) -- and supported by the following secondary fields: name (last and first but not middle initial), date of birth, and gender. In determining a unique identifier, the department first considered the use of social security number because it is a number which has meaning to the individual (is known), and is nearly universal and readily available to the individual. However, the department abandoned the plan to use SSN before the tracking project was implemented because citizens complained -- both verbally and in writing. Concerns included not wanting to draw attention to illegal aliens and considerations of the reasonableness of the number in terms of an individual's willingness to disclose it to school officials for this purpose and how use of the SSN may make it possible to access other unrelated files.

In reviewing Fair's criteria for a personal identifier (permanence, universality, reasonableness with respect to lack of objection to its disclosure, economy, simplicity, availability, having knowledge or meaning to the individual, accuracy, and uniqueness), the HSPT ID received low marks related to permanence and having the property of being easily known or meaningful to the student. The HSPT ID rated high marks on universality and reasonableness, with respect to lack of objection to its disclosure, precisely because it lacked meaning and could not be easily related to other records. The HSPT ID is also economical, simple, accurate, and is secured and safeguarded -- procedures have been implemented which assure that only appropriate school officials can access specific HSPT IDs for their enrolled populations and next access confidential information associated with these students' records of test results, in accordance with concerns regarding data confidentiality (U.S. Department of Education, 1994). Work involving assurance that there is only one number per student includes an HSPT ID update report, an HSPT ID discrepancy report, and multiple opportunities for record changes to correct information on student identifiers based on local educators' reviews of rosters (lists) of their students' test results.

The HSPT ID has been generated within the tracking system on the answer folder for each first-time test taker. Repeat test takers were to use stickers with student identifiers contained in a computer bar code label provided by NCS. District test coordinators can also contact staff at NCS, and after reasonable security checks are completed, obtain the HSPT ID and test results (from previous test administrations) for entering students who have already been tested.

A cohort year designation is to be assigned to a student once and only once. Safeguards are currently being developed to assure that despite grade changes over time, each student is followed based upon the initial (and only) cohort year designation.

Validity Assurances

The department and NCS are currently developing additional computerized procedures to assure the one-to-one correspondence of the HSPT ID to the student. A critical element in the assurance of the validity of the correct identification of each enrolled student as well as pass/fail indicators (for each test section and the total test requirement) is the review of the static roster immediately following the fall eleventh grade test administration.

Recently also, safeguards have been added to the computer system to assure:

- that for a given HSPT ID once a passing score in a particular test section has been obtained by a student, no further information on testing in that test section can be accepted by the cohort tracking system because the first passing score is the official passing score; and

- that for a student who has met the testing requirement by passing all test sections the HSPT ID

number is locked and the system accepts no new information to be associated with that HSPT ID.

Quality Control Procedures for Cohort Reports

Quality control of cohorts reports is a joint effort on the part of the department and NCS. Quality control procedures include visual review of student rosters and statistical report, utilization of a system of SAS programs generated under the same project definitions and decision-making rules by a different programmer in order to check the logic used in the COBOL programs. To date, this quality control has been conducted three times. There is a written quality control protocol which has made it possible to move from the implicit understanding of records linkage methodology and computer systems capabilities to explicit criteria for this particular application. These explicit criteria are identified, clearly articulated, and observable. This protocol has been very useful in that it :

- helped clarify expectations for NCS,
- allowed more department staff to participate fully in the quality control process while minimizing need for specific project orientation or training time, and
- more complete documentation of the quality control effort for each cohort after each test administration. Refinement of these quality control procedures is on-going.

Confidentiality

In addition to the procedures for release of HSPT ID described above, confidentiality is preserved on cohort dynamic out rosters in that students who have left a school are listed without pass/fail indicators for test sections and the total testing requirement.

With respect to public reporting, the department has been very conservative in using a rule of "10" instead of the rule of "three"; in this way individual student test performance information is not discernable from information reported publicly.

Results

Actual test performance results based on this longitudinal study have been reported only once to date (Klagholz, L. et al., 1996). These results were for the first cohort, juniors in October 1995, and followed students through one academic year (two test administrations). The audience of public users seemed to receive the information well and are currently anticipating the December 1997 release of information which is to include the academic progress of the 1995 cohort toward graduation and, in comparison to last year's public release, the academic progress of the second cohort, juniors in October 1996, through their junior year.

While there are a variety of ways to correct and update the cohort tracking master data base, a key (and predominant) method is the first record change process after receiving initial test results. The record change process is an opportunity for correction of erroneous data related to permanent student identifiers (name, date of birth, and gender), personal status identifiers (school enrollment, grade, participation in special programs (such as Special Education, Limited English Proficiency programs, and Title I), and test related information (attendance at time of testing each content area, void classifications, and first time or retest taking statuses).

There were a total of 2,526 record changes processed at this first opportunity for data review. This is not an unduplicated students count, since one student's records might have needed several variables to be

corrected. It is not readily obvious what denominator to suggest in determining rates -- 93,627 for total students enrolled (including students in special programs) would be most appropriate in determining a proportion of the total population to be tested for a cohort year. Another approach, however, would be that record changes as they relate to the cohort tracking project, should be segmented and the number of record changes for the students who had one or more test sections yet to pass after the October test administration would be useful; however, that statistic is not readily available.

Numbers of record changes by reason were as follows:

Student identifiers:	name:	447
	date of birth:	218
	gender:	41
	school:	53
Status:	grade:	367
	Special Education:	631
	Limited English Proficiency:	196
	Title I:	153
Test specific information:	attendance on test days:	43
	void classifications:	57
	first time or retest status	381 .

A systematic error regarding 731 students who were tested for the first time in April 1996 occurred. These students were not appropriately reflected in the dynamic cohort statistics. The computer system has been corrected to handle these cases correctly. This also necessitated tightening the quality control protocol and procedures. Corrected test performance rates for that same time point in the longitudinal study will be released in December 1997. While no one ever wants to release erroneous information, it was interesting to note the order of magnitude: for 51 schools there was no change, for 44 schools the correction increased pass rates by up to 0.8%, and for 136 schools passed rates decreased (by within 1.0% for 104 schools and between 1.1% and 4.2% for 32 schools).

The mobility index was designed to measure the stress on student populations caused by students who change the educational climate by either entering or leaving a particular high school after October of their junior year. This index was considered to be needed to guide the decision as to whether the set of static or dynamic statistics would be more appropriate measures of progress for a given reference group (school, district, state). The mobility index was observed to have a highly negative correlation with test performance. This finding was especially important to educators in communities with high mobility in that it helped these educators quantify the seriousness of the socio-economic problem, and communicate it in understandable terms regarding the consequences of these moves upon the continuity of students' educational experiences and educational progress in meeting performance standards.

Plans for the Future

The department has an ambitious plan to increase the graduation test (and assessments at fourth and eighth grades) to include eight test sections (content areas). The cohort tracking project is to be expanded to include all test sections on the graduation test. Cohort tracking is also to be extended vertically to the other grades for which there is a statewide assessment program.

The possibility of developing a population register for students enrolled in New Jersey's public schools is under discussion. Then the cohort tracking system would be incorporated into a larger department information system for reviewing educational programs, attendance, and school funding as well as outcome measures such as test results. In discussions about a population registry, social security number has been proposed for the linkage variable.

References

Fair, M. and Whitridge, P. (1997). Record Linkage Tutorial, *Records Linkage Techniques -- 1997*, Washington, D.C.: Office of Management and Budget.

Klagholz, L.; Reece, G.T.; and DeMauro, G.E. (1996). 1995 Cohort State Summary: Includes October 1995 and April 1996 Administrations Grade 11 High School Proficiency Test, New Jersey Department of Education.

U. S. Department of Education, (1994). *Education Data Confidentiality: Two Studies, Issues in Education Data Confidentiality and Access and Compilation of Statutes, Laws, and Regulations Related to the Confidentiality of Educational Data*, Washington, D.C.: National Center for Education Statistics, National Forum on Education Statistics.

Chapter 8

Public Attitudes Toward Data Sharing by Federal Agencies

Eleanor Singer and John VanHoewyk, University of Michigan
Stanley Presser, University of Maryland

Abstract

Very little information exists concerning public attitudes on the topic of data sharing among Federal agencies. The most extensive information prior to 1995 comes from questions on several IRS surveys of taxpayers, from questions added to a series of Wisconsin surveys carried out in 1993-95, and from scattered other surveys reviewed by Blair (1995) for the National Academy of Sciences panels. From this review it is clear that the public is not well informed about what data sharing actually entails, nor about the meaning of confidentiality. It seems likely that opinions on this topic are not firmly held and liable to change depending on other information stipulated in the survey questions as well as on other features of the current social climate.

In the spring of 1995, the Survey Research Center at the University of Maryland (JPSM) carried out a random digit dialing (RDD) national survey which was focused on the issue of data sharing. The Maryland survey asked questions designed to probe the public's understanding of the Census Bureau's pledge of confidentiality and their confidence in that pledge. Respondents were also asked how they felt about the Census Bureau's obtaining some information from other government agencies in order to improve the decennial count, reduce burden, and reduce cost. In addition, in an effort to understand responses to the data sharing questions, the survey asked about attitudes toward government and about privacy in general.

Then, in the fall of 1996, Westat, Inc. repeated the JPSM survey and, in addition, added a number of split-ballot experiments to permit better understanding of some of the responses to the earlier survey. This paper examines public attitudes toward the Census Bureau's use of other agencies' administrative records. It analyzes the relationship of demographic characteristics to these attitudes as well as the interrelationship of trust in government, attitudes toward data sharing, and general concerns about privacy. It also reports on trends in attitudes between 1995 and 1996 and on the results of the question-wording experiments imbedded in the 1996 survey. Implications are drawn for potential reactions to increased use of administrative records by the Census Bureau.

Introduction

For a variety of reasons, government agencies are attempting to satisfy some of their needs for information about individuals by linking administrative records which they and other agencies already possess. Some of the reasons for record linkage have to do with more efficient and more economical data collection, others with a desire to reduce the burden on respondents, and still others with a need to improve coverage of the population and the quality of the information obtained.

The technical problems involved in such record linkage are formidable, but they can be defined relatively precisely. More elusive are problems arising both from concerns individuals may have about the con-

fidentiality of their information and from their desire to control the use made of information about them. Thus, public acceptance of data sharing among Federal and state statistical agencies is presumably necessary for effective implementation of such a procedure, but only limited information exists concerning public attitudes on this topic.

A year and a half ago, the Joint Program in Survey Methodology (JPSM) at the University of Maryland devoted its practicum survey to examining these issues. The survey asked questions designed to probe the public's understanding of the Census Bureau's pledge of confidentiality and their confidence in that pledge. It also asked how respondents felt about the Census Bureau's obtaining some information from other government agencies in order to improve the decennial count or to reduce its cost. In addition, in an effort to understand responses to the data sharing questions, the survey asked a series of questions about attitudes toward government and about privacy in general.

Most of these questions were replicated in a survey carried out by Westat, Inc. in the fall of 1996, a little more than a year after the original survey. The Westat survey asked several other questions in addition -- questions designed to answer some puzzles in the original survey, and also to see whether the public was willing to put its money where its mouth was -- i.e., to provide social security numbers (SSN's) in order to facilitate data sharing. Today, I will do four things:

- Report on trends in the most significant attitudes probed by both surveys;
- Discuss answers to the question about providing social security numbers;
- Report on progress in solving the puzzles left by the JPSM survey; and
- Discuss the implications of the foregoing for public acceptance of data sharing by Federal agencies.

Description of the Two Surveys

The 1995 JPSM survey was administered between late February and early July to a two-stage Mitofsky-Waksberg random digit dial sample of households in the continental United States. In each household, one respondent over 18 years of age was selected at random using a Kish (1967) procedure. The response rate (interviews divided by the total sample less businesses, nonworking numbers, and numbers that were never answered after a minimum of twenty calls) was 65.0 percent. The nonresponse consisted of 23.4% refusals, 6.5% not-at-home, and 5.1% other (e.g., language other than English and illness). Computer-assisted telephone interviewing was conducted largely by University of Maryland Research Center interviewers, supplemented by graduate students in the JPSM practicum (who had participated in the design of the questionnaire through focus groups, cognitive interviews, and conventional pretests). The total number of completed interviews was 1,443.

The Westat survey (Kerwin and Edwards, 1996) was also conducted with a sample of individuals 18 or older in U.S. households from June 11 to mid-September. The response rate, estimated in the same way as the JPSM sample, was 60.4% [1]. The sample was selected using a list-assisted random digit dial method. One respondent 18 or over was selected at random to be interviewed.

Trends in Public Attitudes Toward Data Sharing

The most significant finding emerging from a comparison of the two surveys was the absence of change with respect to attitudes relating to data sharing. Indeed, if we are right that there has been little change on these matters, the new survey is testimony to the ability to measure attitudes reliably when question wording, context, and procedures are held reasonably constant -- even on issues on which the public is not well informed and on which attitudes have not crystallized. In 1996 between 69.3% and 76.1%, depending on the agency, approved of other agencies sharing information from administrative records with the Census Bureau in order to improve the accuracy of the count, compared with 70.2% to 76.1% in 1995 [2]. Responses to the Immigration and Naturalization Service, asked about in 1995, and the Food Stamp Office, asked about in 1996, are comparable to those to the Social Security Administration (SSA). Responses are consistently least favorable toward the Internal Revenue Service (IRS).

Westat documents five significant changes ($p < .10$) among 22 questions asked about the Census Bureau on both surveys. First, there is more awareness of the fact that census data are used to apportion Congress and as a basis for providing aid to communities; but second, there is less awareness that some people are sent the long census form instead of the short form. (Both of these changes make sense in retrospect. In the election year of 1996, apportionment was very much in the news; at the same time, an additional year had elapsed since census forms, long or short, had been sent to anyone.) Third, fewer people in 1996 than 1995 said that the five questions asked on the census short form are an invasion of privacy -- a finding at odds with others, reported below, which suggest increasing sensitivity to privacy issues between the two years. This issue will be examined again in the 1997 survey. Fourth, there was a modest increase in the strength with which people opposed data sharing by the IRS. This finding (not replicated with the item about data sharing by SSA) may have less to do with data sharing than with increased hostility toward the IRS. These changes are mostly on the order of a few percentage points. Finally, among the minority who thought other agencies could not get identifiable Census data there was a substantial decline in certainty, although the numbers of respondents being compared are very small.

Trends in Attitudes Toward Privacy

In contrast with attitudes toward data sharing and the Census Bureau, which showed virtually no change between 1995 and 1996, most questions about privacy and alienation from government showed significant change, all in the direction of more concern about privacy and more alienation from government. The relevant data are shown in Table 1.

There was a significant decrease in the percentage agreeing that "people's rights to privacy are well protected" and a insignificant increase in the percentage agreeing that "people have lost all control over how personal information about them is used." At the same time, there was a significant decline in the percentage *dis*agreeing with the statement, "People like me don't have any say about what the government does," and a significant increase in the percentage agreeing that "I don't think public officials care much what people like me think" and in the percentage responding "almost never" to the question, "How much do you trust the government in Washington to do what is right?" The significant decline in trust and attachment to government manifested by these questions is especially impressive given the absence of change in responses to the data sharing questions. We return to the implications of these findings in the concluding section of the paper.

Table 1. -- Concerns about Privacy and Alienation from Government, by Year

Attitude/Opinion	Agree Strongly or Somewhat	

	1995		1996	
People's rights to privacy are well protected	41.4	(1,413)	37.0	(1,198)
People have lost all control over how personal information about them is used	79.5	(1,398)	80.4	(1,193)
People like me don't have any say about what the government does	59.2	(1,413)	62.9	(1,200)
I don't think public officials care much what people like me think	65.4	(1,414)	71.1	(1,202)
How much do you trust the government in Washington to do what is right? (Almost never)	19.2	(1,430)	25.0	(1,204)

Willingness to Provide Social Security Number to Facilitate Data Sharing

One question of particular importance to the Census Bureau is the extent to which people would be willing to provide their social security number to the Census Bureau in order to permit more precise matching of administrative and census records. Evidence from earlier Census Bureau research is conflicting in this regard. On the one hand, respondents in four out of five focus groups were overwhelmingly opposed to this practice when they were asked about it in 1992 (Singer and Miller, 1992). On the other hand, respondents to a field experiment in 1992 were only 3.4 percentage points less likely to return a census form when it requested their SSN than when it did not; an additional 13.9 percent returned the form but did not provide a SSN (Singer, Bates, and Miller, 1992).

To clarify this issue further, the Bureau asked Westat to include a question about SSN on the 1996 survey. The question (Q21) read as follows:

> *"The Census Bureau is considering ways to combine information from Federal, state, and local agencies to reduce the costs of trying to count every person in this country. Access to social security numbers makes it easier to do this. If the census form asked for your social security number, would you be willing to provide it?"*

About two thirds (65.9%) of the sample said they would be willing to provide the number; 30.5% said they would not; and 3.5% said don't know or did not answer the question.

The question about SSN was asked *after* the series of questions asking whether or not people approved of other administrative agencies sharing data with the Census Bureau. Therefore, it is reasonable to assume that responses to this question were influenced by opinions about data sharing, which the preceding questions had either brought to mind or helped to create. And, not surprisingly, there is a relationship between a large number -- but not all -- of the preceding questions and the question about providing one's SSN.

For example, those who would provide their SSN to the Bureau are more likely to believe the census is extremely or very important and more likely to be aware of census uses. They are more likely to favor data sharing. Those who would not provide their SSN to the Bureau are more concerned about privacy issues.

They are less likely to trust the Bureau to keep census responses confidential; they are more likely to say they would be bothered "a lot" if another agency got their census responses; they are less likely to agree that their rights to privacy are well protected; less likely to believe that the benefits of data sharing outweigh the loss of privacy this would entail, and more likely to believe that asking the five demographic items is an invasion of privacy. All of these differences are statistically significant.

Table 2. -- Willingness to Provide SSN and Attitudes to Census Bureau

Attitude/Opinion	Would Not Provide SSN %	Would Provide SSN %
Believes counting population is "extremely" or "very" important	63.8	79.7
Is aware of census uses	43.1	54.8
Would favor SSA giving Census Bureau short-form information	56.3	85.0
Would favor IRS giving Census Bureau long-form information	30.4	61.2
Would favor "records-only" census	45.6	60.0
Trusts Bureau to not give out/keep confidential census responses	45.0	76.7
Would be bothered "a lot" if other agency got census responses	54.1	29.9
Believes benefits of record sharing outweigh privacy loss	36.0	51.1
Believes the five items on short form are invasion of privacy	31.3	13.4

There are also significant relationships between political efficacy, feelings that rights to privacy are well protected, feelings that people have lost control over personal information, and trust in "the government in Washington to do what is right" (Q24a-d) and willingness to provide one's SSN. These political attitude questions, it should be noted, were asked ***after*** the question about providing one's SSN, and so they could not have influenced the response to this question.

Of the demographic characteristics, only two -- gender and education -- are significantly (for gender, $p<.10$; for education, $p<.05$) related to willingness to provide one's SSN. Almost three quarters (71.4%) of men, but only 65.5% of women, are willing to provide their SSN. This is true of 71.2% of those with less than a high school education, 63.9% of those who are high school graduates, 68.7% of those with some college, and 76.8% of those who are college graduates. The same curvilinear relationship is apparent for income: 75.4% of those with family incomes of less than $20,000, 69.6% of those with incomes between $20,001 and $30,000 and $30,001 and $50,000, 68.6% of those between $50,001 and $75,000, and 75.4% of those with incomes over $75,000 say they would be willing to provide their SSN if asked by the Census Bureau to do so.

Table 3. -- Willingness to Provide SSN, by Concerns about Privacy and Alienation from Government

Concern/Alienation	Would Provide SSN %	Would Not Provide SSN %
Disagrees strongly that rights to privacy are well protected	24.2	45.6
Agrees strongly people have lost control over personal information	37.9	54.2
Agrees strongly "people like me" have no say about what government does	27.7	43.7
Agrees strongly public officials don't care much about "what people like me think"	31.2	45.4
Almost never trusts government in Washington to do what's right	19.5	37.8
Privacy loss outweighs economic benefit of data sharing	47.1	56.0
Economic benefit of data sharing outweighs privacy loss	47.9	30.4

From the foregoing, it appears that there are two reasons underlying reluctance to provide one's SSN. *First, there are reasons associated with beliefs about the census:* People who are less aware of the census, who consider it less important, and who are less favorable toward the idea of data sharing are significantly less willing to provide their SSN. Low levels of education are also associated with these characteristics. *Second, however, is a set of beliefs and attitudes concerning privacy, confidentiality, and trust:* People who are more concerned about privacy, who have less trust in the Bureau's maintenance of confidentiality, and who are less trusting of government in general are much less likely to say they would provide their SSN to the Census Bureau. Women are more likely to be concerned about privacy issues than men, and they are also less willing to say they would provide their SSN to the Bureau. In earlier analyses (Singer and Presser, 1996) we found that importance attached to the census, knowledge about the census, and attitudes about privacy were independent factors predicting willingness to have other agencies share data with the Bureau. Though we have not carried out a factor analysis of attitudes toward willingness to provide one's SSN, the relationships described above suggest that the same clusters of beliefs are relevant for this attitude, as well.

We should point out that the question asked on the 1996 survey, about whether or not respondents would be willing to provide their SSN, is not equivalent to a field experiment. The number of people who would provide their SSN if asked to do so in an actual census might very well be higher than the two thirds who said they would do so on this survey, as suggested by the field experiment cited at the beginning of this section. On the other hand, if the issue of privacy became salient prior to the census, the number complying might well be less. Arguing for the second, more cautious, inference is the fact that more than a third of those approached for the survey did not participate, and, since the introduction to the survey informed potential respondents about the topic, the nonparticipants may well have included those more suspicious of government and less inclined to cooperate with any request from government agencies, including the Census Bureau [3].

What Does Confidentiality Mean?

A number of question wording experiments were included in the 1996 Westat survey. The most important of these, from the perspective of understanding data sharing attitudes, had to do with the meaning of the Census Bureau's assurance of data confidentiality to respondents. The short answer to the question, “What does confidentiality mean to the public?” is, “We don't know.” However, in the rest of this paper, we try to summarize what we think we learned.

The 1995 JPSM survey resulted in one very puzzling finding. When asked whether other agencies could get their answers to census questions, *identified by name and address*, 41% said they did not know; of the rest, about 90% said other agencies could get such information (Presser and Singer, 1995). To make things even more puzzling, the better educated were more likely to believe, erroneously, that other agencies could get such data -- virtually the only time, so far as we know, that more education has been associated with more error (Hyman, Wright, and Reed, 1975). Furthermore, the belief that other agencies *could* get such data was associated with *more favorable* attitudes toward data sharing.

It thus seemed fairly clear that our attempt to provide a neutral definition of "confidentiality" in the 1995 instrument had not had the intended effect. Accordingly, we incorporated a four-way split ballot experiment into the 1996 survey.

One quarter of the sample were asked the 1995 question; one quarter, the 1995 question without the DK filter. One quarter were asked, "Do you think the Census Bureau does or does not protect the confidentiality of this (household demographic) information, or don't you know (DK)?" And the final quarter were asked the confidentiality question without the DK filter.

The results are shown in Table 4. The most striking thing about the table is simply the variation in responses, depending on the wording of the question. But the next most startling finding is the difference in responses to the questions asking whether other agencies can get identified data, and whether the Bureau keeps data confidential. Omitting those who answer DK, the percentages who believe responses are NOT shared (or data ARE kept confidential) ranges from 11.5% in Q 7-1 to 69.2% in Q 7-4. Omission of the DK filter reduces the size but does not change the basic form of the relationship. Majorities of the public believe that other agencies can get identified data; they also believe that the Bureau maintains data confidentiality.

Table 4. -- The Effects of Question Wording on Beliefs Regarding Sharing of Responses by Census Bureau

Response	Do you think other government agencies...can or cannot get people's names and addresses along with their answers to the census?		Do you think the Census Bureau does or does not protect the confidentiality of this [household demographic] information?	
	Explicit "Not Sure" %	No Explicit "Not Sure" %	Explicit "Not Sure" %	No Explicit "Not Sure" %
Believe that census responses are shared	47.1	76.9	9.6	20.9
Believe that census responses are not shared	6.1	15.4	12.9	47.0
Not Sure/ Don't Know	46.8	7.7	77.5	32.1
N (unweighted)	310	296	294	315

In passing, we should note that the distribution of answers to the version of the question which is identical to the 1995 question do not differ significantly from the 1995 distribution; and, as in 1995, people who said other agencies CAN get data were significantly more likely to favor data sharing in 1996 as well.

In another effort to understand the meaning of confidentiality to respondents, we asked another split-ballot question near the end of the 1996 survey. One asked whether the Census Bureau was required by law to keep census information confidential; the other, whether the Bureau was forbidden by law from giving identified census information to other agencies.

The responses to the two versions of this question are shown in Table 5. Majorities of those who have an opinion give the correct answer to both questions; but the proportion answering DK is larger, and the proportion giving the correct answer smaller, when the question asks about giving other agencies identified information than when it asks about maintaining confidentiality.

As a follow-up to both questions, we asked those who said the Bureau is required to protect the information or forbidden from disclosing it, *whether or not they trusted the Bureau to uphold the law -- that is, to keep the information confidential, or to refrain from disclosing it to other agencies.* Regardless of which version of Q22 they got, two thirds of those who answered Yes to the factual question about legal requirements said they trusted the Bureau to comply with the law. However, *those who not only say the Bureau is required to keep information confidential but who also trust the Bureau to do so, are significantly more likely to say both that other agencies cannot get the data and that the Bureau keeps data confidential.* Thus, not only knowledge of the law, but also trust in the Bureau's compliance with the law, is implicated in responses to the factual questions about whether the Bureau does or does not protect the data in its possession.

Table 5. -- The Effect of Question Wording on Knowledge of Laws Regarding Sharing of Census Information

Response	Is the Census Bureau forbidden by law from giving other government agencies census information identified by name or address?	Is the Census Bureau required by law to keep census information confidential?	Total
	%	%	%
Yes	28.3	51.1	40.2
No	17.1	11.6	14.2
Dont't Know	54.6	37.3	45.5
N (unweighted)	591	624	121

5

What differentiates those who trust the Bureau to keep information confidential from those who do not?

We found only two demographic characteristics that seemed to make a difference. Women are considerably more likely to say they trust the Bureau than men, and younger respondents are more likely to express trust than older respondents are. Whether this is an effect of age or of cohort is impossible to tell from this cross-sectional survey. None of the other demographic characteristics we examined -- education, race, or income -- make a consistent difference in attitudes of trust.

Finally, we looked at the relation of the beliefs about legal requirements to attitudes about data sharing. People who believe the Bureau is required by law to keep data confidential are significantly more likely to favor data sharing than those who do not. On the other hand, people who believe the Bureau is forbidden from sharing data *with* other agencies are significantly more likely to oppose data sharing *by* other agencies. Whether this results from confusion, or from an application of the norm of reciprocity, or from opposition to all data sharing, is impossible to tell.

Conclusions and Implications

The following conclusions seem to follow from comparison of the 1995 and 1996 surveys:

- Beliefs about the Census Bureau and attitudes toward data sharing have undergone little change since 1995.
- Beliefs about privacy and trust in government have deteriorated since 1995.
- To the public, the belief that the Bureau protects confidentiality does not seem to mean that other agencies cannot get data identified by name and address. What it does mean, we cannot tell from these data.
- In contrast to an implicit Census Bureau hypothesis, knowledge about legal requirements for confidentiality is not enough to convince the public that the Bureau actually protects confidentiality. In

order for knowledge to translate into belief, trust in the Bureau is required. The number of people who both know about legal requirements and trust the Bureau is only about two thirds as great as the number whose factual information is correct. (However, both knowledge and trust are independently related to attitudes toward data sharing.)

We believe these findings have two major implications for future data collection:

- First, we are planning in 1997 to ask both about whether other agencies can get data, and whether the Census Bureau maintains confidentiality, of one third of the sample. Then, everyone will be asked what confidentiality means to them. Only when the sources of misunderstanding are known can the Bureau better communicate its message about data protection to the public.

- Second, future surveys should be used to experiment with arguments that might be presented to the public in favor of data sharing. For example, there is evidence from the 1995 and 1996 surveys that the quality of the data is a more important consideration than cost. Are there orther arguments that are even more persuasive? How can the argument about quality be made even more compelling?

We hesitate to make substantive predictions about the public's acceptance of data sharing at the time of the next census. On the one hand, about two thirds of the public currently to favor this practice, this proportion has remained stable over at least a year, and two thirds say they would be willing to provide their SSN to the Bureau to facilitate such sharing. On the other hand, opposition to data sharing, and to making the SSN available, is strongly related to privacy concerns, and such concerns show a small but significant increase between 1995 and 1996. Thus, it seems possible that if privacy concerns continue to increase, they may erode the support for data sharing that currently exists. The same implication can be drawn from our findings concerning belief in the Census Bureau's assurance of confidentiality. Information about the law is apparently not enough; trust is also required. And the latter is a much more difficult message to communicate effectively.

Acknowledgments

We would like to express our thanks to Jeff Kerwin and Sherman Edwards at Westat for help with some aspects of data analysis and to Randall Neugebauer and his colleagues at the Census Bureau for helpful comments on an earlier draft.

Note

Follow-up to this research appears in Singer, Eleanor and Presser, Stanley (1997). Public Attitudes Toward Data Sharing by Federal Agencies: Trends and Implications, *Survey Research Methods Section Proceedings*, American Statistical Association (in process).

Footnotes

[1] The Westat report gives a response rate of 64.4%, which is based on excluding the number of respondents with language problems (n=126) from the denominator. This group is included in the JPSM count of eligibles. The introduction to the Westat survey differed somewhat from that used by JPSM. It read as follows:

"My name is _____. I'm calling from Westat on behalf of the U.S. Census Bureau in Washington, D.C. We're doing a study of people's opinions on whether government agen-

cies keep information about them private. You were randomly selected for this study from the adults in your household. This survey has been approved by the Office of Management and Budget, Number 0607-0822. Without this approval, we could not conduct this survey. Any questions or comments about the survey may be directed to the Census Bureau. If you would like, when we are done, I will provide you with the address."

The JPSM introduction omitted all references to OMB or the Census Bureau, as well as the sentence about random selection, and introduced the interviewer as calling from the University of Maryland. The sentence about the topic of the study was identical to that in the Westat introduction.

[2] Text and tables use data weighted for number of residential phone numbers in the household and number of persons in the household, poststratified to Census estimates of sex, race, age, education, and region.

[3] If the Bureau used a less specific introduction, the overall response rate to the survey might not change, but nonrespondents might be more representative (less biased) with respect to their attitudes toward government and the Census Bureau.

References

Blair, Johnny (1995). Ancillary Uses of Government Administrative Data on Individuals: Public Perceptions and Attitudes, a paper commissioned by the panel to Evaluate Alternative Census Methods, National Academy of Sciences, College Park, MD: Survey Research Center, University of Maryland.

Hyman, Herbert H.; Wright, Charles; and Reed, John (1975). *The Enduring Effects of Education,* Chicago: University of Chicago Press.

Kerwin, Jeffrey and Edwards, Sherman (1996). The 1996 Survey on Privacy and Administrative Records Use, Staff paper, Rockville, MD: Westat, Inc.

Kish, L. (1967). Survey Sampling, New York: Wiley.

Presser, Stanley and Singer, Eleanor (1995). Public Beliefs about Census Confidentiality, paper presented at the 1995 meetings of the American Sociological Association.

Singer, Eleanor and Miller, Esther R. (1992). Report on Focus Groups, Center for Survey Methods Research, U.S. Bureau of the Census.

Singer, Eleanor; Bates, Nancy; and Miller, Esther R. (1992). Memorandum for Susan Miskura, Bureau of the Census, July 15, 1992.

Singer, Eleanor and Presser, Stanley (1996). Public Attitudes toward Data Sharing by Federal Agencies, paper presented at the Annual Research Conference, Bureau of the Census.

Chapter 8

Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances

Arthur B. Kennickell, Federal Reserve Board

Abstract

Donald Rubin has suggested many times that one might multiply impute all the data in a survey as means of avoiding disclosure problems in public-use datasets. Disclosure protection in the Survey of Consumer Finances is a key issue driven by two forces. First, there are legal requirements stemming from the use of tax data in the sample design. Second, there is an ethical responsibility to protect the privacy of respondents, particularly those with small weights and highly salient characteristics. In the past, a large part of the disclosure review of the survey required tedious and detailed examination of the data. After this review, a limited number of sensitive data values were targeted for a type of constrained imputation, and other undisclosed techniques were applied. This paper looks at the results of an experimental multiple imputation of a large fraction of the SCF data using software specifically designed for the survey. In this exercise, a type of range constraint is used to limit the deviations of the imputations from the reported data. The paper will discuss the design of the imputations, and provide a preliminary review of the effects of imputation on subsequent analysis.

Introduction

Typically, in household surveys there is the possibility that information provided in confidence by respondents could be used to identify the respondent. This possibility imposes an ethical, and sometimes a legal, burden on those responsible for publishing the survey: It is necessary to review the data for items that could be highly revealing of the identity of individuals, and to filter the data made available to the public to minimize the degree of disclosure [1]. A recent issue of the *Journal of Official Statistics* (vol. 9, no. 2, 1993) deals with many aspects of this problem.

The Survey of Consumer Finances (SCF) presents two particularly serious disclosure risks. First, the survey is designed to measure the details of families' balance sheets and other aspects of their financial behavior. Second, the SCF oversamples wealthy families. Because of the sensitive nature of the data collected and because the sample contains a disproportionate number of people who might be at well-known, at least in their localities, disclosure review of the SCF is particularly stringent [2].

There is a growing belief that publicly available records, such as credit bureau files, real estate tax data, and similar files make it increasingly likely that an unscrupulous data user might eventually come closer to identifying an SCF respondent [3]. Several protective strategies have been proposed, but many proposals — truncation, simple averaging across cells, random reassignment of data, etc., — raise serious obstacles for many of the analyses for which the SCF is designed. The prospect of either being unable to release any information, or having to alter the data in ways that severely restrict their usefulness makes it imperative that we explore alternative approaches to disclosure limitation.

Most disclosure limitation techniques attempt to release some transformation of the data that preserves what is deemed to be the important information. Taking this idea to one farsighted conclusion, Donald

Rubin has suggested on several occasions creating an entirely synthetic dataset based on the real survey data and multiple imputation (see, e.g., Rubin, 1993) [4]. My impression is that most people have viewed the idea of completely simulated data with at least suspicion [5]. Such an exercise presents considerable technical difficulties. However, even if it is not possible to create an ideal simulated dataset, we may learn something from the attempt to create one. This paper describes several explorations in this direction.

Multiple imputation has played an important role in the creation of the public datasets for the SCF since 1989. In both the 1989 and 1992 surveys, a set of sensitive monetary variables was selected for a set of cases, the responses to those variables were treated as range responses (rather than exact dollar responses) and they were multiply-imputed using the standard FRITZ software developed for the SCF (see Kennickell, 1991). The approach has been broadened in the 1995 survey based on the work reported here. In the experiments discussed in this paper, several approaches are taken to imputing all of the monetary values in the 1995 SCF.

The first section of the paper provides some general information on the content of the SCF and the sample design and gives a review of the past approach to disclosure review. Because of the importance of imputation in the work reported here, the second section reviews the FRITZ imputation model. The third section discusses the special manipulation of the data for this experiment and presents some descriptive results. A final section summarizes the findings of the paper and points toward future work.

The 1995 Survey of Consumer Finances

The SCF is sponsored by the Board of Governors of the Federal Reserve System in cooperation with the Statistics of Income Division of the IRS (SOI). Data collection for the 1995 SCF was conducted between the months of July and December of 1995 by the National Opinion Research Center (NORC) at the University of Chicago. The interviews, which were performed largely in person using computer-assisted personal interviewing (CAPI), required an average of 90 minutes — though some took considerably longer.

Because the major focus of the survey is household finances, the SCF includes questions about all types of financial assets (checking accounts, stocks, mutual funds, cash value life insurance, and other such assets), tangible assets (principal residences, other real estate, businesses, vehicles, and other such assets) and debts (mortgages, credit card debt, debt to and from a personally-owned business, education loans, other consumer loans, and other liabilities). To meet the analytical objectives of the survey, detailed information is collected on every item. For example, for up to six checking accounts, the SCF asks the amount in the account, the owner of the account, and the institution where the account is held. The actual name of the institution is not retained, but a linkage is established to every other place in the interview where the institution is referenced, and detailed questions are asked about the institution. For automobiles, the make, model, and year of the car are requested along with the details of the terms of any loan for the car. Detailed descriptions of types of properties and business that the household owns are collected, along with information on the financial flows to and from the household and the businesses.

To provide adequate contextual variables for analysis, the SCF also obtains data on the current and past jobs of respondents and their spouses or partners, their pension rights from current and past jobs, their marital history, their education, the ages of their parents, and other demographic characteristics. Data are also collected on past inheritances, future inheritances, charitable contributions, attitudes, and many other variables.

Although the combination of such an broad array of variables alone is sufficient cause to warrant intensive efforts to protect the privacy of the individual survey participants, a part of the SCF sample design introduces further potential disclosure problems. The survey is intended to be used for the analysis of financial variables that are widely distributed in the population — e.g., credit card debt and mortgages — and variables that are more narrowly distributed — e.g., personal businesses and corporate stock. To

provide good coverage of both types of variables, the survey employs a dual-frame design (see Kennickell and Woodburn, 1997). In 1995, a standard multi-stage area-probability sample was selected from 100 primary sampling units across the United States (see Tourangeau, et al., 1993). This sample provides good coverage of the broadly-distributed variables. A special list sample was designed to oversample wealthy households. Under an agreement between the Federal Reserve and SOI, data from the Individual Tax File (ITF), a sample of individual tax returns specially selected and processed by SOI, are made available for sampling [6].

Data Range	*Rounded to Nearest*
>1 million	10,000
10,000 to 1 million	1,000
1,000 to 10,000	100
5 to 1,000	10
-5 to -1,000	10
-1,000 to -10,000	100
-10,000 to -1 million	1000

Negative numbers smaller than -1 million truncated at -1 million
Negative numbers between -1 and -5 unaltered

The area-probability design raises no particularly troubling issues beyond the need to protect geographic identifiers that is common to most surveys. However, the list sample raises two distinct problems. First, it increases the proportion of respondents who are wealthy. Such people are likely to be well-known at least in their locality, and because of the relatively small number of such people, it is more likely that data users with malicious intent could match a respondent to external data if sufficient information were released in an unaltered form. Second, because SOI data have been used in the design of the sample, there is a legal requirement that SCF data released to the public be subjected to a disclosure review similar to that required before the release of the public version of the ITF.

Generally, the SCF data have been released to the public in stages. This strategy has allowed us to satisfy some of the most immediate demands of data users, while allowing time to deal with more complex disclosure issues. Once a variable has been released, no amount of disclosure review can retrieve the information, and it can be much more difficult to add variables later because of the possible interactions of sensitive variables. In the past, staged release has allowed users to build a case for including additional variables, and we have been able to accommodate many such requests.

In 1992, the last year for which final data were released at the time this paper was written, the internal data were altered in the following ways for release [7]. First, geography, which was released at the level of the nine Census regions, was altered systematically; observations were sorted and aligned by some key indicators, and geography was swapped across cases.

Second, unusual categories were combined with other related categories — e.g., among owners of miscellaneous vehicles, the categories “boat,” “airplane,” and “helicopter” were combined. Third, a set of cases with unusual wealth or income were chosen, and a random set of other cases was added to the group. For these cases, key variables (for which complete responses were originally given) were multiply imputed subject to range constraints that ensured that the outcomes would be close to the initially reported values. Fourth, a set of other unspecified operations was performed to increase more broadly the perceived uncertainty associated with all variables in every observation; these operations affect both actual data values and the “shadow” variables in the dataset that describe the original state of each variable [8]. As a final step, all continuous variables were rounded as shown in Table 1. Generally, it is impossible to tell with certainty from the variables observed by a user of the public dataset which variables may have been altered and how they were altered.

Table 1.—Rounding of Continuous Variables

A similar strategy is being followed for the 1995 SCF. The one significant change is in the imputation of data for the cases deemed "sensitive" and the random subset of cases described in step three. For the 1995 survey, all monetary data items in the selected cases will be imputed. Depending on the reception of the data by users, this approach may be extended in the 1998 SCF.

FRITZ Imputation Model

Because the principal evidence reported in this paper turns critically on the imputation of monetary variables, it is important to outline some of the more important characteristics of the FRITZ model, which was originally developed for the imputation of the 1989 SCF and has been updated for each round of the survey since then. This discussion focuses on the imputation of continuous variables (see Kennickell, 1991).

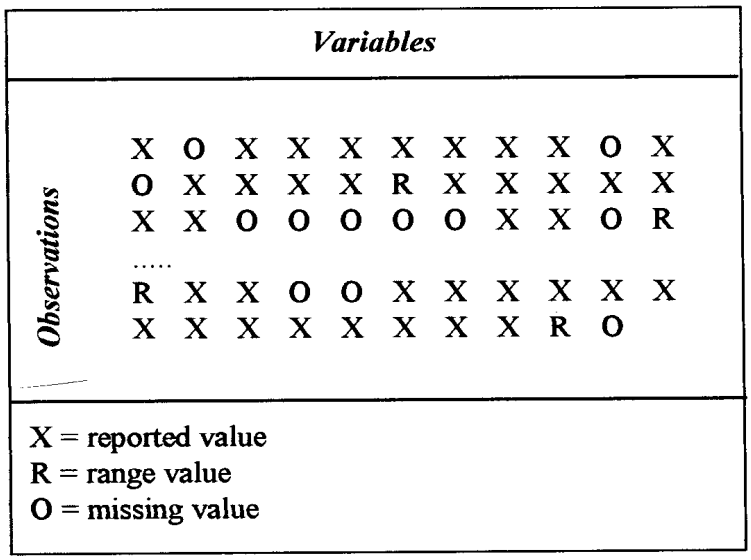
Figure 1 shows a hypothetical set of observations with various types of data given. In the figure, "X" represents complete responses, "R" symbolizes responses given as a type of range, and "O" indicates some type of missing value. In the SCF, there is a lengthy catalog of range and missing data responses, and this information is preserved in the shadow variables. Respondents in the 1995 SCF had the option of providing ranges in many ways: as an arbitrary volunteered interval (e.g., between 2,546 and 7,226), as a letter from a range card containing a fixed set of intervals (e.g., range "G" means 5,001 to 7,500), or as the result of answering a series of questions in a decision tree the intervals of which varied by question [9]. Data may be missing because the respondent did not know the answer, refused to answer, because the respondent did not answer a question of a higher order in a sequence, because of recording errors, or other reasons.

The FRITZ system is an iterative multiple imputation model based on ideas of Gibbs sampling. The system acts on a variable-by-variable basis, rather than simultaneously drawing a vector of variables [10]. Within a given iteration, the most generally applied continuous variable routine is, in essence, a type of randomized regression, in which errors are assumed to be normally distributed [11].

One factor that distinguishes the model from the usual description of randomized regression imputation models is the fact that the FRITZ model is tailored to the missing data pattern of each observation. In Figure 1, all of the missing data patterns shown are different, and they are not monotone (Little, 1983). For most continuous variables, the program generates a covariance matrix for a maximal set of variables that are determined to be relevant as possible conditioning variables. For a given case, the model first determines whether a particular variable should be imputed. Given that the variable should be imputed, the FRITZ

model computes a regression for the case using the variables in the maximal set that either are not originally missing or are already imputed within the particular iteration for the case. Finally, the model draws from the estimated conditional distribution until an outcome is found that satisfies any constraints that may apply. Constraints may take several forms. When a respondent has given a range response to a question, FRITZ uses the range to truncate the conditional distribution. Constraints may also involve cross-relationships with other variables, or simply prior knowledge about allowable outcomes. Specification of the constraints is very often the most complex mechanical part of the imputations.

Figure 1. — Hypothetical Missing Data Patterns



As noted, once a variable has been imputed, its value is taken in later imputations as if it were originally reported by the respondent. In a given imputation, variables which were originally reported as a range but are not yet imputed within the iteration, are given special treatment. Range reports often contain substantial information on the location of related variables, and one would like to use this knowledge in imputation. In the ideal, it is not difficult to write down a general model that would incorporate many types of location indicators. However, in practice such a model would quickly exhaust the degrees of freedom available in a modestly sized survey like the SCF. In practice, we adopt a compromise solution. Values reported originally as ranges are initialized at their midpoints, and these values are used as conditioning variables for other imputations until the final choice within the range is imputed.

The FRITZ model produces multiple imputations. For simplicity, the strategy adopted is to replicate each observation five times and to impute each of these "implicates" separately. Because different implicates may be imputed to take very different paths through the data, this arrangement allows users to apply standard software to the data.

The iteration process is fairly straightforward. In the first iteration, all the relevant population moments for the imputation model are computed using all available data, including all non-missing pairs of data for the covariance calculations. As imputations progress in that iteration, the covariance estimation is based on increasingly "complete" data. In the second iteration, all population moments are computed using the first iteration dataset, and a new copy of the dataset is progressively "filled in." In each successive iteration, the process is similar. Generally, the distribution of key imputations changes little after the first few iterations. Because the process is quite time-consuming, the model for the 1995 SCF was stopped after six iterations [12].

Experiments in Imputation for Disclosure Limitation

In this section, I report on three experiments in using multiple imputation for disclosure avoidance (summarized in Figure 2). In these experiments every monetary variable for every observation in the survey was imputed [13]. In the first experiment, all complete reports of dollar values were imputed as if the respondent had originally reported ranges which ran from ten percent above the actual figures to ten percent below that figure. In keeping with our usual practice of using midpoints of ranges as proxies for location indicators in imputation, the original values were retained until the variable was imputed. The second experiment also retained the reported value for conditioning, but imposed no range constraints on the allowed outcomes other than those required for cross-variable consistency. The third experiment treated the original values as if they were completely missing (that is, they were unavailable as conditioning variables) and, like the second experiment, imposed no prior bounds on the imputations; other monetary responses that were originally reported as ranges were also treated as completely missing values for purposes of conditioning, but their imputed values were constrained to lie within the reported ranges.

Figure 2. — Design of Experiments

Experiment	*Range Constraints*	*Use Original Value as Initial Location Indicator*
1	±10%	Yes
2	None	Yes
3	None	No

For several reasons, these experiments fall short of Rubin's ideal that one impute an entire dataset conditioning only on general information — even possibly using only distributional data external to the actual sample. First, the experiments deal only with the dollar variables in the SCF. Second, all complete responses other than monetary responses are used as conditioning variables. Third, the imputations of range responses are constrained to lie within the reported ranges, even in experiment three. Finally — and most probably importantly — the results are specific to the particular specification of the FRITZ model. Inevitably there are deep compromises of theory made in implementing almost any empirical system. For imputation, such compromises may be less pressing when the proportion of missing data is relatively small, as is usually the case in the SCF. These compromises may cause larger distortions when much larger fractions of the data are imputed. A key question in evaluating the results here is how well the system performs under this more extreme condition. Because we also have the originally reported values, it is possible to make a direct evaluation of the performance of the model.

Despite the shortcomings of the three experiments, they seem very much in the sprit of Rubin's proposal. Because the experiments show the effects of progressively loosening the constraints on imputation, I believe the results should provide useful evidence in evaluating the desirability of going further in developing fully simulated data.

The mechanical implementation of these experiments was reasonably straightforward. In the first experiment, the shadow variables of all complete reports of dollar values were set to a value which would normally indicate to the FRITZ model that the respondents had provided distinct dollar ranges. Values equal to the points ten percent above and 10 percent below the reported value were placed in the appropriate positions in a file that the model normally assumes contains such information. In the second and third experiments, a special value was given to the shadow variable to indicate that there were no range constraints on the imputations other those that enforce cross-variable consistency. In experiments one and two, the initial values of complete responses were left in the dataset at the beginning of imputation; during the course of imputations, these values were used for conditioning until they were replaced by an imputed value, which was used to condition subsequent imputations. In experiment three, values originally reported completely were set to a missing value, and the usual midpoints of range responses were also set to a missing value. Thus, no dollar variables in the third experiment were available for conditioning until they were imputed. In each of the experiments, the imputations were treated as if they were the seventh iteration of the SCF implementation of FRITZ. Thus, estimates of the population moments needed for the model were computed using the final results of the sixth iteration.

In the absence of technical problems — far from the case with the work for this paper for which the imputation system was subject to a massively larger than normal stress — each version of the experiment would require approximately three weeks to run through the entire dataset on a fast dedicated Sun server. More importantly, each execution would also require about 2 gigabytes of disk space for the associated work files. The process could probably be made at least somewhat more efficient, but the time available for debugging such a potentially complex change was limited. A compromise has been adopted here. The first of the eight modules of the SCF application of FRITZ was run for all of the experiments. This module deals largely with total household income and various financial assets.

Figures 3 through 6 show descriptive plots of data from the three experiments for the following four variables: total income, amount in the first savings account, the amount of Treasury bills and other Federal bonds (referred to hereafter as "T-bills"), and the total value of financial assets [14]. The first three of these variables are intended to span a broad set of types distributions; total financial assets, a variable constructed from many components, is included to show the effects of aggregating over the potentially large number of responses to questions about the underlying components. The impression from looking at a broader set of variables is very similar. Each of the figures is divided into two sets of three panels. The top three panels show the distribution for experiments one through three, of the (base-10) logarithm of the originally reported values less the average across the five implicates of the logarithm of the corresponding imputed values ("bias"), where the distribution is estimated as an unweighted average shifted histogram (ASH). The bottom three panels are ASH plots for the three experiments, of the distribution over all cases of the standard deviation of the multiply-imputed values within observations.

For experiment one, the distribution of bias has a mode at approximately zero for all the variables. This is not surprising given that the outcome is based on models estimated using reported data for these observations. In the case of income, savings balances, and T-bills, the distribution of bias is fairly concentrated, with the 10th and 90th percentiles of the distribution corresponding to a bias of only about 5 percent (±0.02 on the scale shown). The distributions of bias for savings accounts and T-bills are

Figure 3a: ASH Plots of Distribution over Observations of Log_{10}(Actual Value)-Mean(Log_{10}(Imputation)) within Observations, Total Household Income, Experiments 1-3.

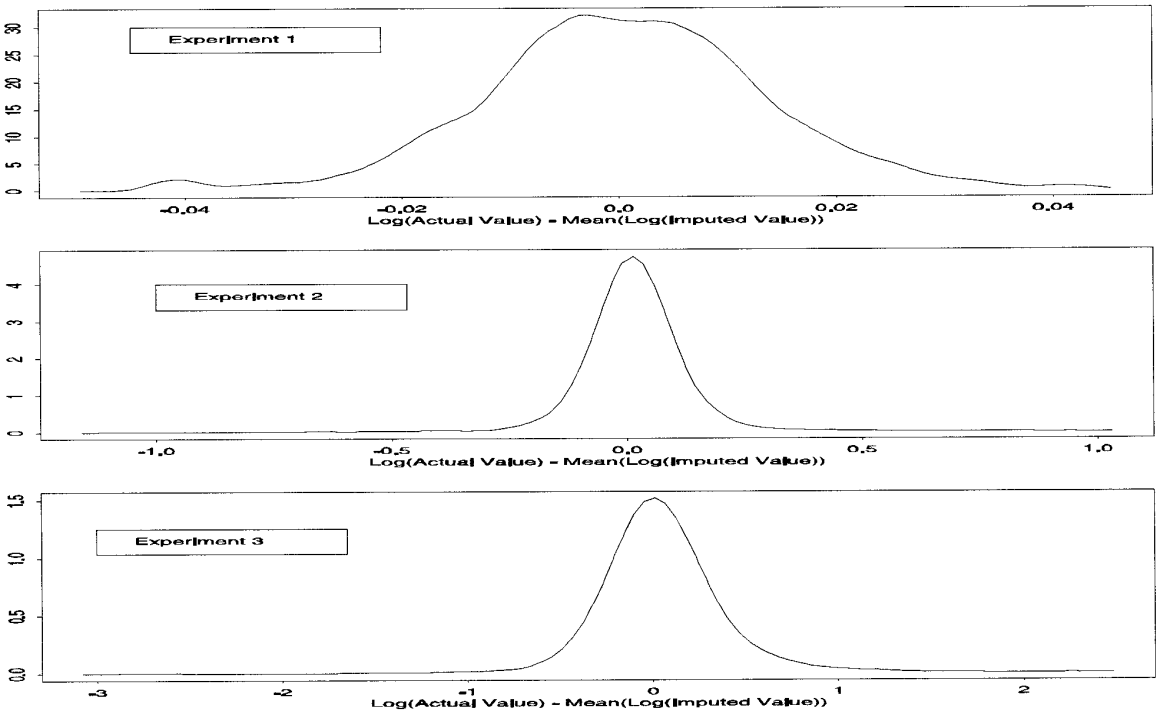


Figure 3b: ASH Plots of Distribution over Observations of the Standard Deviation of Log_{10}(Imputation) within Observations, Total Household Income, Experiments 1-3.

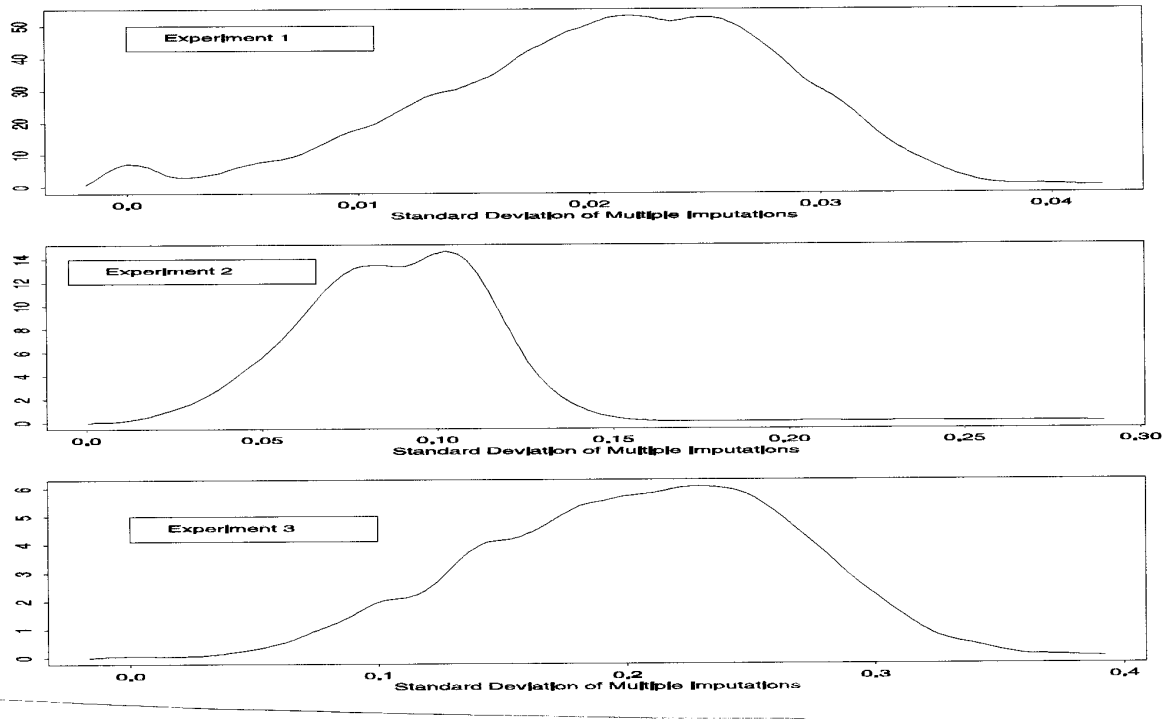


Figure 4a: ASH Plots of Distribution over Observations of Log_{10}(Actual Value)-Mean(Log_{10}(Imputation)) within Observation, Balance in First Savings Account, Experiments 1-3.

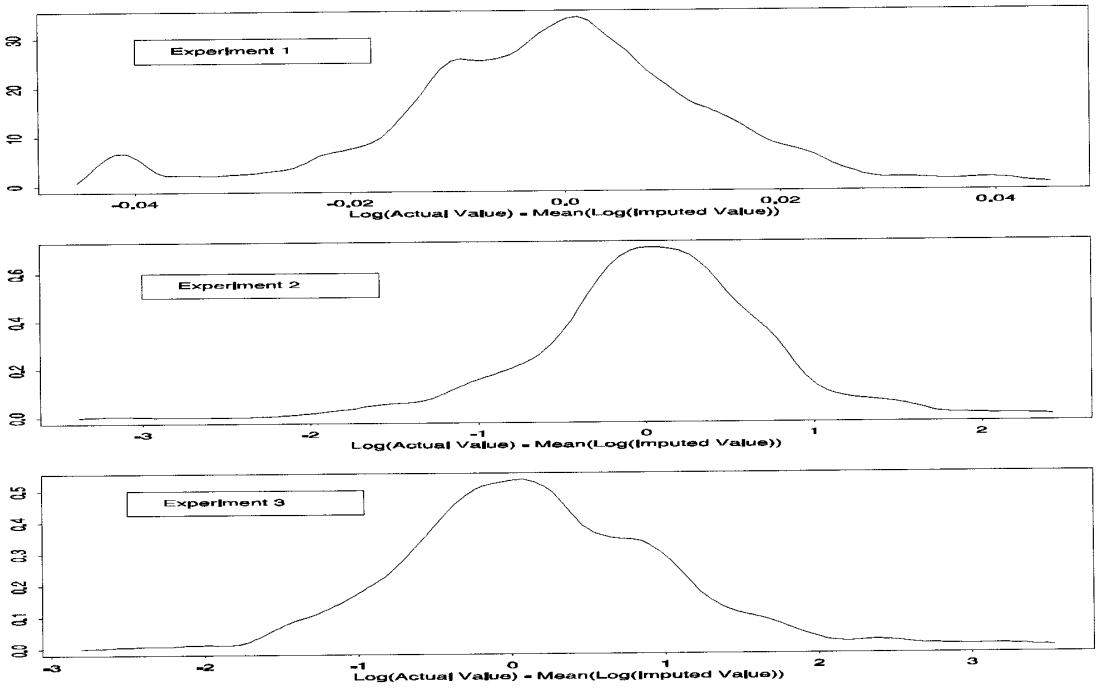


Figure 4b: ASH Plots of Distribution over Observations of the Standard Deviation of Log_{10}(Imputation) within Observations, Balance in First Savings Account, Experiments 1-3.

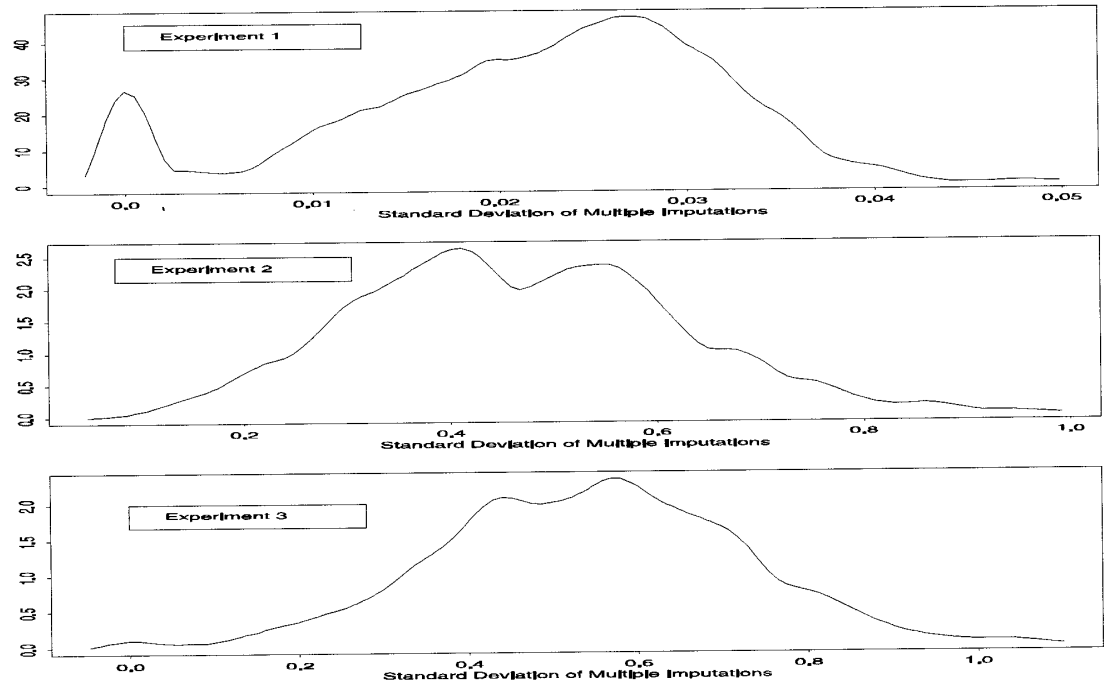


Figure 5a: ASH Plots of Distribution over Observations of Log_{10}(Actual Value)-Mean(Log_{10}(Imputation)) within Observations, Face Value of T-Bills, Experiments 1-3.

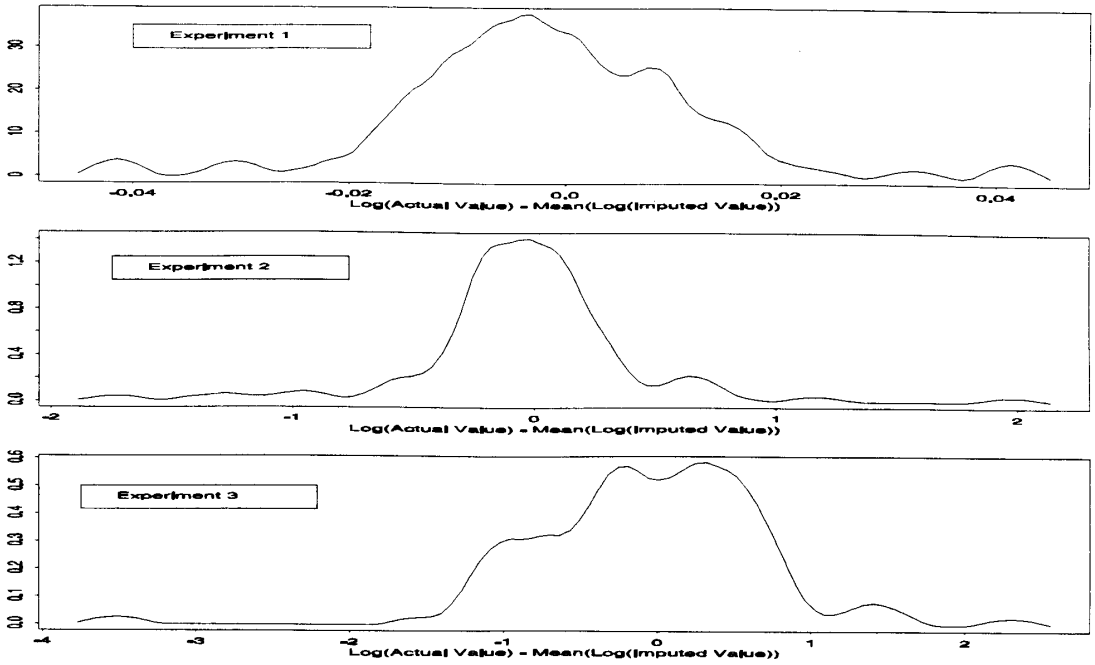


Figure 5b: ASH Plots of Distribution over Observations of the Standard Deviation of Log_{10}(Imputation) within Observations, Face Value of T-Bills, Experiments 1-3.

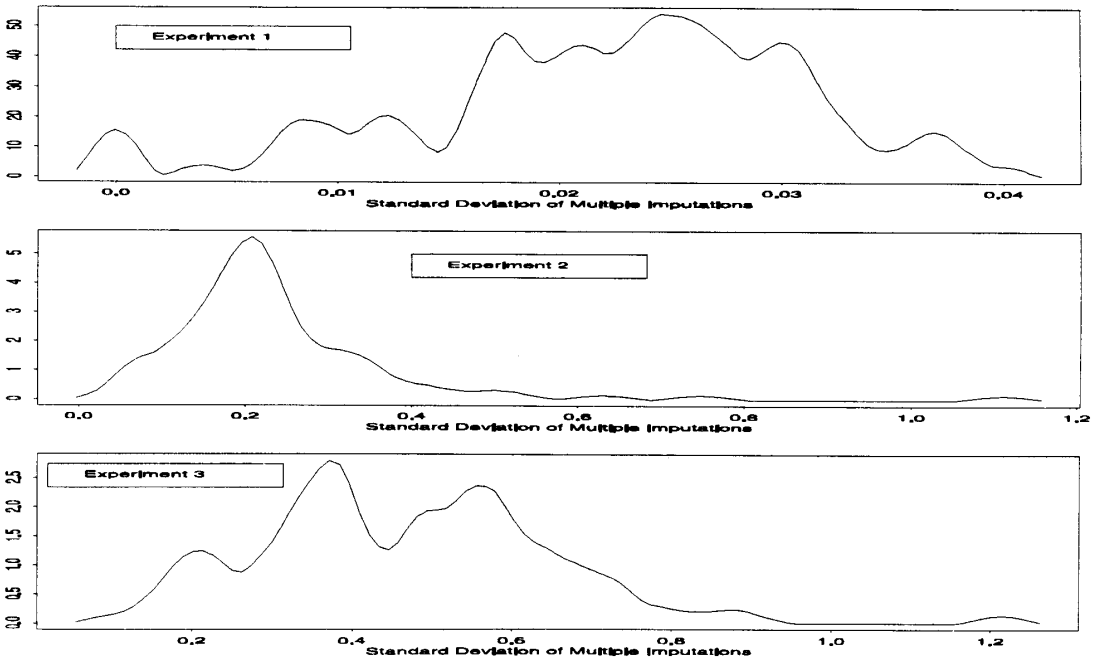


Figure 6a: ASH Plots of Distribution over Observations of Log_{10}(Actual Value)-Mean(Log_{10}(Imputation)) within Obervations, Total Financial Assets, Experiments 1-3.

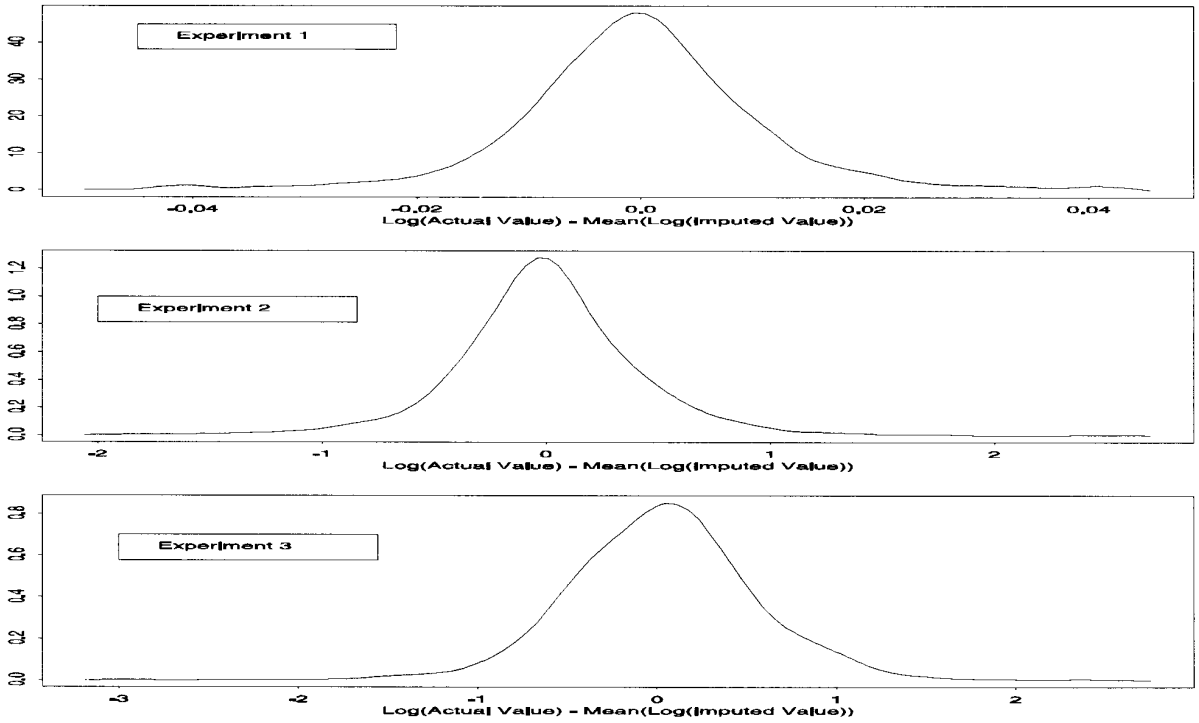
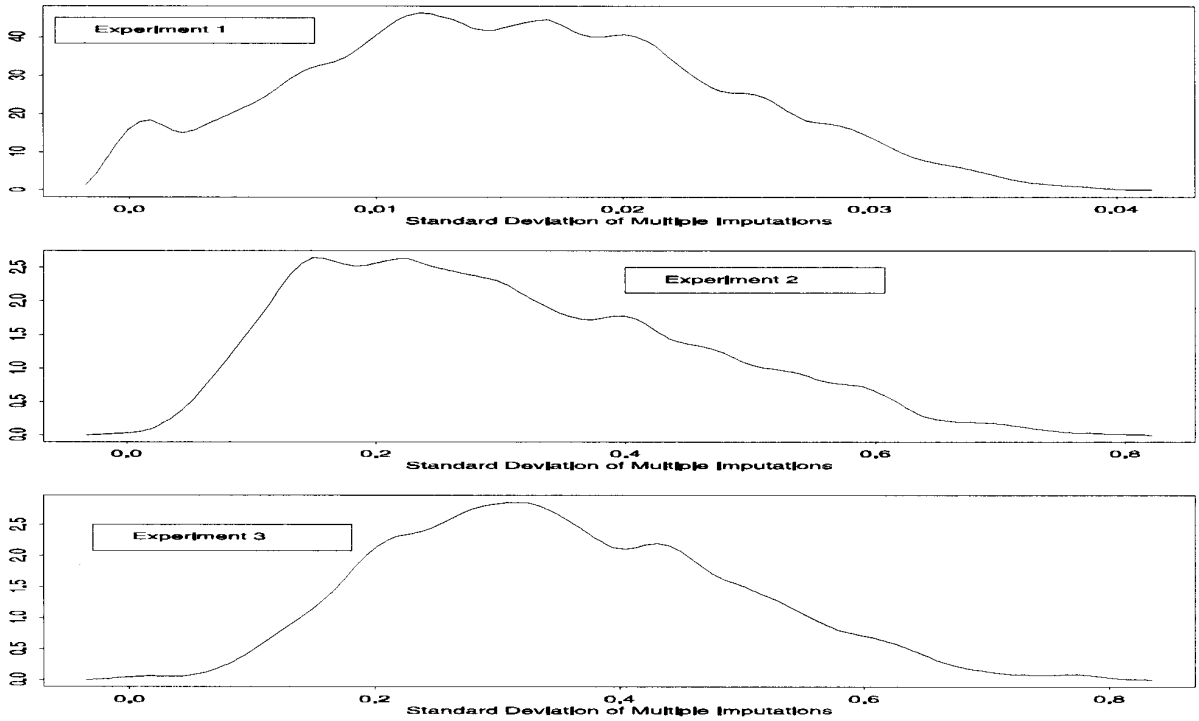


Figure 6b: ASH Plots of Distribution over Observations of the Standard Deviation of Log_{10}(Imputation) within Observations, Total Financial Assets, Experiments 1-3.



relatively "lumpy," largely reflecting the smaller samples used to estimate these distributions: about 1,200 observations were used for the savings account estimate and only about 110 observations were used for the T-bill estimate, but about 2,900 were used to estimate the distribution for total income. Reflecting the integration over possibly many imputations, the distribution of bias for total financial assets is quite smooth. In every case shown, there is some piling up of cases at the outer bounds corresponding to ±10 percent (about ±0.04 on the log scale). The FRITZ model is allowed to draw as many as 400 times from the predicted conditional distribution of the missing data before selecting the nearest endpoint of the constraint. Thus, it is likely that these extreme observations are ones for which the models do not fit very well. Not surprisingly, examination of selected cases suggests that these observations are more likely to have unusual values for some of the conditioning variables in the imputation models. The median variability of the imputations within implicates shown by the ASH plots of the distributions of standard deviations, is about ±6 percent for income, savings accounts, and T-bills. The variability within implicates is substantially lower for the sum of financial assets, reflecting offsetting errors in imputation.

In the second experiment, the relaxation of the simple range constraint in experiment one has the expected effect of increasing the variability of the bias, and increasing the standard deviation of imputations within implicates. In the case of total household income, the bias corresponding to the 90th percentile of the bias distribution jumps to about 25 percent. The effect is even larger for the other variables (the bias is nearly 300 percent at the 90th percentile for total financial assets). It is somewhat surprising just how much these values increase given that the imputations are potentially conditioned on a large number of reported values [15].

In the third experiment with the removal of the reported values used for conditioning in experiment two, the range of the bias rises further. The 90th percentile of the bias distribution is about 140 percent for total income, and about 400 percent for total financial assets.

Because these results are reported on a logarithmic scale, it is possible that they could be unduly influenced by changes that are small in dollar amounts, but large on a logarithmic scale. The data do not provide strong support for this proposition. For income, scatterplots reveal that the logarithmic bias appears to be approximately equally spread at all levels of income for experiments one and two [16]. In the third experiment, the dominant relationship is similar, but there are two smaller groups that deviate from the pattern: a few dozen observations with actual incomes of less than a few thousand dollars are substantially over-imputed on average, and a somewhat larger number of observations with actual incomes of more than $100,000 are substantially under-imputed. The data suggest a similar relationships across the experiments for the other variables as well.

To gauge the effects of the experiments on the overall univariate distributions of the four variables considered, Figures 7-10 show quantile-quantile (Q-Q) plots of the mean imputations against the reported values on a logarithmic scale. Across these variables, the distribution is barely affected by experiment one. In the second experiment, the results are a bit more mixed. For total income and total financial assets, there is some over-imputation of values less than a few thousand dollars, and slight under-prediction at the very top. For T-bills, the relationship is much noiser, but not strikingly different. However, for savings accounts, the Q-Q plot is rotated clockwise, indicating that the imputed distribution is under-imputed at the top and over-imputed at the bottom. All of the simulated distributions deteriorate in the third experiment, though the distribution of total financial assets appears the most resilient [17].

Figure 7: Q-Q Plots of Imputed Distribution vs. Actual Distribution, Total Household Income, Experiments 1-3.

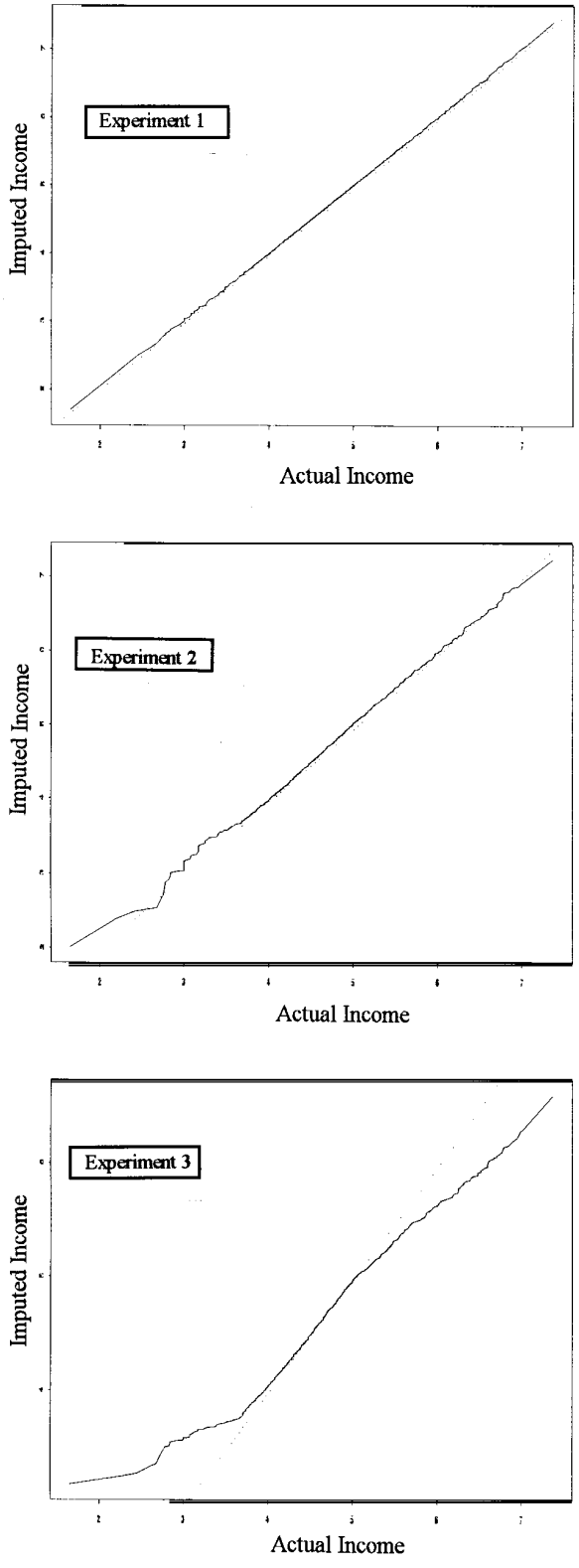


Figure 8: Q-Q Plots of Imputed Distribution vs. Actual Distributrion, Balance in 1st Savings Account, Experiments 1-3.

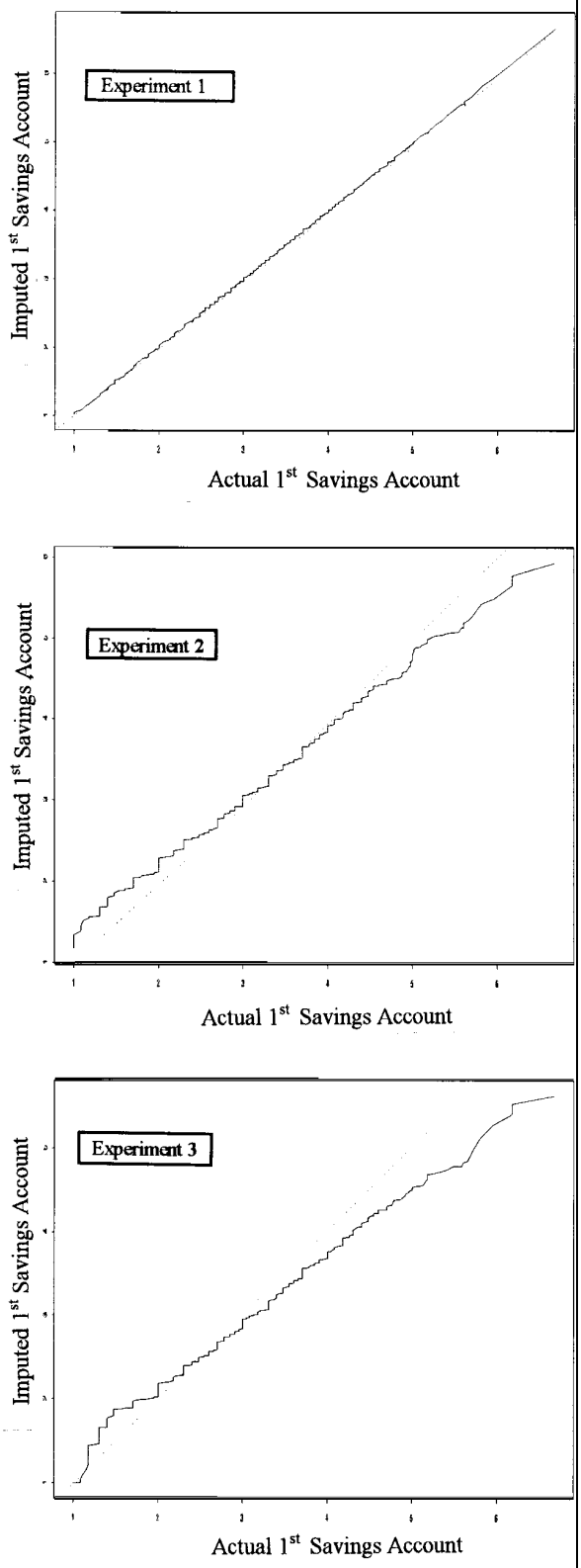
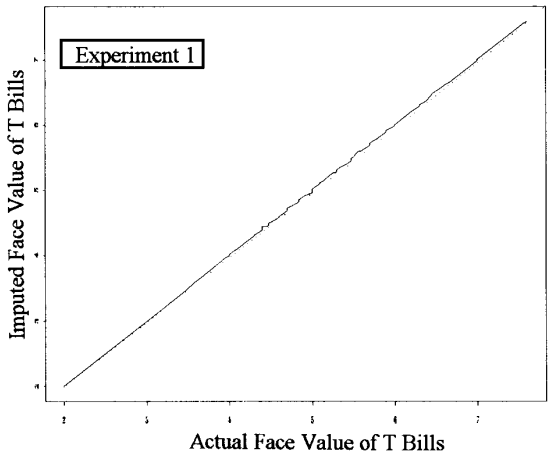
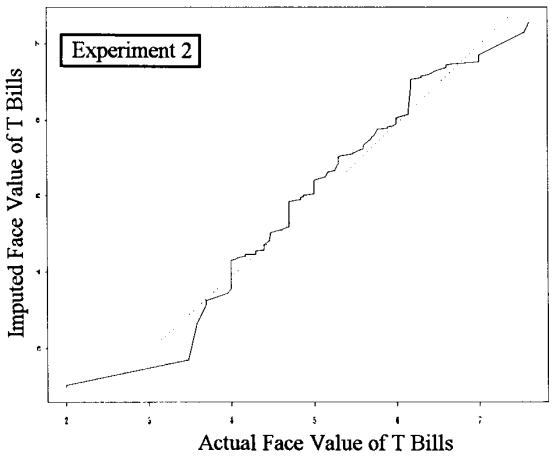


Figure 9: Q-Q Plots of Imputed Distribution vs. Actual Distribution, Face Value of T-Bills, Experiments 1-3.





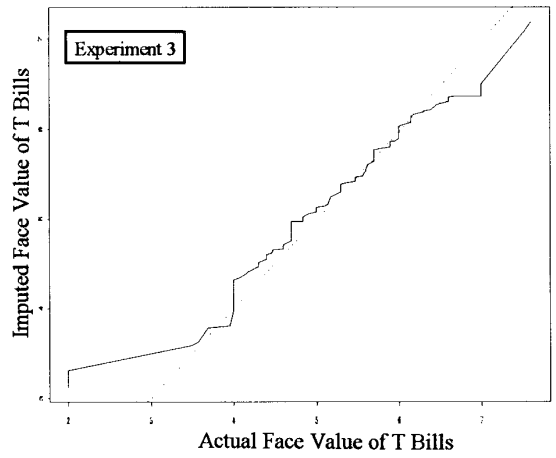
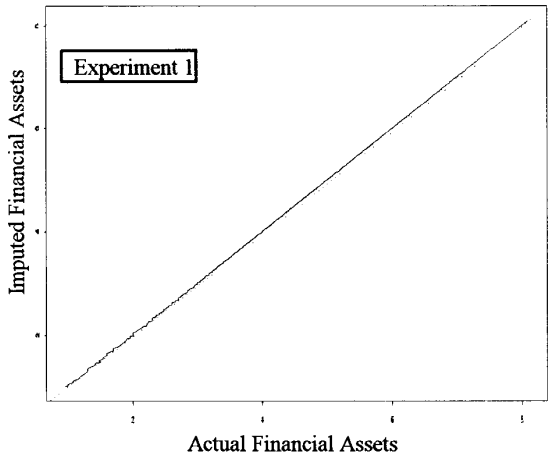
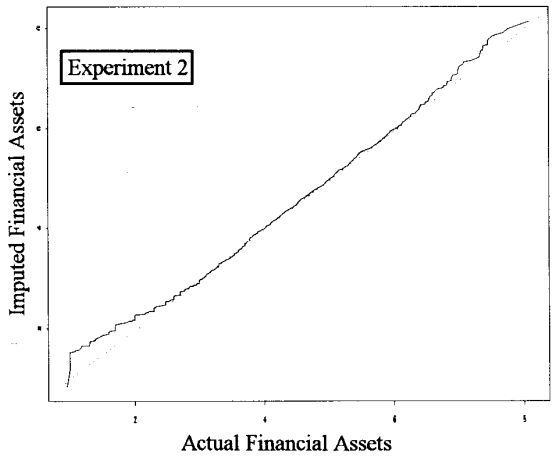
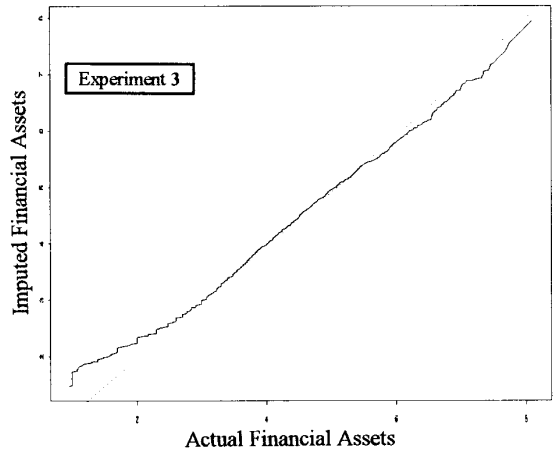


Figure 10: Q-Q Plots of Imputed Distribution vs. Actual Distribution, Total Financial Assets, Experiments 1-3.







Univariate and simple bivariate statistics are important for many illustrative purposes, but for the SCF, as is the case for many other surveys, the most important uses of the data over the long run are in modeling. Table 3 presents the coefficients of a set of simple linear regressions of the logarithm of total household income on dummy variables for ownership of various financial assets and the log of the maximum of one and the value of the corresponding asset. This model has no particular importance as an economic or behavioral characterization. It is intended purely as a descriptive device designed to examine the effects of the variation across the experiments on the partial correlations of a set of variables imputed in all the experiments. Two types of models are shown: one set includes all observations regardless of whether the variables included were originally reported completely by the respondent, and the other model includes only cases for which every variable in the model was originally reported completely. The regressions were run using data from each of the three experiments, as well as data from the final version of the sixth (final) iteration of the imputation of the main dataset [18].

Experiments one and two perform about equally well in terms of determining the significance of coefficients in both variations on the basic model. However, data from the first experiment misclassify one variable as not significant, and data from the second experiment misclassify some variables as significant. The third experiment implies both type one and type two errors. The R^2 of the regressions changes little except in the third experiment, where this value drops about 10 percent. Overall, none of the experiments do dramatically worse than the original data. Given the structure of the FRITZ model and the degree to which the variables in these regression models were mutually interdependent, it would be very surprising if the outcome were otherwise. However, such regressions are only the beginning of what many economist would consider applying to the data, and it is possible that more complex models or methods of estimation would give different results.

Summary and Future Research

By design, experiment one is virtually guaranteed to induce minimal distortions, but it also leaves the outcomes near the original values. Unfortunately, just knowing that an outcome is in a certain range may already be sufficient information to increase too much the probability of identifying some of the very wealthy respondents in the SCF. My ex ante choice of contenders among the experiments was the second one, in which imputations condition on actual values, but there is no prior constraint on the outcome that is connected to the original value. Ex post, I find the results relatively disappointing. Certainly, the reported outcomes of the third experiment look least attractive. There may be ways of more globally constraining or aligning the outcomes of experiments two and three, but I suspect the choice of method would depend critically on a ranking of the importance of the types of analyses to be performed with the data. I hope that someone in the SCF group or elsewhere will be able to take the next step.

One technical question that appears potentially troublesome is how to estimate sampling error in a fully simulated dataset [19]. It is possible, in theory, to simulate records for the entire universe, but even in this case there would still be sampling variability in the imputations. This variation may be a second order effect in normal imputation, but we need to deal with the issue carefully if we expect to simulate all the data. Perhaps we could find an approximate solution in independently multiply imputing each of a manageably small number of replicates — implicates of replicates; each replicate would require population estimates from a corresponding replicate selected from the actual data in a way that captured the important dimensions of variability in the sample. Another possibility might be to compute variance functions from the actual data.

Table 3. — Regression of Logarithm of Total Household Income on Various Variables, Original Data and Experiments 1-3, Using all Observations and Using Only Observations Originally Giving Complete Responses to all Variables in the Model

	All Observations Included				Only Complete Responders Included			
	Orig.	Exp. 1	Exp. 2	Exp. 3	Orig.	Exp. 1	Exp. 2	Exp. 3
Intercept	2.64*	1.92*	2.56*	3.76*	2.83*	2.87*	3.43*	6.60*
	0.75	*0.75*	*0.74*	*0.69*	*1.09*	*1.09*	*1.02*	*1.09*
Have checking	0.18*	0.20*	0.25*	0.21*	0.17*	0.18*	0.18*	0.15*
	0.03	*0.03*	*0.03*	*0.03*	*0.04*	*0.04*	*0.04*	*0.04*
Ln($ checking)	0.25*	0.27*	0.30*	0.26*	0.26*	0.27*	0.27*	0.23*
	0.01	*0.01*	*0.01*	*0.01*	*0.02*	*0.02*	*0.02*	*0.02*
Have IRA/Keogh	0.16*	0.18*	0.18*	0.17*	0.07	0.06	0.12	0.08
	0.05	*0.05*	*0.05*	*0.05*	*0.07*	*0.07*	*0.07*	*0.07*
Ln($ IRA/Keogh)	0.10*	0.11*	0.11*	0.10*	0.07*	0.07*	0.10*	0.08*
	0.02	*0.02*	*0.02*	*0.02*	*0.03*	*0.03*	*0.03*	*0.03*
Have savings acct.	0.01	0.02	0.01	0.01	-0.03	-0.03	-0.02	-0.03
	0.04	*0.04*	*0.04*	*0.04*	*0.04*	*0.05*	*0.05*	*0.05*
Ln($ savings acct)	0.03	0.03	0.03	0.04	0.00	0.00	0.01	0.01
	0.02	*0.02*	*0.02*	*0.02*	*0.02*	*0.02*	*0.02*	*0.02*
Have money market acct.	0.02	0.03	-0.04	-0.11	0.11	0.12	0.01	-0.07
	0.07	*0.07*	*0.07*	*0.07*	*0.09*	*0.10*	*0.10*	*0.10*
Ln($ money market acct.)	0.03	0.03	0.00	-0.02	0.05	0.05	0.01	-0.02
	0.03	*0.03*	*0.03*	*0.03*	*0.04*	*0.04*	*0.04*	*0.04*
Have CDS	0.24*	0.26*	0.31*	0.27*	0.22*	0.22*	0.27*	0.23*
	0.08	*0.08*	*0.08*	*0.08*	*0.10*	*0.11*	*0.11*	*0.11*
Ln($ CDS)	0.07*	0.07*	0.09*	0.08*	0.07	0.07	0.09*	0.07
	0.03	*0.03*	*0.03*	*0.03*	*0.04*	*0.04*	*0.04*	*0.04*
Have savings bonds	-0.02	-0.01	-0.05	-0.09	-0.10	-0.10	-0.12	-0.13
	0.04	*0.04*	*0.05*	*0.04*	*0.06*	*0.06*	*0.06*	*0.05*
Ln($ savings bonds)	0.02	0.02	0.00	-0.02	-0.03	-0.03	-0.05	-0.04
	0.02	*0.02*	*0.02*	*0.02*	*0.03*	*0.03*	*0.03*	*0.03*
Have other bonds	0.62*	0.65*	0.51*	0.63*	0.68*	0.66*	0.54*	0.35*
	0.09	*0.09*	*0.08*	*0.09*	*0.14*	*0.14*	*0.13*	*0.14*
Ln($ other bonds)	0.26*	0.27*	0.22*	0.25*	0.27*	0.26*	0.22*	0.15*
	0.03	*0.03*	*0.03*	*0.03*	*0.05*	*0.05*	*0.04*	*0.05*
Have mutual funds	0.06	0.07	0.09	-0.02	0.18	0.17	0.20*	0.00
	0.07	*0.07*	*0.07*	*0.05*	*0.09*	*0.09*	*0.09*	*0.06*
Ln($ mutual funds)	0.04	0.05	0.05*	0.01	0.10*	0.09*	0.10*	0.03
	0.02	*0.02*	*0.02*	*0.02*	*0.03*	*0.03*	*0.03*	*0.03*
Have annuity/trust	0.02	0.02	0.03	0.01	-0.04	-0.04	-0.07	-0.07
	0.04	*0.04*	*0.04*	*0.02*	*0.05*	*0.05*	*0.05*	*0.05*
Ln($ annuity/trust)	0.04*	0.04*	0.04*	0.02	0.01	0.01	0.01	-0.29
	0.01	*0.01*	*0.01*	*0.02*	*0.02*	*0.02*	*0.02*	*0.24*
Have whole life insurance	-0.70	0.11	0.14*	0.19*	-0.61	-0.63	0.17*	0.2*
	0.17	*0.08*	*0.05*	*0.05*	*0.25*	*0.26*	*0.07*	*0.06*
Ln($ cash value life ins.)	0.10*	0.01	0.02*	0.01	0.09*	0.09*	0.03	0.02
	0.02	*0.01*	*0.01*	*0.01*	*0.03*	*0.03*	*0.04*	*0.03*
R^2	0.40	0.39	0.40	0.37	0.43	0.43	0.42	0.36

* = significant at the 95% level of confidence.
Simple regression standard errors are given in italics below each estimate.

The experimental results reported in this paper say at least as much about the nature of the SCF imputations as they do about the possibility of creating a fully simulated dataset. Although the imputation models have been refined over three surveys now, the results of experiments two and three, in particular, suggest that there is room for improvement. Indeed, a number of changes were instituted in the process of getting the experiments to produce meaningful data, and other changes will be implemented during the course of processing the 1998 SCF. Other changes, including the possibility of using empirical residuals, deserve further attention. However, I am not optimistic that there are many major improvements in our ability to impute the SCF data waiting to be discovered. There is a difference in what one can accept in imputing a relatively small fraction of the data and what is acceptable for the whole dataset. With fully simulated data, we are left with a difficult tradeoff between noise (however structured) and potential disclosure.

Disclosure limitation techniques have a Siamese twin in record linkage techniques. As one side progresses, the other side uses closely related ideas to follow. This conference has played an important part in highlighting this relationship and the need for coordination. Perhaps if we work hard together, there may be a chance that we will find a way to allow users to analyze disclosure-limited data using record linkage ideas to sharpen inferences. There may also be a payoff in more routine statistical matching, which is really just another form of imputation.

A large problem in planning all disclosure reviews is how to accommodate the needs (but not necessarily all the desires) of data users. I expect that users will express considerable resistance to the idea of completely simulated data. Some statisticians may be troubled about how to address questions of estimating sampling error with such data. Among economists, there are substantial pockets of opposition to all types of imputation, and some researchers have raised carefully framed questions that need to be addressed equally carefully. For example, if unobserved effects are a serious issue (and they often are in econometric modeling), then imputation must consider the distortions it may induce if such latent models are ignored; the question becomes much more pressing if all of the data are imputed. Given the choice between having no data or having data that are limited in some way, most analysts will likely opt for some information. However, to avoid developing disclosure strategies that yield data that do not inform interesting questions for users, it may be important to engage users in the process where possible.

Acknowledgments

The author wishes to thank Kevin Moore and Amy Stubbendick for a very high level of research assistance. The author is also grateful to Gerhard Fries, Barry Johnson, and R. Louise Woodburn for comments, and to Fritz Scheuren for encouragement in this project.

Footnotes

[1] As Fienberg (1997) argues, releasing any information discloses something about the respondent, even if the probability of identification is minuscule.

[2] See Fries, Johnson and Woodburn (1997a) for a summary of the disclosure strategies that have been developed for the survey.

[3] Ivan Fellegi emphasized a similar point in his address to this conferences.

[4] For example, Rubin (1993) says "Under my proposal, no actual unit's confidential data would ever be released. Rather, all actual data would be used to create a multiply-imputed synthetic microdata set of artificial units..."

[5] However, Fienberg and Makov (1997) have proposed creating simulated data for the purpose of evaluating the degree of disclosure risk in a given dataset and Feinberg, Steele and Makov (1996) have examined the problem of simulating categorical data.

[6] Use of the ITF for the SCF is strictly controlled to protect the privacy of taxpayers. For the 1995 SCF, SOI provided NORC with the names and addresses of a sample selected from a copy of the ITF purged of name and address information at the Federal Reserve. NORC contacted respondents, but had no means of linking to the tax data. The SCF group alone at the Federal Reserve is allowed access to both survey data and tax data, but no names were available, and use of these tax data at the Federal Reserve is strictly limited to activities connected with sampling, weighting, and other such technical issues.

[7] See Fries, Johnson, and Woodburn (1997b) for details and information about the effects of the alterations on the data.

[8] The shadow variables are used as a formal device in documentation, and they inform the imputation software about which variables should be imputed. The shadow variables contain information about various types of editing that may have been performed to reach the final value, whether it was reported as one of a large number of types of range outcomes, whether it was missing for various reasons, or whether its outcome was affected by other processes.

[9] The collection of range data in the 1995 SCF is described in detail in Kennickell (1997).

[10] For an excellent example of a simultaneously determined system, see Schafer (1995). Geman and Geman (1984) discuss another type of structure involving data "cliques."

[11] In general, continuous variables are assumed to follow a conditional lognormal distribution. For continuous variables, the program assumes by default that errors should be drawn within a bound of 1.96 standard errors above and below the conditional mean.

[12] For the 1995 data, the process required about ten days per iteration, which is down from about four weeks per iteration in 1989.

[13] There are 480 monetary variables in the SCF, but it is not possible for a given respondent to be asked all of the underlying questions.

[14] The sets of observations underlying the charts include only respondents who gave a complete response for the variable, or, in the case of financial assets, who gave complete responses for all the components of financial assets. For many sub-models of the SCF implementation of FRITZ, general constraints are imposed for all imputations to ensure values that are reasonable (e.g., amounts owed on mortgage balloon payments must be less than or equal to the current amount owed); in the actual data, these constraints are occasionally violated for reasons that are unusual, but possible. When reimputing these values subject to dollar range constraints in experiment one, a small number of imputations violated the bounds imposed. To avoid major restructuring of the implementation of the FRITZ model for the experiments, these instances are excluded from the comparisons reported here. In each of the figures, the set of observations is the same across all six of the panels. For the income plots, households reporting negative income have been excluded.

[15] For example, total income is the first variable imputed, and all reported values (or midpoints of ranges) for variables included in the model for that variable are used to condition the imputation.

[16] For disclosure reasons, the scatterplots supporting this claim cannot be released.

[17] In the cases examined, this result also holds if the data are separated by implicates rather than averaged across implicates.

[18] The five implicates were pooled for these regressions. Standard errors shown in the table are simple regression standard errors that take no account of imputation or sampling error; the degrees of freedom were altered in the standard error calculation to reflect the fact that there were five times as many implicates as observations.

[19] Fienberg, Steele and Makov (1996) also address this question.

References

Fienberg, Stephen E. (1997). Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistics Research, working paper, Department of Statistics, Carneige Mellon University, Pittsburgh, PA.

Fienberg, Stephen E. and Makov, Udi E. (1997). Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data, working paper, Department of Statistics, Carneige Mellon University, Pittsburgh, PA.

Fienberg, Stephen E.; Steele, Russell J.; and Makov, Udi E. (1996). Statistical Notions of Data Disclosure Avoidance and their Relationship to Traditional Statistical Methodology: Data Swapping and Loglinear Models, *Proceedings of the 1996 Annual Research Conference and Technology Interchange*, Washington, DC: U.S. Bureau of the Census, 87-105.

Fries, Gerhard; Johnson, Barry W.; and Woodburn, R. Louise (1997a). Analyzing Disclosure Review Procedures for the Survey of Consumer Finances, paper for presentation at the 1997 Joint Statistical Meetings, Anaheim, CA.

Fries, Gerhard; Johnson, Barry W.; and Woodburn, R. Louise (1997b). Disclosure Review and Its Implications for the 1992 Survey of Finances, *Proceedings of the Section on Survey Research Methods*, 1996 Joint Statistical Meetings, Chicago, IL.

Geman, Stuart and Geman, Donald (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6, 6 (November), 721-741.

Kennickell, Arthur B. (1991). Imputation of the 1989 Survey of Consumer Finances, *Proceedings of the Section on Survey Research Methods*, 1990 Joint Statistical Meetings, Atlanta, GA.

Kennickell, Arthur B. and Woodburn, R. Louise (1997). Consistent Weight Design for the 1989, 1992 and 1995 SCFs, and the Distribution of Wealth, working paper, Board of Governors of the Federal Reserve System, Washington, DC.

Kennickell, Arthur B. (1997). Using Range Techniques with CAPI in the 1995 Survey of Consumer Finances *Proceedings of the Section on Survey Research Methods*, 1996 Joint Statistical Meetings, Chicago, IL.

Little, Roderick J.A. (1983). The Nonignorable Case, *Incomplete Data in Sample Surveys*, New York: Academic Press.

Rubin, Donald B. (1993). Discussion of Statistical Disclosure Limitation, *Journal of Official Statistics*, 9, 2, 461-468.

Schafer, Joseph (1995). *Analysis of Incomplete Multivariate Data*, Chapman and Hall.

Tourangeau, Roger; Johnson, Robert A.; Qian, Jiahe; Shin, Hee-Choon; and Frankel, Martin R. (1993). Selection of NORC's 1990 National Sample, working paper, National Opinion Research Center at the University of Chicago, Chicago, IL.

The views presented in this paper are those of the author alone and do not necessarily reflect those of the Board of Governors or the Federal Reserve System. Any errors are the responsibility of the author alone.

Chapter 8

Sharing Statistical Information for Statistical Purposes

Katherine K. Wallman and Jerry L. Coffey
Office of Management and Budget

Abstract

Congress has recognized that a confidential relationship between statistical agencies and their respondents is essential for effective conduct of statistical programs. However, the specific statutory formulas devised to implement this principle in different agencies have created difficult barriers to effective working relationships among these agencies. The development of mechanisms to establish a uniform confidentiality policy that substantially eliminates the risks associated with sharing confidential data will permit significant improvements in data used for both public and private decisions without compromising public confidence in the security of information respondents provide to the Federal government.

Initiatives of the Statistical Policy Office to enhance public confidence in the stewardship of sensitive data and to permit limited sharing of confidential data for exclusively statistical purposes received a substantial impetus in the 1995 reauthorization of the Paperwork Reduction Act. The Act strongly endorses the principles embodied in statistical confidentiality pledges and charges OMB to promote sharing of data for statistical purposes within a strong confidentiality framework.

This paper discusses the history, the promise, and the current status of initiatives to strengthen and improve data protection while promoting expanded data sharing for statistical purposes. The most recent efforts include the OMB Federal Statistical Confidentiality Order, the Statistical Confidentiality Act (SCA), and companion legislation to the SCA, that would make complementary changes to the Internal Revenue Code.

Introduction

A promising initiative to improve the quality and efficiency of Federal statistical programs is a legislative proposal that would allow the sharing of confidential data among statistical agencies under strict safeguards. The development of this approach has been a painstaking, careful process that has been supported and nurtured by Administrations of both parties over many years.

The Administration's *Statistical Confidentiality Act* and two companion initiatives -- the *OMB Federal Statistical Confidentiality Order* and an *amendment to the Internal Revenue Code* -- address two issues that are vital to ensuring the integrity and efficiency of Federal statistical programs and, ultimately, the quality of

Federal statistics. These are

- the unevenness of current *statutory* protections for the confidential treatment of information provided to statistical agencies for exclusively statistical purposes; and

- the barriers to effective working relationships among the statistical agencies that stem from slightly different statutory formulas devised to implement the principle of confidentiality for statistical data in different agencies.

The proposed legislation would establish policies and procedures to guarantee the consistent and uniform application of the confidentiality privilege and authorize the limited sharing of information among designated statistical agencies for exclusively statistical purposes.

Initiatives Span More Than Two Decades

Efforts to address confidentiality concerns with regard to Federal statistical data have a history that extends for more than 25 years. Such efforts have been endorsed on both sides of the aisle in the Congress. The roots of the policies in the Administration's current Statistical Confidentiality Act reflect the work of three Commissions that examined statistical and information issues during the Administrations of Presidents Nixon and Ford. In 1971, the President's Commission on Federal Statistics recommended that the term *confidential* should always mean that disclosure of data in a manner that would allow public identification of the respondent or would in any way be harmful to him should be prohibited; this commission also recommended that consideration should be given to providing for interagency transfers of data where confidentiality could be protected.

In July 1977, the Privacy Protection Study Commission stated that "no record or information... collected or maintained for a research or statistical purposes under Federal authority... may be used in individually identifiable form to make any decision or take any action directly affecting the individual to whom the record pertains..." Later, in October of that year, the President's Commission on Federal Paperwork endorsed the confidentiality and functional separation concepts, but applied them directly and simply to statistical programs, saying that:

- Information collected or maintained for *statistical purposes* must never be used for administrative or regulatory purposes or disclosed in identifiable form, except to another statistical agency with assurances that it will be used solely for statistical purposes; and

- Information collected for *administrative and regulatory purposes* must be made available for statistical use, with appropriate confidentiality and security safeguards, when assurances are given that the information will be used solely for statistical purposes.

The policy discussions generated by the three Commissions came together during the Carter Administration in a bipartisan outpouring of support for the Paperwork Reduction Act (PRA), which largely addressed the *efficiency* recommendations of the Paperwork Commission. The legislative history of that Act recognized the unfinished work of fitting the functional separation of statistical information into the overall scheme.

The first attempt to deal with the issues of confidentiality and sharing of statistical data was made by the Carter Administration's Statistical Reorganization Project (popularly known as the "Bonnen Commission"). This effort paralleled the legislative development work by OMB that became the Paperwork Reduction Act. The initiative identified a group of statistical agencies that could serve as protected environments -- or *enclaves* -- for confidential data and attempted to create a harmonized confidentiality policy by synthesizing the several prescriptions in existing laws. The initiative was left behind by the fast-track PRA for two reasons:

- First, each new prescription to solve problems in one agency raised new questions in other agencies, so that objections to the language increased as the draft legislation became longer and more complex.

- Second, the approach failed to appreciate that some large databases -- e.g., Census and tax files -- represented more significant risks and, thus, needed more elaborate confidentiality protection than other files.

During the first Reagan Administration, this prescriptive formula became more and more complex, as attempts were made to incorporate comments from both statistical and nonstatistical agencies. The draft proposal eventually was withdrawn when it became apparent that almost no one could understand how all of the myriad definitions and exceptions fit together.

While the proposed approach did not succeed, the effort did draw attention to many subtle weaknesses in existing law and led to new statutes and amendments during the second Reagan Administration. In particular, stronger statutory protections were enacted for the National Center for Health Statistics, the National Agricultural Statistics Service, and the National Center for Education Statistics. At the same time, the concept of a government-wide law for statistical confidentiality and data sharing received a complete overhaul.

A new strategy was presented to the statistical agencies during the Bush Administration. It had five important features that were missing from earlier efforts:

- It was designed to work with the tools already available in the PRA -- promoting data sharing, but providing for functional separation to ensure that the statistical data are only shared for statistical purposes.

- It was designed to be robust with respect to reorganizations within the statistical system. Since every major statistical agency had been involved in one or more reorganizations since 1970, it became apparent that any successful strategy would have to work well in any reasonable organizational environment.

- It was built around a procedural strategy that gives due deference to the precepts of existing law that are tailored to specific risks and builds on agency experience in implementing that body of law. The idea was to adopt a general confidentiality policy consistent with existing law and provide the tools -- data sharing agreements, coordinated rules, and consistent Freedom of Information Act (FOIA) exemptions -- to address those risks.

- It provided a means for the major statistical agencies to work closely with other agencies in their areas of expertise. While only the Statistical Data Centers would have broad access to data, any agency that collects its own statistical data can act as a full partner in improving those data under the terms of a data sharing agreement.

- It strengthened the Trade Secrets Act. This universal confidentiality statute consolidated provisions of tax law, customs law, and *statistical* law, but the statistical implications had been ignored. The new proposal set uniform policies for *confidential statistical data*, increasing penalties and addressing questions of *agents.*

This fresh start – based on a precedent-setting data sharing order involving the Internal Revenue Service, the Census Bureau, and the Bureau of Labor -- had strong support within the Administration. But the effort failed to reach closure.

The basic strategy developed during the Bush Administration was later expanded and refined during the first term of the Clinton Administration. Criteria for the Statistical Data Centers (SDCs) were incorporated into the Statistical Confidentiality Act, and every statistical agency that could meet these tests was added to the list of SDCs -- bringing the total from four agencies to eight. The relationship to the PRA was fine-tuned, as well, and this process identified some improvements to the PRA that were adopted in the 1995 amendments to that Act.

The final step in the recent initiative involved negotiating a complementary amendment to the *Statistical Use* section of the tax code [26 USC 6103(j)]. This change actually facilitates increased security for taxpayer information, by targeting and, thus, limiting the wholesale disclosures permitted under current law. It permits multi-party sharing agreements, so that specific statistical data sets that include tax data can be shared under IRS security procedures with other SDCs.

What Factors Argue for Success Now?

After more than two decades, why should we think that these efforts will be any more successful than those of the past? Perhaps it comes down to what can be called the "Three E's:"

- *Experience*. -- Over the past 25 years we have learned a considerable amount. The current proposal builds on the experience OMB and the agencies gained through earlier efforts.

- *Environment*.-- The Federal statistical system is faced with growing fiscal resource constraints. At the same time, the 1995 Paperwork Reduction Act extends requirements for reducing burdens imposed on respondents to Federal surveys. Yet another factor that has affected agency views is the increasing number of proposals for consolidating statistical agencies.

- *Enthusiasm*.—Last but not least, the statistical agencies appear to be in a "can do" mood – enthusiastically supporting the development and passage of legislation that will even out statutory confidentiality protections and permit data sharing for statistical purposes.

Whatever the reasons, the agencies have come together on the Administration proposal now embodied in Statistical Confidentiality Act and its companion pieces.

The Statistical Confidentiality Act

As the centerpiece of this effort, the Statistical Confidentiality Act has two principal functions:

- To ensure consistent and uniform application of the *confidentiality privilege*; and

- To permit limited *sharing of data* among designated agencies for exclusively statistical purposes.

A limited number of Federal statistical agencies would be designated as Statistical Data Centers. The eight agencies that currently meet the criteria to become SDCs are the Bureau of Economic Analysis (BEA), Bureau of the Census, Bureau of Labor Statistics (BLS), National Agricultural Statistics Service (NASS), National Center for Education Statistics (NCES), National Center for Health Statistics (NCHS), the Energy End-Use and Integrated Statistics Division of the Energy Information Administration (EIA), and the Science Resources Studies Division of the National Science Foundation (NSF).

A key component of the legislation is *functional separation*, whereby data or information acquired by an agency for *purely statistical purposes* can be used only for statistical purposes and cannot be shared in identifiable form for any other purpose without the informed consent of the respondent. If a designated SDC is authorized by statute to collect data or information for any nonstatistical purposes, such data or information must be distinguished by rule from those data collected for strictly statistical reasons.

The procedural strategy for implementing the legislation would be carried out via written data sharing agreements between or among statistical agencies. The Statistical Data Centers would provide information on actual disclosures and information security to OMB for inclusion in the annual report to Congress on statistical programs. OMB would also review and approve any implementing rules to ensure consistency with the purposes of the SCA and the PRA.

Companion Legislation

In addition to the Statistical Confidentiality Act, special amendments have been proposed to the Statistical Use subsection of the Internal Revenue Code -- Section 6103 (j). These amendments would authorize limited disclosure of tax data to agencies which have been designated as Statistical Disclosure Centers. In addition, the Research and Statistics Division at the Federal Reserve Board has been added to the group of agencies covered under the IRS companion Bill.

The amendment would provide access to tax return information to construct sampling frames and for related statistical purposes as authorized by law. Names, addresses, taxpayer identification numbers, and classifications of other return information in categorical form could be provided for statistical uses. These latter data are not to be used as direct substitutes for statistical program content, but rather can be applied using statistical methods such as imputation to improve the quality of the data. Class sizes or ranges for such data -- e.g., for income -- will vary by purpose.

The amendment is designed to protect taxpayer rights and maintain proper oversight and control over tax return disclosures, while allowing carefully targeted expansion of access to tax return information for statistical purposes only.

The Statistical Confidentiality Order

As an integral step to foster passage of these legislative proposals, OMB felt it was critical to move ahead with efforts to clarify and make consistent government policy protecting the privacy and confidentiality interests of individuals and organizations that provide data for Federal statistical programs. With that aim in mind, OMB developed and sought public comment on an Order that assures respondents who supply statistical information that their responses will be held in confidence and will not be used against them in any government

action. The Order also gives additional weight and stature to policies that statistical agencies have pursued for decades and includes procedures to resolve a number of ambiguities in existing law. Following the public review process, the Federal Statistical Confidentiality Order went into effect on June 27, 1997.

What Opportunities Will Attend Passage of the Legislation?

For more than a decade, we have worked within the constraints of existing law to make limited comparisons between similar data sets in different agencies. We have set in motion a series of limited exchanges tailored to conform to current law, but they cannot address all of the problems. Moreover, such exchanges could be cut short by an unfavorable interpretation of any one of the dozens of statutes involved. In each of these cases, extraordinary efforts have been required to accomplish even limited data exchanges. Based on these experiences, we believe that even modest exchanges of information could, in the future, unearth and eliminate important errors in existing economic series, enable significant consolidations of overlapping programs (with comparable reductions in costs), and permit substantial reductions in reporting burden imposed on the public.

As the possibility of a law to permit data sharing in a safe environment has become more credible, statistical agencies have begun to identify potential improvements to current operations and programs that this law would permit. These include possibilities such as the following:

- Integrated database concepts for information on particular segments of the economy and society, such as educational institutions (NCES, NSF, and Census), health care providers (NCHS, Census, and some program-specific agencies), and agricultural establishments (NASS, Census, and the Economic Research Service at the Department of Agriculture), would improve the consistency and quality of data while reducing current data collection costs.

- Collaboration on sampling frames would improve accuracy and reduce maintenance costs. A more efficient division of labor would make it possible to maintain high quality frames at minimum cost, both for list frames (Census, BLS, NASS) and for area frames (NASS, Census, NCHS). This approach would avoid duplicate expenditures and improve quality. Coordination and shared use of relisting information (updates) in large multi-stage designs could also reduce frame maintenance costs.

- Targeted frames – or sample selection services – from improved master frames could reduce duplicative expenditures in agencies that must currently pay the cost of independently developing these resources for specific surveys.

- Access to specific data details that can resolve uncertainties in particular analyses – e.g., anomalies that arise in the Gross Domestic Product estimation process – would reduce errors in macroeconomic statistics without imposing additional burden.

- Coordination of sample selection across agencies could reduce the total reporting burden that falls on any one household or company (and, thus, improve the level of respondent cooperation).

What Systemic Problems Will the Act Address?

- *The Statistical Confidentiality Act creates a credible government-wide confidentiality umbrella*.-- The public will know that the entire government stands behind the pledges of statistical confidentiality offered by the SDCs or any agency engaged in joint statistical projects with the SDCs.

- *The SCA creates the legal presumption that data collected for most purposes may be used in a safe environment for statistical purposes.*-- This is one of the critical insights of the Privacy and Paperwork Commissions.

- *The SCA provides consistent FOIA policies for all the SDCs.*-- This was controversial 15 years ago, but now six of the eight agencies designated as SDCs already have in place statutes that meet the requirements of Section (b)(3) of FOIA.

- *The SCA permits the data sharing authorities of the PRA to work without compromising confidentiality*.-- By establishing the functional separation principle in law, the SCA facilitates the use of PRA mechanisms to promote and manage data sharing for exclusively statistical uses.

- *The SCA provides a privacy-sensitive alternative to the creation of universal databases, which each Department has proposed at one time or another to support its own policy interests.*-- Statistical methods -- particularly sampling -- coupled with secure data sharing provide a natural hedge against the big database (i.e., dossier building) mentality that puts privacy at risk.

In short, the Statistical Confidentiality Act permits the SDCs and their statistical partners to share both expertise and data resources to improve the quality and reduce the burden of statistical programs, while preserving privacy. Moreover, no matter how the organizational boxes for the ideal Federal statistical system are drawn, this legislatin will permit the components of the statistical system to manage their data as if they were a single, functionally-integrated organization.

Current Status of the SCA and Related Initiatives

Culminating efforts that literally have spanned decades, the Statistical Confidentiality Act initially was introduced on a bipartisan basis in the House of Representatives in 1996. Late in 1997, the Administration's proposed legislation was included in a broader bill, S. 1404, introduced on a bipartisan basis in the Senate. With growing bipartisan support in both houses, hopes are high that the SCA will soon become law. The complementary amendment to the Internal Revenue Code is also pending before Congress, with broad bipartisan support. OMB is working with the House and Senate to attain re-introduction and successful action on the legislation during 1998.

In addition to these legislative approaches to foster efficiency and quality in Federal statistical programs, the agencies are actively exploring other means of expanding collaboration to improve the effectiveness of the Federal statistical system. Recently the Interagency Council on Statistical Policy (ICSP), under the leadership of the Office of Management and Budget, has broadened efforts of the principal Federal statistical agencies to coordinate statistical work -- particularly in areas where activities and issues overlap and/or cut across agencies. One by-product of these efforts was the establishment in 1997 of the Interagency Confidentiality and Disclosure Avoidance Group, under the auspices of OMB's Federal Committee on Statistical Methodology. This working group discusses common technical issues involving privacy, confidentiality, and disclosure limitation. The group is currently working on developing a set of generic guidelines for disclosure review, which could be adapted for use by other agencies.

It is our hope and expectation that both the statistical confidentiality legislation and the subsequent cooperative efforts will go a long way towards solving some of the challenges the Federal statistical agencies have encountered in a decentralized environment.

Chapter 9

Linking Records on Federal Housing Administration Single-Family Mortgages

Thomas N. Herzog and William J. Eilerman
U. S. Department of Housing and Urban Development

Abstract

Over the years, we have developed a number of ad hoc record linkage procedures to correct serious data problems on the Federal Housing Administration's (FHA) primary database of single-family mortgage records. This work describes a number of the procedures used and illustrates the results of these efforts. One effort resulted in the identification of thousands of duplicate mortgage records. The subsequent deletion of these duplicate records from the database saved FHA several million dollars. A second effort resulted in the identification of thousands of mortgage records on terminated loans which the database erroneously indicated were active mortgages. This effort enabled FHA to more accurately predict its future premium income as well as to improve other analytic studies of these Federal mortgage insurance programs.

Chapter 9

Using Microsoft Access to Perform Exact Record Linkages

Scott Meyer, Statistics Canada

Abstract

The author describes how using Microsoft Access to perform record linkage may be a viable alternative to specially designed record linkage software for certain applications. Access was pursued since it is fairly easy to use, flexible, interactive, and reasonably fast when performing simple queries. The major drawbacks and minor difficulties will also be discussed. Examples will be drawn from a project which involved linking court records to police records for selected Canadian cities.

A description of the project to link the data from the court and police surveys will be given. The motivation for beginning the linkage and the long term goals will be discussed. The history of the project will be briefly reviewed. The author will then focus on Access, and how it is considered to be an effective alternative to methods previously used for this project. The advantages and disadvantages will be presented.

The strengths of Access include: flexibility -- the criteria which must be met for the records to be considered matches are fully controlled and easily altered by the user, plus it is simple to select subsets of large files which can then be easily explored; availability -- Access, being part of Microsoft Office Suite, is available to many users; speed -- for our application the queries took very little time to run, making the session highly interactive; ease of use -- Access is easy to learn, and even fairly complicated queries can be done with only "point and click" actions with no knowledge of how to program in SQL required.

The primary disadvantage is that there is no probabilistic matching based on the theory of Fellegi and Sunter (1969). This is a significant drawback; however, for many applications, exact matching is enough to meet the project's goals.

Lastly, some results of linking court and police records using Access will be presented.

Introduction

The goal of this project is to combine court data from the Adult Criminal Court Survey (ACCS) which collects provincial court data and the Revised Uniform Crime Reporting Survey (UCR2) which collects police data on criminal incidents. The populations of the two surveys overlap a great deal. By linking the two files, we can map an offender's movement through the criminal justice system from the point of arrest to the point of sentencing. The ACCS provides data about the decision (guilty or not guilty) plus full sentencing information. The UCR2 survey provides details about the criminal incident, the accused, and, for violent offences, the victim(s). It is anticipated that linked data will provide answers to some interesting questions asked in the justice community. For example, is there a difference in the types and lengths of sentences for accused charged in spousal versus non-spousal assaults? Does the location of a break and enter or act of vandalism affect the severity of the sentence?

This report will show that using Access has the potential to be an effective method to combine data. A detailed explanation of how the linked data sets were created using Access queries is not included here, instead linkage results and discussion of some of the advantages and disadvantages of using Access are presented. The first two sections provide a description of the preprocessing of the survey data. This is followed by some statistics obtained for the city of Regina and some explanation of possible reasons for nonmatches in this study. The disadvantages and advantages of using Access are then presented, and a short summary and some conclusions are given in the final section.

Data Sources

Police and court data from the city of Regina was used in this study. Specifically, ACCS charges which had a date of offence between July 1, 1993 and December 31, 1993, were loaded into an Access table. Similarly, UCR2 records from the Regina police department which had a report date in the same six months were loaded into an Access table. Most of these police records also had a date of offence between July 1, 1993 and December 31, 1993, but there were incidents where the date of offence was many years prior. Although these UCR2 records will likely not match, they do not distort the linkage rate since this calculation is based on the percentage of ACCS records for which a match to a police record is found. Regina was chosen since the coverage of the two surveys currently overlaps only in Quebec and Saskatchewan. All previous record linkage studies have been done using only Quebec data.

Preprocessing of the ACCS and UCR2 Data

Before the linkage could be performed, some preprocessing of both raw data files was done. The goal was that each charge on the court file would link to its corresponding violation on the police file. This is a significant change from prior linkage attempts where many charges were linked internally or "rolled-up" into one ACCS record (Brown, 1995; Cooley, 1996).

The raw ACCS data comes from every courthouse in the city and there is one record for every appearance. Since the unit of count used in ACCS published tables is the charge, a program was available which converted the raw data into its one record per charge equivalent. This file with one record per charge was loaded into Access.

The raw UCR2 data is reported by the municipal police department of each city and consists of three files, the Incident, Accused, and Victim files. All three files contain two variables which uniquely identify any incident, the *respondent code* and *incident file number*. These codes allow the files to be easily joined. The Incident file contains information about the criminal incident including the location, date and time of offence, value of property stolen, etc. The Accused file contains a record for each accused that has been identified. Variables such as date of birth, sex, and ethnic status (aboriginal or non-aboriginal) appear on the Accused record. Similarly, the Victim file contains information about each victim of a violent offence. Variables appearing on the Victim file include level of injury, age, and relationship to the accused. An incident may involve more than one violation (up to four are captured on a single UCR2 Incident record). For example, one UCR2 incident could involve both theft and mischief violations. Also, there may be more than one accused involved in a single incident, and for violent offences, a single incident could involve multiple victims.

In order to make the UCR2 data more compatible with the ACCS data structure, a violation file was created. This file had one record for each accused/violation. For example, if an incident involved two accused and three violations, there would be six violation records created, one for every accused/violation combination. This UCR2 preprocessing made the linkage between ACCS charges and UCR2 violations more straightforward than in previous attempts. The UCR2 violation file was loaded into Access along with

the Victim file. The three Access tables (the ACCS charge table, the UCR2 violation table, and the UCR2 victim table) together form an Access database.

In addition to creating ACCS charge and UCR2 violation and victim files, derived fields were added. The most important of these fields was the Common Offence Classification (COC). The COC consists of 28 codes which represent broad categories of crime. The three digit ACCS offence code and the four digit UCR2 violation code were each mapped to their corresponding COC codes. For example, a COC of "1" represents homicide and related offences; a COC of "2" represents attempted murder; and "3" represents robbery. Previously, the offence and violation codes were not used in the linking strategy. By incorporating the COC into the linkage procedure, there is higher confidence that the court record and police record relate to the same event. This minimizes the incidence of the "false-positive" matches encountered in earlier work.

Within Access, the variables available on both files which were used for linkage are Soundex (encrypted name), date of birth, date of offence, sex and COC. The premise of Soundex coding is that names which sound alike (regardless of spelling) are assigned the same code consisting of one letter followed by three numbers. The Soundex codes are created when the survey records are extracted from the local databases, therefore, the full names of the accused are not available from either survey. Records which link exactly on all five of these variables are deemed to represent the same criminal violation and subsequent court charge.

Results for Regina

For Regina, the overall match rate based on an exact match for all five variables was 58%. This is calculated from 3105 of the 5360 charge records from the court data linking to violation records from police data. As a second step, constraints could have been relaxed and another link using only unmatched records from the first step could have been done. However, the goal was not to create one large linked file but rather to produce Access tables which could be used to link records for specific research questions.

The following example illustrates the use of Access to examine one specific problem concerning assaults. There is a potential linking problem because common assault and major assault have different COC codes. Table 1 shows that when using the restriction that the COC codes must agree exactly, 62% of the 442 assault records on the ACCS table were linked. By allowing major assaults to be linked to common assaults, and vice versa, an additional 10% of the records were successfully linked. These are likely to be true matches and the match rate for Regina assaults was increased to 72%. Further steps were then taken to expand the linked file by allowing other constraints placed on the linking variables to vary. For instance, allowing some range for the date of offence, rather than matching exactly, brought together records that, in all likelihood, refer to the same UCR2 violation and ACCS charge.

Table 1 reveals that a match rate of around 85% was achieved while still maintaining high confidence in the quality of the links. Confidence in the quality of the matches declines as more differences among the linking variables are allowed. Judgement of the analyst is important in deciding whether to increase the size of the linked Access table at the risk of allowing "false positive" errors to be made. Table 1 shows the results of following one particular linking strategy for Regina data. Other equally effective strategies may be used, and an analyst is free to reorder the steps, add or omit steps, or decide how much relaxation of matching constraints is appropriate within any particular step (e.g., using 10 days instead of 35 for the date of offence range). The method used to create the analytical Access table will depend on the application and on the input data being used. For example, if some linking variables from certain jurisdictions are known to have data quality problems, then constraints on these variables can be relaxed at an early stage.

Table 1. -- Linkage Rates Using Various Strategies -- Regina Assault Charges (N = 442)

Link #	Soundex	Date of Birth	Date of Offence	COC	Sex	# of new matches	Cumulative # of matches	Cumulative match rate
1	exact	exact	exact	exact	exact	276	276	62%
2	exact	exact	exact	same family[1]	exact	42	318	72%
3	exact	exact	within 35 days	same family	exact	32	350	79%
4	exact	close[2]	exact	same family	exact	11	361	82%
5	exact	close	within 35 days	same family	exact	3	364	82%
6	close[3]	exact	exact	same family	exact	11	375	85%
7	exact	exact	exact	same family	disagree	2	377	85%

[1]Same family = there was no distinction between major and common assaults. Links to UCR2 violations of sexual assault were also permitted, but there were only two matches of this type.
[2]Close for Date of Birth = agreed on two of year, month, and day, or had a month/day reversal.
[3]Close for Soundex = agreed on first letter and first digit of Soundex code.

Possible Reasons for Unlinked ACCS Records

Court and police records may not link for two distinct reasons. There may be no corresponding police record for the charge or there exists a record, but it can not be matched to the charge based on the linking variables and strategy.

There are several possible reasons why no corresponding police record exists. First, there are problems of geographical (jurisdictional) coverage. For example, persons who were charged by the Royal Canadian Mounted Police (RCMP) will have no UCR2 record because the RCMP does not currently report to the UCR2 survey. It is estimated from UCR aggregate data that the proportion of charges laid by the RCMP for Regina is around 5%.

Another aspect of coverage difficulties is charges pertaining to offences which are court related and may not involve the police at all, for example, offences against the administration of justice. Charges for these offences often have no corresponding police record.

There is also the possibility that the police record would be located in another city or province. Since the database includes only the records for one city, the ACCS charge record would remain unmatched. In future, databases which include records from larger areas, such as provinces or regions, could be produced and these would provide the opportunity to link records for an individual who offends in one city and is tried in another.

A fourth reason is that a UCR2 record exists, but due to the restrictions put on the report date when

preparing the Access table, it was excluded. This will be investigated further, and some adjustments to the table preparation method may be required.

The reasons for failing to find true matches when both records exist are harder to describe. Name changes, keying errors on Soundex (incorrect first letter), and missing data are examples of data quality problems on the source files that can result in nonmatches.

Microsoft Access as a Record Linkage Tool

While this report shows that Access can be an effective tool, there are, as with any software, some problems or difficulties. Obstacles and drawbacks encountered in this study will be discussed first, followed by a summary of some advantages of using Access.

These are some difficulties with using Access:

- Although Access does allow for some inexact matches, there is no real probabilistic matching based on weighting. It is possible to use weights when doing exact matching, however, assigning weights in Access is difficult, and this is a major drawback. Probabilistic matching based on the theory of Fellegi and Sunter (1969) is possible with Statistics Canada's GRLS system (Felx, 1995). Further, GRLS and other record linkage software allow the use of sophisticated comparison rules (e.g., string comparator metrics), which would be very difficult, if not impossible, to imitate using Access.

- When the linkage is performed, duplication can occur. If two UCR2 violation records have exactly the same values for all matching variables, they would both be linked to the same charge record. In a sense, the charge record has been duplicated. This duplication is not a problem when simply counting the number of successful matches, or when the UCR2 records are very similar. The difficulty occurs when the UCR2 records differ in important ways. For instance, an analyst is interested in comparing sentencing for break and enters (B & E) committed against businesses to sentencing for break and enters committed against personal homes. If one break and enter charge links to two different UCR2 violations, and one violation is against a business and the other is against a home, then the charge is difficult to classify. Should the sentence length be used in the mean sentence length calculations for business B & E, residential B & E, both, or neither? The analyst must be aware of this possibility, and, when doing analysis, these ambiguous records may have to be excluded or handled in some other fashion.

- Access does have some mathematical functions (sum, average, max/min value, etc.), but to do more sophisticated statistical analysis with the linked data set, it would have to be exported to a statistical software package.

- Though Access is easy to use, careful attention to detail is required. Queries with seemingly small differences can produce vastly different results. Careful design of queries is needed to ensure that the final result is what was intended. Novice users, not knowing what kind of output to expect, may not immediately recognize flawed queries. Also, depending on the linking strategy used, a relatively long sequence of steps may be involved. Though each step is fairly easy to perform, the entire procedure can become quite complicated.

- The study used Access 2.0 for Windows and there are some important technical limitations. The speed, and hence the convenience, of using Access is affected by the power of the PC that it is running on. Some important Access limitations are listed here. The maximum database size is 1 Gigabyte; the maximum number of tables plus queries in the database is 32,768; the maximum number of fields per record/table is 255; the maximum number of tables used in a query is 32; the

maximum number of sorted fields per query is 10. In the Regina study these maximum capabilities were not generally restrictive. One problem encountered was by continually using the output from one query as the input to the next, after several layers of depth, the error message "query is too complex" would appear. This is avoided by saving the output from an intermediate step as a table, then using this newly created table, rather than the output from the query, as the input to subsequent queries. MS Access Version 7.0, which is now available, may have greater capacities. For large applications, MS SQL server could be adopted as the underlying relational database management system, while the user interface would still be MS Access.

These obstacles are not terribly severe. The advantages of Access, which are listed below, outweigh the problems or difficulties.

- The greatest advantage of using Access is the flexibility. As mentioned, the criteria which must be met for the records to be considered matches is fully controlled and easily altered by the analyst.

- Also, there is flexibility when creating the linked analytical file with respect to which variables are included. Since only the selected variables will be written to the linked table, the analyst is able to work with an uncluttered data set. In addition, the analysis of nonmatched records from any Access table is very easy. A simple built-in query wizard will provide the analyst with the unmatched records. Patterns among the unmatched records may be discovered by reviewing them visually or via subsequent queries on the unmatched data set. For instance, match rates may, for some reason, be lower for certain courtrooms within the city. If a situation like this is discovered, it can be further investigated.

- Another asset of Access is its availability. In particular, it is available to analysts in the Canadian Centre for Justice Statistics (CCJS), and generally, it is a component of the ubiquitous MS Office Suite.

- Another benefit of using Access is its speed. How quickly a query runs depends on the computer's hardware, the size of the tables which are being queried, and the complexity of the query. Using a pentium computer, queries on the Regina database took only a few seconds to complete. The result is a highly interactive session where one can quickly learn about the data while creating the linked table.

- Since it runs in the PC environment, Access is inexpensive to use. The only cost is an up-front cost of purchasing the software/software licences.

- Another asset of Access is its ability to use data from and provide data to a number of sources (e.g., spreadsheets, other database software, flat file, etc.). Since the preprocessing was done using SAS, and further analysis requiring sophisticated statistical procedures may be done, it is important that Access be able to import and export the files. Indeed, the import and export capabilities of Access are quite good, thus lending compatibility with other packages.

- Lastly, Access is easy to learn and use and requires no special programming skills to use effectively. Table 1 shows the results of a sequence of queries. This was done to show how relaxing the constraints can increase the number of matches. In practice, the analyst would not usually run several follow-up queries. It is more likely that a single complex query which achieves much the same result would be run. The drawback of a single secondary query which follows the exact match is that for the added records, it is not immediately obvious why they failed to match on the first attempt. For example, records which did not match exactly on the COC and records which did not match exactly on date of offence would be added at the same stage, and it would not be obvious

how many records were in each group. A complicated query which would allow inexact matching on any one of date of birth, date of offence, COC, sex, or Soundex needs to be prepared only once, after which it can be re-used (if some consistent table naming convention is used). In this way, the analysts who are new to using Access and not confident about preparing their own queries can still perform an effective record linkage using these pre-written queries.

In summary, there are both positive and negative aspects to using MS Access to link ACCS and UCR2 records. Weighing these various considerations, Access appears to be a viable and practical way to link records, and meets the goals of this project.

Conclusions

The preliminary work using Access to perform the record linkage is very encouraging. This report focuses on one application, linking adult criminal court records to police records, but Access could also be used for other CCJS record linkage projects. Some possibilities are: youth court to youth corrections (YCS-YCCS), youth court to police (YCS-UCR2), and the marriage of these, police to youth court to youth corrections (UCR2-YCS-YCCS). The match rates achieved for the UCR2-ACCS linkage in Regina were similar to previous studies, but the interpretation of the resulting file is easier. This is an advancement in the record linkage work done in the past three years, since for the first time meaningful analysis of the linked file seems possible.

References

Brown, C. (1995). *Record Linkage Feasibility Study: Uniform Crime Reporting Survey/Adult Criminal Court Survey*, Internal Statistics Canada Report.

Cooley, D. (1996). *Record Linkage Feasibility Study: UCR2/ACCS - Part II*, Internal Statistics Canada Report.

Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, 64, 1183-1210.

Felx, P. (1995). *Feasibility of Using CANLINK for Linkage: An Application in the Canadian Centre for Justice Statistics*, Internal Statistics Canada Report.

Chapter 9

Analysis of Immigration Data: 1980-1994

Adam Probert, Robert Semenciw, and Yang Mao, Health Canada
Jane F. Gentleman, Statistics Canada

Abstract

This paper describes the record linkages being carried out at Statistics Canada to link data for all immigrations to Canada from 1980 through 1994 to income tax files (for live follow-up) and then to data for all deaths in Canada. The files involved are very large. As an example, the number of immigrants in 1980 was 143,432, and the number in 1994 was 222,538.

Numerous studies of immigrants have been published around the world, but the vast majority of them are missing information on the entry date to the country. Our immigration data will not have this limitation. They will be used to follow up immigrants to see how living in Canada has impacted their health and how this is affected by the length of time they have lived in Canada. This project will study how cause-specific mortality varies with country of origin and length of residence in Canada, to aid in disease control and prevention. The study of disease patterns in persons from different geographical areas is an epidemiological technique that can provide important clues to the causes of disease. Such studies can show the potential for preventive actions if a risk pattern from one population can be transposed to another.

Introduction

Record linkages are presently being carried out at Statistics Canada to link data for all immigrants to Canada from 1980 through 1994 to income tax data (for live follow-up) and then (for the earliest years) to mortality data. This paper is a description of the data bases and the rationale for the project.

This project will study how cause-specific mortality varies with country of origin and length of residence in Canada, to aid in disease control and prevention. The study of disease patterns in persons from different geographical areas is an epidemiological technique that can provide important clues to the causes of disease. Such studies can show the potential for preventive actions if a risk pattern from one population can be transferred to another. One of the earliest of these studies involved Japanese migrants to Hawaii. From the differences in stomach cancer rates among the immigrant and native populations the researchers were able to implicate diet as a risk factor for stomach cancer.

Immigration Data

Numerous studies of immigrants have been published around the world, but the vast majority of them are missing information on the entry date to the country. Without this information, the amount of exposure to life in the new country is unknown, so an "exposure-response" relationship cannot be studied. Our immigration data, with the landing date, will not have this problem.

Immigrants comprise a large proportion of the Canadian population. For example, the number of immigrants in 1980 was 143,432, and the number in 1994 was 222,538. According to the 1991 Census,

there were approximately four million immigrants in Canada, or 16% of the population. The health status of such a large segment of the population should be investigated.

The Immigration Data Base has existed in machine-readable form since 1980. It contains information on every landed immigrant to Canada, as of the actual date of landing. In contains data on education level, intended occupation, medical class (a summary variable providing a baseline medical status), language ability and, of course, name, sex and date of birth. It also contains a couple of unique identifiers: visa number, which is unique for every landed immigrant, and family identification number, which is given to all members of a family who immigrate on the same date. The database, for the most part, is complete. The least complete variable for 1980 immigrants is date of birth, which is missing in approximately 1,000 out of 143,476 records (0.7%).

The present study will use probabilistic record linkage to the Canadian Mortality Data Base (CMDB), maintained at Statistics Canada, to link to almost three million immigrant records. If this linkage proves successful, then future linkages to the cancer incidence and tuberculosis data bases will be considered. The linked data will be used to follow up immigrants to see how living in Canada has impacted their health and how this is affected by the length of time they have lived in Canada.

Preliminary analysis will be performed on the immigration data before the linkage to the mortality data. Trends in immigration over the 15-year period will be examined. Specifically, the number of immigrants by country of birth, age, sex, education, medical class and intended occupation will be described over the 15 years. The analyses to be performed on the linked data involve Poisson or logistic regression models of outcome (mortality, cancer or tuberculosis) and exposure variables (length of time in Canada, age, country of birth, medical class at arrival, etc.).

Data Limitations

Three challenges have to be dealt with in analyzing these data. Studies of immigrants must deal with what is termed re-migration, i.e., immigration followed by emigration. Of all the landed immigrants to Canada, about 30% will later emigrate, most likely to the United States or return to their home country. The second challenge pertains to immigrants who will not link to the mortality database, that is, who have not died in Canada. If they do not link to the mortality database then it is not known whether they are alive in Canada or deceased in another country. Based on the initial data, one would not be able to estimate the time spent in Canada if there was no record of death. A third issue concerns female immigrants. For those who have changed their name through a change in marital status, it may prove to be extremely difficult to find a match to the mortality database. Whether or not data for both sexes can be analyzed, the 1980 immigration data are expected to yield the most useful results, as the follow-up period during which mortality could occur is longest for this group.

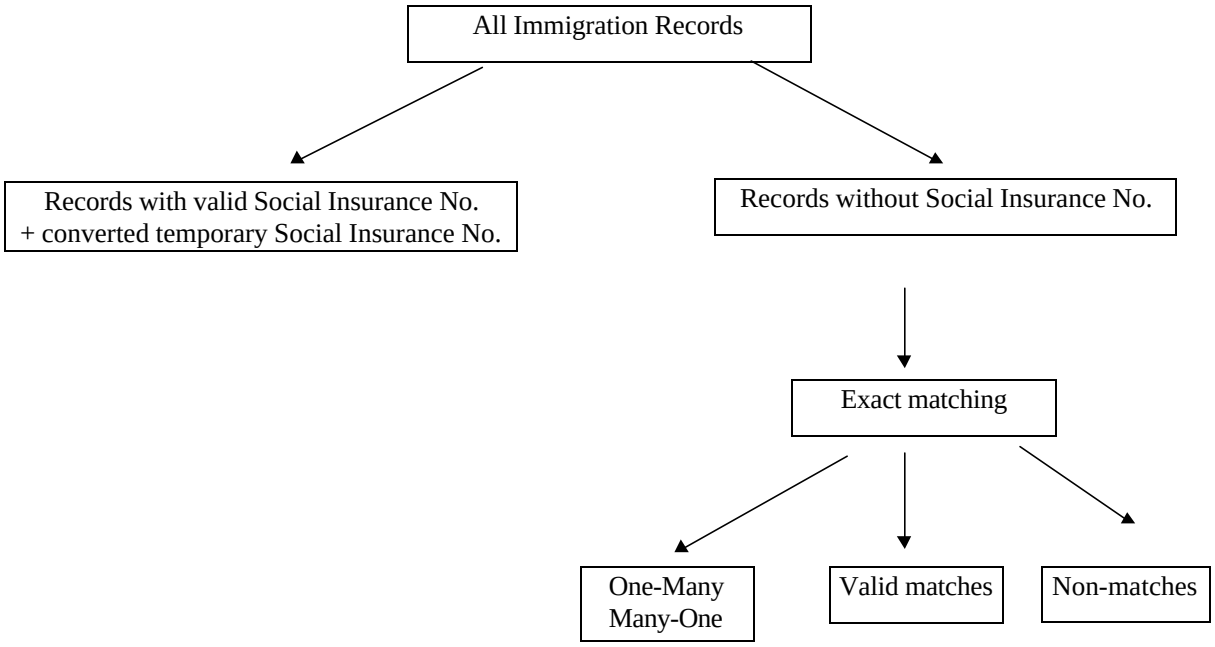
Record Linkage

To address these problems a record linkage to the income tax files was suggested, not necessarily to obtain tax information. The main purpose of linkage to tax files is live follow-up, i.e., accumulation of evidence that the immigrant remained in Canada. This is critical information because a significant number of immigrants leave Canada subsequent to immigration, as mentioned previously. An extremely useful by-product of this linkage will be the date of death for those immigrants who died since 1980; this will facilitate the second stage of linkage -- to mortality data.

As the first stage of linkage, the immigration data for each year have been linked to all income tax files for 1980-1994 (including tax files for years before immigration to Canada, because it is possible to file before immigrating). Of all the 3 million immigration records, only half had a valid Social Insurance Number

(SIN). This included those who had a temporary SIN, which was later converted to a permanent number. The SIN is the Canadian counterpart of the American Social Security Number. For each immigrant, exact matching was used to find that person on the tax file for any year, based on surname, first four characters of given name, date of birth (year, month, day), and sex. Once the immigrant was found on any tax file for any year, the SIN was known and could be used to find the same person on tax files for other years. As part of the regular income tax form, immigration date, emigration date and date of death all appear in addition to the regular tax information, if applicable. See Figure 1 for a diagram of the tax linkage procedure.

Figure 1. — Income Tax Linkage Procedure



At the stage of exact matching, duplicates are created. For those records where there is a prefix to the surname (e.g., De La or Von), there is a duplicate record created (and flagged) for the surname without the prefix. From this procedure there were approximately 1,000 records that yielded many-to-one or one-to-many linkages; these were ignored for this linkage. All records with a SIN are then linked using that SIN to all the income tax files. It is at this step where the linear record of a landed immigrant's stay in Canada will be found. Regardless of the outcome of the income tax linkage procedure, all immigration records will be incorporated in the mortality linkage.

Results

Table 1 contains some of the results of the tax linkage. Examining the 1980 data, one can see that there were 143,432 landed immigrants, of which 44, 486 linked to the 1980 tax files and 67,782 linked to the 1994 files. Note that the tax files mostly contain data for filers aged 15-65 (about 100,000 out of 143,432 1980 immigrants). In other words, about 68% of those who landed in 1980 filed a tax return in 1994. Also, 152 people who filed a tax return in 1980 did not become landed immigrants until 1994. It is possible, for example, for people present in Canada on business or student visas to pay tax before becoming landed immigrants.

Table 1. — Preliminary Results from Income Tax Linkage

Landing Year	Number of Immigrants	Found 1980 Tax Form	Found 1994 Tax Form
1980	143,432	44,486	67,782
1981	128,735	4,648	62,956
1982	121,253	2,776	60,771
…			
1993	255,087	222	118,795
1994	222,538	152	86,943

Next Steps

Initially, only the 1980 immigration-tax data (1980 immigrant files linked to 1980-1994 tax files) will be linked to the CMDB using the commercially available Automatch linkage software. The CMDB is a record of all deaths since 1950. The database is mostly complete, with coverage varying between 98% and 100% for most variables. The completeness differs over time and among provinces. Linkage to the CMDB will be done using probabilistic methods. The variables Surname, Given name(s), Date of Birth, Sex and Other Name will be used for the linkage. Marital status and Country of Birth may also be used depending on the success of the previous pass. All of the names will be converted to NYSIIS format to aid in the name-matching process. With foreign names there may be more spelling/typographical errors that NYSIIS coding can alleviate. Reversing of the name and birthdate fields will be allowed to control for those errors where the first and last name or day of birth and month are switched.

As mentioned previously, linkage problems may be encountered for females because of name changes subsequent to immigration and because of the decreased propensity of females to file tax forms. To increase the chances of finding females on the mortality database, all surnames of females (maiden names and married names) found on the income tax records will be captured and used in the death linkage.

Analyzing the output from the linkage will involve many steps. First, we will examine the risk of disease in immigrants compared to their country of birth, controlling for age and sex and examining trends over time. Second, we will examine the Canadian rates of disease for Canadian-born persons. Third, unique to analyses of data of this nature, risk of death by disease and by duration of residence in Canada as well as age at migration will be analyzed. This analysis will also examine the differences by country of birth, occupation, education and other factors.

Conclusion

To summarize, the immigration database offers the opportunity for new research into immigrant health. By linking to Canadian income tax records, we could know when a landed immigrant is no longer a resident of Canada, something that is not available in most immigrant studies. We should also be able to account for some name changes that occur in female immigrants. From the linkage to the mortality database, we will be able to examine the risk of death by country of birth, length of stay, age, sex, education and other demographic variables. All of these analyses will aid in identifying trends and the etiology of specific diseases.

Chapter 10

Record Linkage Methods Applied to Health Related Administrative Data Sets Containing Racial and Ethnic Descriptors

Selma C. Kunitz, Clara Lee, and Rene C. Kozloff, Kunitz and Associates, Inc.
Harvey Schwartz, Agency for Health Care Policy and Research

Abstract

In response to the lack of easily retrievable clinical data to address health services and medical effectiveness questions, especially as they relate to racial/ethnic minorities, the Center for Information Technology (CIT), Agency for Health Care Policy and Research (AHCPR) recently sponsored a project on record linkage methodology applied to automated medical administrative datasets containing racial and ethnic identifiers (Contract 282-94-2005). The primary objectives of the project were to:

> *link patient-level related datasets that contain racial and ethnics descriptors;*
> *and assess the value of the linked data to address medical effectiveness research questions that focus on the quality, effectiveness, and outcomes from care for minority populations.*

KAI, AHCPR's contractor, received approval from the State of New York's Department of Health to utilize the Statewide Planning and Research Cooperative System (SPARCS) files Discharge Data Abstract (DDA) and Uniform Billing files (UBF), which contain all acute hospital discharge and claims data, the SPARCS Ambulatory Surgical files, and the Cardiac Surgery Reporting System (CSRS) files, a research dataset. KAI received files for the 1991, 1992, and 1993 time periods. The files were successfully linked by patient, "visits" across the time periods. While the linked data appear to be of high quality, the process of obtaining and linking the data is lengthy. Additionally, these administrative health care data sets contain millions of records that document all hospital stays and thus, identifying appropriate subpopulations for a particular research question is a time and resource-consuming effort.

While the administrative health care datasets may be useful in answering questions about charges, length of stay, and other health service issues, their current utility may be less useful in answering clinical questions for minority populations. These datasets can be used to explore potential associations among diagnoses, treatment, and outcome variables. However, understanding the mediating factors and the decision-making variables that result in patient care may not be possible. For example, the results of diagnostic tests such as angiograms are not generally recorded in these datasets, thus limiting the ability to carefully subgroup patients by disease severity. With consideration for the potential utility of these datasets, however, there are several recommendations that emanate from the study.

This talk will briefly describe the research questions posed, linkage process, findings, and recommendations for additional action and policy considerations.

Introduction

The ability to link automated health data records is of critical importance in our rapidly changing health care system. In a managed care and cost containment environment, researchers require reliable and valid data collected over time and across providers that describe patient characteristics and the location, process, cost, quality, and outcome of care to analyze which procedures are effective and produce satisfactory patient outcomes. Approaches and methods to linking records across time and providers are needed to provide information to policy makers, health plans, practitioners, consumers, and patients to make decisions about accessing, using, and paying for care, as well as the effectiveness of that care.

Background

In response to the lack of easily retrievable clinical data to answer medical effectiveness questions, especially as they relate to racial/ethnic minorities, the Agency for Health Care Policy and Research (AHCPR) sponsored a project on "Record Linkage Methodology Applied to Linking Automated Data Bases Containing Racial and Ethnic Identifiers to Medical Administrative Data Bases" under AHCPR contract number 282-94-2005 (Kunitz and Associates, Inc., 1996). This linkage demonstration project contributed to AHCPR's research goals by reviewing and adding to record linkage methodology; illustrating the value of this methodology; assessing the need for further development; and providing guiding principles to developers. The primary objectives of this record linkage methodology project were to: link two patient- level related data sets that contain racial and ethnic descriptors; and assess the value of the linked data to address medical effectiveness research questions that focus on the quality, effectiveness, and outcomes from care for minority populations.

Data Sets

AHCPR's contractor, KAI, a health research firm, identified data sets to use for assessing the value of linking administrative health related data bases to support medical effectiveness research in minority populations. KAI received approval from New York State's Department of Health (NYSDOH) to utilize the Statewide Planning and Research Cooperative System (SPARCS) files Discharge Data Abstract (DDA) and Uniform Billing files (UBF), the SPARCS Ambulatory Surgery files, and the Cardiac Surgery Reporting System (CSRS) files. KAI received files for 1991, 1992, and 1993. The selected systems and data files are briefly described as follows:

- **SPARCS (State Wide Planning and Research Cooperative System)** is a system maintained by the NYSDOH. The Discharge Data Abstract files (DDA) contain all acute hospital discharge data and the Uniform Billing Files (UBF) contain all acute hospital billing records. Data about surgeries performed at hospital-based ambulatory care centers and certified diagnostic and treatment free-standing centers are maintained in the Ambulatory Surgery files. The data are used for planning and research. Three of the files extracted from SPARCS for this project were the DDA, UBF, and the Ambulatory Surgery file.

  NYSDOH staff combined the acute hospital DDA and UBF data files by individual hospital stay for this project. Thus, we received both matched and unmatched records from the DDA and UBF for 1991 - 1993. Because the files were selected based on DDA variables, unmatched records are those that are in the DDA file but do not have a corresponding match in the UBF. A completeness level of 95% is typically achieved in SPARCS files, a figure that is supported by our research, as seen in Figure 1. Yet, those records which are unmatched may reflect not only missing UBF records, but also incorrect information which may have hindered the original matching process performed by the

NYSDOH.

Figure 1. -- Matching Rates Between DDA and UBF Records

Year	Total	Matched	Unmatched	
			N	%
1991	626,222	594,302	31,290	5%
1992	699,246	663,323	35,923	5%
1993	714,583	677,778	36,805	5%

KAI obtained 31,290 unmatched records out of a total of 626,222 for 1991, 35,923 unmatched out of a total of 699,246 for 1992, and 36,805 unmatched records out of 714,583 for 1993. The selection process did not enable KAI to receive data which was in the UBF file but absent in DDA. In addition, we received the Ambulatory Surgery files for these three years.

- **Cardiac Surgery Reporting System (CSRS)** is a voluntary reporting system of all in-hospital cardiac surgeries. It contains risk factors, clinical descriptors and procedure data and is used as a research data set. We received these files for 1991-1993.

Figure 2 summarizes the size of the original data files. The DDA/UBF files contain between 2.5 and 3 million records each year. The ambulatory surgery files were not segregated by year and contain slightly more than two million records. The CSRS data files are also summarized by year and contain a considerably smaller number of records because of the more narrow focus of the records on cardiac surgery.

Figure 2. -- Summary of Sizes of Complete Data Files

Data Set	Year(s)	Number of Records
SPARCS	1991	1,687,521
	1992	1,677,948
	1993	1,660,109
Ambulatory Surgery	1991-1993	2,121,542
Cardiac Surgery Reporting System (CSRS)	1991	19,783
	1992	21,592
	1993	22,491

Research Question

One of the primary goals of this project was to determine whether a medical effectiveness research

question could be successfully addressed by the linked data. The selected *research question* for this project relates risk factors, treatment and outcome of cardiovascular disease to minority status:

> *Are the racial/ethnic differences in mortality and morbidity from coronary heart disease related to racial/ethnic differences in treatment?*

The working *hypothesis* stated that minorities are less likely to receive surgical treatment for coronary artery disease and, therefore as a group, experience higher incidence of cardiovascular morbidity and mortality than the majority U.S. population. The cohort was to be extracted from the linked SPARCS and CSRS data sets. The linked data sets were to contain records for 3 years, 1991-1993. Males and females aged 45-75 who were assigned a diagnosis of ischemic heart disease (ICD-9 codes 410 - 414) were to be included.

Confidentiality Approach

One of the primary issues in acquiring the New York State files was data confidentiality. Technically, the problems of confidentiality of data are often addressed by suppressing, encrypting or compressing information. In these data sets primary identifiers such as name, address, and telephone number were removed or suppressed from the files and secondary identifiers such as Medical Record Number (MRN), Admission Number, and Physician License Numbers (PLNs) were encrypted consistently across files and years to aid the matching process. Typically, confidentiality restrictions hinder the matching of large data sets. Identifiers such as name, address, and medical record number are important in order to be confident that the correct linkages are being made. If only demographic data and broad geographic identifiers are available such as gender, race, age and zip code, then a large group of people may have the same characteristics with the result that their records inaccurately matched.

Cardiac Subset -- Identification and Issues

The original research plan specified the use of ICD-9 codes 410 - 414 to address the research question. The low yield on initial matches, however, indicated that we needed to expand these codes to obtain a more complete record match between the SPARCS and CSRS files. Therefore, for the linkage process, the codes were expanded to include: 390.xx - 459.xx - disease of the circulatory system; 212.7x - benign neoplasm of the heart; 745.xx - bulbus cordis anomalies and other cardiac anomalies; 861.0x - injury to the heart without open wound to thorax; 861.1x - injury to the heart with open wound to thorax; 901.xx - injury to thoracic aorta; and 996.0x - mechanical complication of cardiac device. Figure 3 summarizes the number of potential patients on the DDA/UBF files using the ischemic heart disease codes (ICD-9 410 - 414) and an expanded set of codes.

Figure 3. -- Universe of Patient Records In DDA/UBF

DDA/UBF File Year	Initial Universe of ICD-9 Codes - 410-414	Expanded Universe of ICD-9 Codes - 390-459
1991	170,779	626,222
1992	189,198	699,246
1993	190,497	714,583

Record Linkage

The linkage software used for this project was MatchWare Technology Incorporated's (MTI) *Automatch*, developed by MTI's founder, Matthew Jaro (Jaro, 1997). MTI was KAI's subcontractor and its linkage experts collaborated with KAI's clinical researchers in conducting this project.

Several steps were involved in the data preparation process prior to performing the record matching or linking process. Fields that are common to the files had to be identified and recoded, where necessary, for potential use in the linkage process. Common person and event fields included for all three data sets were MRN, sex, date of birth, patient county, hospital identification number, diagnosis, procedure code and date, and Physician License Number (PLN). Fields common to two of the three files included age, patient zip code and state, admit date, discharge date, and payor.

As an example of recoding needs, race codes on the CSRS files were converted to correspond to SPARCS codes as shown in Figure 4.

Figure 4. -- Race Code Conversions

Description	SPARCS Race	CSRS Race
Asian or Pacific Islander	1	8
Black	2	2
Hispanic	3	8
Native American	4	8
Other	5	8
White	6	1

Linkage Objective

The linkage objective was to build a longitudinal, comprehensive patient history that captured clinical encounters over time and across care settings. Thus, records for the same patient were linked in two ways: matches were performed within each of the three data sets; and matches were performed between the DDA/UBF files and CSRS and between the DDA/UBF and Ambulatory Surgery files.

Steps in Record Linkage

Steps in the linkage process included identifying duplicate records; running preliminary matches as an iterative process to determine which fields yielded the most appropriate matches; identifying appropriate cut-off weights; and running the final linkage.

Duplicate records were identified on each of the files with no file having duplicates that exceeded 1% of the records. Automatch's method for determining most effective variables and probability weights to match across files were evaluated in preliminary iterative match runs. The process was iterative and consisted of selecting key variables for each match strategy, producing preliminary matched pairs, examining matched pairs with marginal match weights, and revising the parameters to better discriminate between apparent true and false matches. For the final matches specific probabilities of agreement were determined based on the preliminary matches. The match cutoff weight was chosen so that the estimated absolute odds of a true match for record pairs with that match weight were 95:5; i.e., a confidence level of .95 of a true match.

Linkage Data Quality Analysis

The linkage results were reviewed for data reliability and validity. First, the same variables on linked and unlinked records were compared to assess internal consistency and reliability. Agreement was 99% or greater for all variables except for date of principal procedure (67%) and admission number (83%); MRN, zip code, county, and other procedure each exhibited an agreement rate of 93%. Principal procedure as well as other procedure differences may reflect differences in reimbursement categories that were changed on the UBF for payment advantages. Admission number and MRN are scrambled by computer and any clerical error such as a transposition of numbers in the original MRN yields an inconsistent scrambled MRN. Likewise, transposition of numbers in zip code and county can yield mismatches.

The DDA and UBF responses for linked and unlinked records were then compared for the same patient. The responses are fairly consistent across DDA and UBF subfiles and between linked and unlinked records with slight differences in reimburser and diagnoses, which could be a function of the research question reflected in the linked files.

The DDA variables were selected for matching and were compared for linked and unlinked patient records, because of their tendency to be more reliable in the clinical area. In the linked records, patients are older (age $\geq$ 65 - 71% versus 59% for unlinked records), most likely reflecting the research question which focuses on cardiac diagnoses. Racial characteristics are similar as are ethnicity and gender.

Linked and unlinked records for Ambulatory Surgery patients were also compared. Analysis showed a greater percentage of the linked records to have a higher proportion of angina as the primary diagnosis while in the unlinked files there was a higher proportion of arterial disease, perhaps reflecting procedures performed in ambulatory surgery, i.e., angiograms. There were more Medicare reimbursers in the linked records which is consistent with differences in age groups. Other fields show no differences. Linked records compared with unlinked records for the CSRS patients showed a greater proportion of persons over 65, most likely reflecting the diagnostic groups of research interest. There were no gender, race, or ethnicity differences in the linked and unlinked records, reflecting similar patient populations.

The general consistency between the DDA and UBF subgroups and the consistency between linked and unlinked records within each of the data sets demonstrate the reliability of the matching and indicates that the linked records generally reflect the file population.

Racial Subsets

Responses across racial subgroups for DDA variables were reviewed. As expected, more Blacks, Asians and other minorities are treated in the New York City area (over 70%) than other parts of the state. Payment also differs, with a higher proportion of Whites on Medicare (69% versus 46% for Blacks and Others and 40% for Asian Americans. A higher proportion of Blacks and other minorities have Medicaid as the primary reimburser (Blacks -- 26%, Whites -- 5%, Asian Americans -- 25%, Other -- 27%). Blacks have a higher proportion of diabetes (4% versus 1% for Whites, 2% for Asian Americans and Other) and hypertension diagnosis (5% versus 1% for Whites, and 2% for Asian Americans and Other), and a slightly lower proportion of myocardial infarctions (Whites -- 11%, Blacks -- 7%, Asian Americans -- 10%, Other -- 11%) as principal diagnosis. Responses for other variables for linked and unlinked records by racial and ethnic categories are consistent, indicating that the linked file is a representative subset of the larger file.

Research Subsets

The research subsets, defined as the original diagnoses categories, 410.xx - 414.xx, were examined next. Comparing the DDA and UBF records on the SPARCS data set for linked and unlinked records indicates that age is higher on the linked records (age $\geq$ 65 = 75%) than on the unlinked records ($\geq$ 65 = 59%), reflecting the cardiac procedure research question. Also reflecting the research question is the larger number

of patients on Medicare in the linked data set (74% versus 59% in the unlinked data set). Comparisons between linked and unlinked records in the ambulatory surgery research files indicates no significant differences between the two subsets.

A review of responses for racial and ethnic subgroups for the linked and unlinked subsets in the DDA research file indicates that in both Whites are significantly older (78% Whites in the linked subset and 66% Whites in the unlinked subset are 65 or older). In the other racial categories, however there is a larger proportion under 65 (Blacks -- 42%; Asian Americans -- 33%; and Other -- 56%) in both linked and unlinked subgroups. The age differences between White and minority racial subgroups are also reflected in the proportion of patients on Medicare. There do not appear to be other major differences between White and minority subgroups. These trends are also reflected in the differences between Hispanic and non-Hispanic subgroups.

Linked Data Sets and the Research Question

Preparing the data to answer the research question was a complex process despite the fact that record linkage had taken place. The primary reason for the complexity of the process is that the research question focuses on outcome while the linkage focused on diagnosis. The linkage focus on diagnosis appears logical because it is how patients are generally categorized for health services and clinical research. However, medical effectiveness questions often focus on outcomes and thus, within diagnoses, outcome is an important patient characteristic. The research question, while resulting in a complex subject identification procedure, was typical of many medical effectiveness questions. The amount of time, then, needed for progressing from a linked data set to analyses for outcomes research, is several months and should be built into the research planning process.

Data and Linkage Issues

Several issues related to health care data sets and application of linkage methodology were identified:

- **Purpose. --** The purpose of the primary data collection endeavor impacts on the quality of specific variables and on their utility for linkage and their relevance for addressing a medical effectiveness question. For example, primary diagnosis frequently differed between the DDA and UBF subfiles. The diagnoses in the DDA is driven by clinical practice while in the UBF it is driven by reimbursement. Variables such as age and date of birth, gender, county of residence, hospital identification number, MRN, admission date, and procedure date may not be consistent across billing and discharge administrative records as well as the research records for several reasons: accuracy is not important for billing, discharge, and some research; an individual's high anxiety state; and family members reporting information under stress. Further, discharge abstracts generally reflect clinical diagnoses more accurately, while billing data typically reflect charge justification.

- **Encryption. --** Encrypting the Medical Record number (MRN), admission numbers, and physician license numbers degrades the efficiency of the matching software. The matching software used in this study can take into account slight differences among identifiers such as transposition of characters and adjust the match for them. However, since the encryption process scrambles identifiers or assigns a sequential number to records, the software is not dealing with actual numeric identifiers, which may have typographical errors. Thus this feature of the software is not useful for electronically encrypted or created numbers. The degradation was demonstrated in the first matching pass between the ambulatory surgery file and the DDA/UBF file. The MRN in the DDA/UBF file is defined as ten characters and was encrypted as such. The MRN in the Ambulatory Surgery file is defined as seventeen characters in which the first ten characters actually contain the MRN and the last seven characters are spaces. When the initial match between the DDA/UBF file and the Ambulatory

Surgery file took place there were no matches. The resolution involved the recreation of the Ambulatory Surgery File using only the first ten characters of the MRN in the encryption process. If however, the MRNs had been provided without being encrypted, the software could have adjusted for the spaces at the end of the original MRN in the Ambulatory Surgery File.

- **Race and Ethnicity Codes. --** The race and ethnicity codes are not always accurate as demonstrated by all observations for a particular New York State hospital which contained a race code of 5 and ethnicity code of 2 for all patient records. Additionally, the state SPARCS programmer indicated that there were software problems for RACE and ETHNICITY for certain hospitals that affected accuracy.

- **Dependent Relationships Among Variables. --** Certain pairs of patient and provider variables are strongly dependent on each other. For example, MRN is frequently hospital-specific and physicians are generally associated with only a few hospitals, thus PLN (Physician License Number) and Hospital Identification Number are also strongly dependent as shown statistically by *chi square* and *uncertainty coefficient* tests. The *Automatch* software requires that only one member of each dependent pair is used as a match variable because of relative odds of a true match calculation. For example, if both date of birth and age were used in a matching process, the calculated match weight would overstate the relative odds of a true match by exactly the contribution of the second occurrence. While date of birth and age represent the same concept, hospital and physicians may be logically independent entities although statistically associated. The nature of association in health related records should be considered in the matching process and perhaps, a different statistical approach used for these data.

- **Matching Variables. --** A related issue is determining what variables provide the greatest yield during the blocking and matching procedures. Linking is generally dependent upon person identifiers such as name and address, and date of birth, as well as on procedure and diagnosis codes from health related records. Since name and address were omitted from the files used to preserve personal privacy, other variables assumed greater importance. The clinical research staff, experienced with clinical data, recommended the use of age and date of birth, gender, county of residence, hospital identification number, MRN, admission date, and procedure date. The researchers pointed out that procedure and diagnoses codes can vary between administrative and clinical data sets because of reimbursement interests and are more likely to be accurate in clinical files. Identification of the variables most appropriate for linking health related files is still an open research issue.

- **Type and Number of Variables Utilized for Linking. --** Personal identifiers such as name and address are frequently used in census and vital statistics linkage efforts. Since these variables are not present on the health files, other variables that appear in several files and have a high probability of accuracy must be identified. Some examples are hospital identification number, admission date, and zip code. Additionally, linkage software experts often argue for numerous variables upon which to link. We found that the health- related data sets were more frequently linked with fewer discrepancies in the matching records when fewer variables are used. Thus the percentage of "true" matches was higher with fewer variables or, conversely, the number of false positives was lower. However, the total number of matched records was fewer.

- **Experience from Other Applications. --** Experience and assumptions gathered from other applications of linkage methodology such as census data cannot necessarily be applied to health- related data. Thus, for health- related data, multidisciplinary teams of linkage software programmers and health researchers need to develop appropriate linkage algorithms and to identify variables pertinent for linking these files.

Findings

Despite time delays and other issues, the files were successfully linked and the data were used to address the above hypothesis that pertains to care among minority populations. General findings are as follows:

- **Data quality** in the administrative and research files generally appears high and the data are potentially useful for health services research.

- **Both the linkage process and the analytic phase** for large data sets are lengthy and resource consuming. The practicality of linking large health- related data sets needs to be balanced against the number of years the data will be useful. If data can be used to support research for three to five years, then the linkage overhead expense may be justifiable. Costs of linking large data sets, then need to be balanced against the potential benefits.

- **Linking is only the first step** when the data are to be used to address research questions. The linkage process identifies a set of unique indexes for each of the patient records in each of the linked files. Depending upon the focus of the research question, it is necessary to carefully review the data files and the index files, which consumes both time and computer processing. Since the data files for large data sets must reside on mainframe computers, it also is a costly process.

  In this project, in which those subjects with the same diagnoses who received cardiac surgery are compared to those who did not, patients with relevant diagnoses had to be identified to form a subgroup from the SPARCS DDA/UBF files. The subgroup had to then be identified on the index files, determined whether linked or not linked to the CSRS file, and then found on the CSRS files. These steps precede any analytic procedures and represent the complexity of data management procedures that are associated with the analysis of the linked files.

- **Utility of administrative data sets** in answering medical effectiveness questions is variable. Clearly, identifying diagnoses, treatment, and outcome at a general level is possible and meaningful. The data set can be used to explore potential associations among diagnoses, treatment, and outcome variables. However, understanding the mediating factors and decision making variables that result in a patient proceeding to surgery or not may not be possible. For example, the results of an angiogram for a patient with ischemic heart disease are not recorded in SPARCS DDA/UBF or in CSRS. Thus, understanding why some patients who have angiograms proceed to surgery and others do not

is not possible.

Recommendations

This project yielded the following recommendations:

- **Utilize Linking Techniques for Projects With a Three- to Five-Year Life. --** Because of the time, labor, and financial costs of linking large data sets, it would appear practical to utilize linking techniques for data that can be analyzed over a period of three to five years.

- **Continue Methods Research. --** Issues in data dependence and optimal variables for use in linking health related data sets should be addressed in additional research projects.

- **Multidisciplinary Teams. --** The need for utilizing multidisciplinary teams composed of health researchers, programmers, and linkage experts was demonstrated in the linkage process.

- **Linking Prior to Research Use. --** Future efforts may enlarge record linkage before data are released from the agency that holds authority for the data to avoid degradation of data from scrambling or encryption. Linking prior to release across agencies raises issues of data sharing, protection of privacy, and other operational issues that must be addressed.

- **Recognize Time Needed for Research. --** Research efforts using linked data sets must allocate sufficient time and manpower resources to identify and extract the suitable subpopulation for a specific research question.

Selected References

Kunitz and Associates, Inc. (1996). Record Linkage Methodology Applied to Linking Automated Data Bases Containing Racial and Ethnic Identifiers to Medical Administrative Data Bases. Unpublished Final Report.

Jaro, Matt (1997). MatchWare Product Overview, *Record Linkage Techniques – 1997*, Washington, DC: National Academy Press.

Schwartz, H.; Kunitz, S.; Jaro, M.; Therlault, G.; and Kozloff, R. (1996). Studying Treatment Variation among Minority Populations via Linked Administrative and Clinical Data Sets, *Proceedings of the Section on Social Statistics, American Statistical Association.*

Chapter
10

Matching Census Database and Manitoba Health Care Files

Christian Houle, Jean-Marie Berthelot, Pierre David, and Michael C. Wolfson, Statistics Canada; Cam Mustard and Leslie Roos, University of Manitoba

Abstract

Introduction*: In the current economic context, all partners in health care delivery systems, be they public or private, are obliged to identify the factors that influence the utilization of health care services. To improve our understanding of the mechanisms that underlie these relationships, Statistics Canada and the Manitoba Centre for Health Policy and Evaluation have set up a new database. For a representative sample of the population of the province of Manitoba, cross-sectional microdata on individuals' health and socio-economic characteristics were linked with detailed longitudinal data on utilization of health care services.*

Data and methods*: The 1986-87 Health and Activity Limitation Survey, the 1986 Census and the files of Manitoba Health were matched (without using names or addresses) utilizing a CANLINK software. In the pilot project, 20,000 units were selected from the Census according to modern sampling techniques. Before the files were matched, consultations were held and an agreement signed by all parties to establish a framework for protecting privacy and preserving the confidentiality of the data.*

Results*: A match rate of 74% was obtained for private households. A quality evaluation based on the comparisons of names and addresses over a small subsample established that the overall concordance rate among matched pairs was 95.5%. The match rates and concordance rates varied according by age and household composition. Estimates produced from the sample accurately reflected the socio-demographic profile, mortality, hospitalization rate, health care costs, and consumption of health care by Manitoba residents.*

Discussion*: The match rate of 74% was satisfactory in comparison with response rates reported by the majority of population surveys. Because of the excellent concordance rate and the accuracy of the estimates obtained from the sample, this database will provide an adequate basis for studying the association between socio-demographic characteristics, health and health care utilization in province of Manitoba.*

Introduction

A number of studies have clearly shown that there is a link between an individual's socio-economic status and the probability of his or her death during a given period of time (Wolfson et al., 1993; Marmot, 1986; Wilkins et al., 1991). Other studies have shown that the prevalence of certain diseases varies greatly depending on the socio-economic characteristics of the area in which an individual resides (Anderson, 1993, Dougherty, 1990). In addition, several Canadian surveys have already provided cross-sectional data on individuals' health status and socio-economic status, along with self-reported information

on the use of health services, e.g., General Social Survey of 1991 (Statistics Canada, 1994a), Ontario Health Survey of 1990, Enquête Santé Québec of 1987 and 1992-93, Health and Activity Limitation Survey of 1986 and 1991 (Statistics Canada, 1988), Canadian Health and Disability Survey of 1983-84 (Statistics Canada, 1986a), Canada Health Survey of 1978-79 (Health and Welfare Canada, 1981). However, to our knowledge, there is no Canadian longitudinal database that combines information on health, use of health services, and socio-economic characteristics. In an effort to meet this information need, Statistics Canada and the Manitoba Centre for Health Policy and Evaluation (MCHPE) set up a joint pilot project to evaluate the possibility of creating such a database using existing data sources.

The primary objective of the pilot project was to evaluate the feasibility of combining the following three data sources: the 1986 Census of Population, the 1986-1987 Health and Activity Limitation Survey (HALS) , and the Manitoba Health (MH) longitudinal file on health care service utilization. The database resulting from this combination will enable researchers to explore new directions with respect to health determinants. In this article, we describe the matching of files, the selection of the sample for analysis purposes and the results, which show the representativeness of the database created and validate the techniques employed.

Confidentiality and Right to Privacy

When creating a database from both administrative and survey data, it is essential to ensure the confidentiality of the data and prevent any unwarranted intrusion into individuals' privacy. In accordance with the policies of the collaborating agencies, certain procedures were undertaken prior to matching these data sets. They include consultations with the Privacy Commissioner of Canada, the Faculty Committee on the Use of Human Subjects in Research at the University of Manitoba, and Statistics Canada's Confidentiality and Legislation Committee. In addition, Manitoba Health's Committee on Access and Confidentiality was informed of the project.

Following these consultations, and in accordance with the formal policies of Statistics Canada, the Minister responsible for Statistics Canada authorized the matching as outlined below:

- A pilot project for evaluating the feasibility and utility of data matching.
- It was explicitly stated that individuals' names and addresses would not be used for matching purposes, nor would they appear in the database.
- The matching would be done entirely on the premises of Statistics Canada by persons sworn in under the Statistics Act.
- Only a sample of 20,000 matched units would be used for purposes of research and analysis; and
- Access to the final data would be strictly controlled in accordance with the provisions of the Statistics Act. In addition, all activities with the linked data set are covered by a memorandum of understanding including Statistics Canada, the University of Manitoba and the Manitoba Ministry of Health.

Data

The detailed questionnaire (questionnaire 2B) of the 1986 Census of Population contains extensive socio-economic information including variables such as family composition, dwelling characteristics, tenure, ethnic origin and mother tongue, as well as a number of variables relating to income and educational attainment (Statistics Canada, 1986b). This questionnaire was filled by persons residing in Manitoba on

June 3, 1986 in a proportion of approximately one household in five. The other households completed a short form designed solely for enumerating the population. Thus, the file used for matching purposes consisted of 261,861 records. The individuals represented by these records lived in two types of dwellings: private or collective. While this article focuses primarily on the private household component, there were in 1986 more than 26,161 persons in Manitoba living in a collective dwelling according to the Census. Examples of collective dwellings are hospitals, hospices, nursing homes, institutions for the physically handicapped, orphanages, psychiatric institutions, hotels/motels, work camps, jails, Hutterite colonies, military residences, religious institutions, student residences and YMCAs.

The 1986 HALS was a postcensal survey that sought to identify individuals who, because of their health, were limited by the type or amount of daily activities that they could perform. A postcensal survey refers to a question from the Census (in this case, Question 20 on disabilities) which serves to enrich the survey sample by identifying a high proportion of the target population. An appropriate questionnaire was then completed for each person sampled. For HALS, the Manitoba population living in private households and having disabilities was studied on the basis of a sample of 5,480 persons representing a population of 150,857 persons having at least one disability. The data set created, contained information on individuals' health and functional limitations as well as on type of employment, educational level, transportation, housing and recreation. Since the survey was of the self-reporting type, the data represent the situation of respondents from their viewpoint rather than from an administrative or clinical viewpoint.

The MH longitudinal file, for its part, contains information on visits to physicians, stays in hospital, diagnoses, surgical procedures, admission to personal care (nursing) home, health care received at home, the date and cause of death, and other data on health care utilization. A number of innovative studies in health care research have used this file (Roos et al., 1987, Shapiro and Roos, 1984). For this pilot project, a register of persons covered by Manitoba health insurance was identified from June 1986, using the date of commencement of health insurance coverage and the date of cancellation of coverage. The register contained 1,047,443 records.

Methods

The matching project was divided into three main stages. The first stage consisted of pairing individuals belonging to three distinct data sources. The second stage consisted of assessing what proportions of the pairs formed represented the same individual. The third stage consisted of selecting a sample of 20,000 matched units used to create the database for analysis purposes. In this section, we shall deal with the methodology used in each of these stages in turn.

Matching

The CANLINK system (Smith, 1981; Fellegi and Sunter, 1969) developed at Statistics Canada, was used for the pairing stage. CANLINK is a probabilistic matching software that pairs records from two sets of data by using the discriminatory power of the common variables available. The software weights the pairs of records according to the degree of concordance of the values observed and also takes account of the probability of random concordances. The files paired were that of the 2B sample from the 1986 Census covering the province of Manitoba and the file of persons registered with MH in June 1986, containing a subset of the variables available. Only these two files were involved in the probabilistic matching, since the 1986-1987 HALS sample was drawn from the Census 2B sample (Dolson et al., 1987), and all HALS records were already paired to those in the Census by a unique identifier.

The individual records which were paired came from two files, one containing the records of 261,861 individuals living in Manitoba (derived from the 2B file of the 1986 Census), and the other, containing the records of 1,047,443 persons (a derivative of the Manitoba Health file). The strategy adopted for identifying

pairs representing the same individual (good pairs) consisted of dividing up the two data sets into blocks and forming only pairs of individuals belonging to the same block.

The pairs of records were compared only if all the blocking variables concurred. It was therefore necessary to choose carefully so as not to eliminate at an early stage a great number of "good" pairs. It will be recalled that the most discriminant variables, namely surnames, given names and addresses, were not used in this study. Because of this constraint, we were forced to choose other combinations of variables that were limited in discriminatory power and then apply innovative techniques.

Two matching phases were carried out. First, after examining various possible definitions, we defined a block as a set of four individual characteristics, namely a person's sex, year of birth, month of birth and postal code. In the second matching phase, the definition was relaxed in order to form more pairs of individuals. The exact year and month of birth were replaced by the person's age, which made it possible to compare an individual with a greater number of candidates. In addition, the area covered by the geographic variable in urban settings was expanded by a factor of approximately three, with the postal code being replaced by the census enumeration area.

Through these matchings the census file was divided into three subsets: records which had clearly matched (definites), those which had matched but for which the discriminatory power of the available variables raised a doubt (based on CANLINK criteria (possibles)), and those which had never matched.

While the information on family structure was used in the matching process, the CANLINK system compared only two individuals at a time, without taking account of matches obtained for other family members. We had to define a series of rules in order to ensure the consistency of matchings within a given family and between two matching phases (David et al., 1993).

Evaluation of the Concordance of the Pairs Formed

The evaluation pursued two objectives. First, it was important to determine the degree of accuracy with which we had associated the Manitoba Health data with the Census data (definite matches only). Then it was necessary to assess whether the rules that had been developed for rejecting certain "possible" matches were adequate.

A sample of 1,000 families was drawn, representing 2,102 matched individuals. As stratification variables, we used urban/rural area as determined by Census, family composition (person living alone, couple with child, couple without child, multiple family) and matching status (definite or possible). MH extracted the names and addresses of all these individuals and their family members. It should be understood that this identifying information was not used to determine the validity of specific matches, but only to estimate actual matching rates at aggregated levels. Names and addresses were compared manually with those on the mircrofilmed 2B questionnaires kept at Statistics Canada.

Sample Selection

As the project involved three databases, the sampling frame was derived from the 2B file from the 1986 Census. The sample size was already set at a maximum of 20,000 units and the database created had to combine information from the census files and the MH file, as well as information from the HALS file. The HALS sample used the individual as its sampling unit, whereas the analysis of the overall population of Manitoba used the household as defined by the Census. Several options were considered in order to try to construct a single database, however the complexity of the analysis would have negated any potential gains in accuracy. To ensure a balance between simplicity of analysis and an effective design, the selection process consisted of constructing two independent databases: the first, to study the link between disability, socio-economic status and health, and the second, to analyse the general population of Manitoba. To

maximize the use of the 20,000 units, it was also necessary to take account of the overlap between these two databases.

Owing to the complex sample design of HALS, the relatively small number of individuals sampled and the importance of this database from an analytical standpoint, matched individuals with disabilities were all selected. These accounted for 4,434 basic units. This sample formed the first database, used for the analysis of persons with at least one functional disability.

There were therefore 15,566 units left to form the general population database plus the expected number of units overlapping the two databases. Still pursuing the objective of optimizing the sample design, an evaluation indicated that stratification was appropriate. Stratification has several advantages. First, it serves to reduce the overall variance of the estimates. Second, it ensures a standard of quality for estimates relating to subgroups of interest in the population. Third, stratification can result in improved accuracy for cases in which non-sampling error can be taken into account. Finally, stratification is especially effective when the stratification variables are correlated with the target variable.

Since a number of studies have established links between socio-economic status and health, it was natural to use socio-economic variables to construct the strata. In addition, there was no disadvantage to using the household as the sampling unit, since socio-economic status is generally the same for all members of a given household. Since it was the 1986 Census file that entirely determines the composition of this population, all the stratification variables were either taken directly from that file or derived from it. The final number of strata for private households was 611. The total number of units drawn was 16,387. These represented 46,670 persons.

Finally, it is common practice to adjust sampling weights so that the totals estimated by the sample will reflect as accurately as possible the counts of the population studied. With post-stratification, the counts can be adjusted for categories for which the number of units was insufficient to create a real stratum but which were of sufficient analytical importance to justify the use of special techniques. These techniques changed the initial weight subject to the constraints of minimum change (Kovacevic, 1995). For private households, the counts by age group, rural or urban geographic area, marital status and sex were used to adjust the weights to the individual level, while rural or urban geographic area, household size and tenure played the same role for adjusting at the household level.

Results

Results for Matching

Despite the conservative approach applied to the initial matches, overall, 74% (174,476 out of 235,700) of individuals from the census living in a private household were matched with an individual in the Manitoba file. This rate varied according to geographic mobility, age, marital status and family size.

The factors that had the greatest influence on the match rate were all related to individuals' geographic mobility. Hence, the following groups of individuals were more difficult to match: young adults (between 20 and 25 years of age: Figure 1), persons who had changed their place of residence between the 1981 and 1986 censuses (Table 1), and divorced or separated persons (Table 2). Among these groups, frequent changes of address and family structure made concordance between the two data sources more difficult than among less mobile groups. The reason for this is that since the Census figures date from June 3, 1986, and some MH data are dated December 31, 1986, there was more likely to be an information lag with respect to mobile individuals.

Figure 1. -- Match Rate by Age
Private Households Only

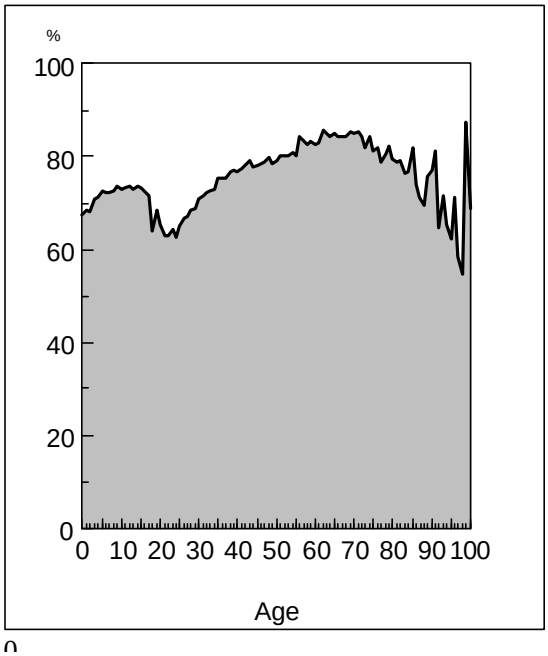


Table 1. -- Match Rate According to Mobility:
Private Households Only

Mobility	Match Rate %
Same household	81.7
Same CD	65.8
Other CD	62.5

CD: Census Division, a geographic unit used by the Census. Manitoba is made up of twenty-three Census Divisions.

Table 2. -- Match Rate According to Marital Status:
Private Households Only

Marital Status	Match Rate %
Married	78.5
Widowed	74.5
Single	71.2
Divorced	61.4
Separated	43.4

The effect of age on the match rate was not surprising. Children under fifteen years of age and adults between thirty and sixty years of age had better rates, owing to their more stable situation. Among individuals over 85 years of age, there was greater variability in the data due to the rate of institutionalization and the small number of cases.

Among individuals who did not move between the 1981 and 1986 censuses (same household), one might have expected an even better match rate. The rate of 81.7% is perhaps an indication that using the methodology described thus far, there is an upper limit of around 80% on matches, given that the files are not totally free of errors.

Intuitively, family size is correlated in two opposite ways with the match rate. While a large family has an intrinsic constraint on the mobility of the family nucleus, some members of the family will periodically attach or re-attach themselves to this nucleus. Table 3 indicates that match rate dropped off significantly as family size increased.

Table 3. -- Match Rate According to Family Size:
Private Households Only

Family Size	1	2	3	4	5	6	7	8	9	10+
Match rate %	66.4	78.5	74.7	79.8	77.1	70.0	55.9	51.4	39.2	46.7

Results for Evaluation of Concordance of Pairs Formed

Table 4 shows that overall, more than 95% of the definite matches retained represented the same individual. As the sample of 20,000 units was drawn from definite matches only, this meant that the matching was of exceptional quality. Also, due to the fact that the rate of concordance of names among possible pairs was only 40% indicated that the enforcement of a conservative methodology was justified and prudent, as they prevented a large proportion of bad matches.

Household size was closely related to the concordance rate. Persons living alone and those living in households of eight or more persons exhibited a lower concordance rate, namely 86.8% and 90.8% respectively. These results would seem to be due to the small number of discriminant variables available for persons living alone and the fact that in the case of large households, there was often more than one family within the household.

Table 4. -- Rate of Concordance of Names According to Various Groupings:

Private Households Only

Match Status	Concordance on Names %	Standard Error* %
"Possible" match"	40.1	2.4
"Definite" match"	95.5	0.5
Indian reserves	95.3	1.5
Household of 1 person	86.8	2.5
Household of 2 to 7 persons	96.5	0.5
Household of 8 or more persons	90.8	2.0

** The design effect is ignored in calculating the standard error.*

A final point to be observed is that ***matched*** inhabitants of Indian reserves [1] had a concordance rate equivalent to that of persons living off reserve.

Results for Sample Selection

Overall, we were pursuing two specific objectives in designing the stratification. First, it was necessary to come as close as possible to having a self-weighted design so as to allow for the use of existing computer applications. The costs of custom applications and the time required to develop them would have been a major handicap for any subsequent analysis. The first objective was attained by avoiding oversampling of strata to the extent possible and by maintaining a certain uniformity of weights within each stratum formed. The second objective was to use socio-economic variables in the process of forming strata. Therefore, when stratum sizes permitted, we used variables derived from income, education level, family structure, age and geography.

There were major conceptual differences with respect to the definition of the populations represented by the Census and by the MH file. Only persons having a usual place of residence in Manitoba on June 3, 1986 were enumerated in that province. The MH file was made up of all persons who were covered by the health insurance plan. Some of these persons no longer lived in Manitoba or may not have indicated a change in their status, resulting in some overcoverage. The MH file contained no information on residents of several categories of collective dwellings for which medical services were provided by the federal government, such as military camps and some Indian reserves, whereas the Census considered these persons to be residents of Manitoba.

For purposes of comparison, we excluded persons living in nursing homes (an institutional collective dwelling) from the MH counts in the tables that follow. According to the census definition, persons who had stayed for 180 days or more in a health care institution were considered institutionalized, and therefore excluded. Despite efforts to make the two universes uniform, the fact remains that we managed only to approximate the counts of persons living in institutions. Consequently, the populations compared represent ***approximately*** those persons living in a private household or a ***non***-institutional collective household.

As Table 5 shows, despite major conceptual differences, the sizes of the two populations by age group were quite comparable. Overall, the estimated total sizes of the two populations differed by only 0.1%, although males were underestimated by 1.2% and females overestimated by 1.4%. It was also observed that the greatest differences were amongst younger individuals.

Table 5. -- Accuracy of the Sample by Age Group Versus MH:
Private And Non-Institutional Collective Households

Age	Males MH	Difference from Sample %	Females MH	Difference from Sample %	Total MH	Difference from Sample %
0 to 4 years	32 743	2.57	31 105	1.98	63 848	2.28
5 to 14 years	78 076	1.47	73 912	2.87	151 988	2.15
15 to 24 years	86 722	-1.24	82 971	1.61	169 693	0.15
25 to 44 years	165 783	-2.84	159 458	1.92	325 241	-0.51
45 to 64 years	96 989	-1.71	98 997	0.92	195 986	-0.38
65 years and +	57 904	-1.34	74 129	-1.10	132 033	-1.20
Total	518 217	-1.20	520 572	1.39	1 038 789	0.10

Tables 6, 7 and 8 compare the mortality rate, medical care utilization and hospital care utilization by whether they were estimated from our sample or from the MH file. It should be noted that the death rates reported in the literature (Statistics Canada, 1994b) were slightly higher than those presented in Table 6, with the difference increasing with age. This may be explained by the fact that our files exclude individuals living an institutional collective dwellings, who exhibit a higher mortality rate than persons living in private households.

Table 6. -- Annualized Mortality Rates Based on the Period from June 1986 to May 1989:
Private and Non-Institutional Collective Households

Age	Annual Mortality Rate* (x 1,000) MH	Annual Mortality Rate* (x 1,000) Sample	95% Confidence Interval for the Sample
0-4	0.51	0.02	(0, 0.18)
5-44	2.49	2.04	(1.54, 2.54)
45-49	3.23	2.78	(0.57, 4.99)
50-54	5.12	3.68	(1.03, 6.33)
55-59	8.42	8.91	(4.80, 13.02)
60-64	12.43	10.48	(6.01, 14.95)
65-69	18.96	19.03	(12.69, 25.37)
70-74	28.63	24.59	(16.76, 32.42)
75-79	42.77	43.09	(30.85, 55.33)
80 and over	76.98	73.67	(57.41, 89.93)
Total	6.71	6.11	(5.39, 6.83)

**Number of estimated deaths (over the three years period) divided by estimated population total times 3.*

Overall, the mortality rate estimated by the matched sample (6.11) was lower than the one derived from the MH file (6.71), but this difference was not statistically significant at a 95% confidence level. It should be noted that the confidence intervals derived from our sample contained the value calculated by MH for all age groups except children aged 0 to 4. While the number of deaths in this category were relatively small, the difficulty in matching children under one year of age may be related to this underestimate. Additionally, this also indicates that any analysis specific to children from 0 to 4 years of age be conducted with caution, especially where the prevalence of a disease or condition was low.

Table 7. -- Number and Costs of Medical Services, 1986-87 Fiscal Year:
Private and Non-Institutional Collective Households

Selected Type of Practice	Number of Services			Costs of Services ($)		
	MH	Sample	Relative Difference (%)	MH	Sample	Relative Difference (%)
Internal medicine	699,542	702,735	0.46	19,060,658	18,904,922	-0.82
Paediatrics	416,157	449,122	7.92	6,932,217	7,359,290	6.16
Psychiatry	154,279	146,704	-4.91	8,970,489	8,468,584	-5.60
Surgery	406,907	409,097	0.54	21,772,057	21,230,743	-2.49
Ophthalmology	339,334	357,273	5.29	10,017,371	10,506,461	4.88
Radiology	623,712	653,850	4.83	9,564,330	9,970,191	4.24
Pathology	2,941,244	3,126,365	6.29	21,369,502	22,489,399	5.24
Obstetrics and Gynaecology	294,288	328,728	11.70	8,774,785	9,151,231	4.29
General practice	4,762,316	4,858,641	2.02	75,806,649	76,545,594	0.97
Totals*	10,938,103	11,351,540	3.78	193,386,798	195,908,988	1.30

** Totals includes some Type of Practice not showned on this table.*

For most categories of medical practice, the estimates drawn from the sample were fairly close to those presented by MH, both for the number of services and for the costs generated in providing these services. The accuracy achieved was all the more remarkable as no post-stratification was carried out at any level to adjust the consumption of health care services to the MH figure.

Table 8. -- Number and Duration of Hospital Stays, 1986-87 Fiscal Year:
Private and Non-Institutional Collective Households

Age Group	Number of Stays			Duration of Stays		
	MH	Sample	Relative Difference (%)	MH	Sample	Relative Difference (%)
0-64	100,127	96,303	-3.82	538,616	499,665	-7.23
65 and over	43,226	41,318	-4.41	452,172	414,555	-8.32
Total	143,353	137,621	-4.00	990,788	914,220	-7.73

Table 8 shows the results of the comparison of the number and duration of hospital stays. Taking the conceptual differences between the two data sources into account, it is deemed to be satisfactory to achieve an accuracy of the estimates within 10%. A larger underestimate for the duration of stays, than for the

number of stays, indicates that longer stays were prone to underestimation. This situation may be explained by the difficulty in identifying residents of institutional collective dwellings on the MH files.

Discussion

With the methodology presented in this article, approximately 74% of the census file corresponding to private households could be matched with the MH file, using mainly age, sex, postal code, family size and family structure. This rate of 74% is satisfactory when compared to the response rate reported in a number of surveys. For example, response rates for the Nova Scotia Nutrition Survey were 79.7% among located respondents and 60.0% for the total sample (Maclean, 1993). The Manitoba Heart Health Survey registered response rates of 77.1% among located respondents and 60.8% for the total sample (Young et al., 1991).

Obviously, considering the various types of errors possible with matching on a large scale, it is not realistic to expect a matching rate of 100%. It is inevitable that the success rate of any probabilistic matching exercise be affected by erroneous data, lags in the collection or updating of the information, as well as conceptual differences between the data sets. Furthermore, while non-matched individuals exhibited different characteristics from matched individuals, rich socio-demographic information concerning the non-matched population was available from the Census. This information was used to select a sample of matches representative of the entire population.

In 95.5% of cases, the pairs formed did associate with the data on an individual's health care utilization and with the socio-economic data on the same individual in the 1986 Census. This rate of accuracy is exceptional, considering that surnames, given names and birth dates were not used in the matching process.

The accuracy obtained in estimating various indicators associated with the consumption of health care (such as mortality, number and costs of medical services, number and duration of hospital stays) justifies the care with which the matching and sampling methods were developed.

In light of the match rates obtained, the rates of concordance of names and the accuracy of the estimates, it can be said that not only is the new database unique in Canada but also that the quality of the data coded in it greatly exceeds that of many surveys based on interviews.

At a time when health expenditures exceed 10% of the GDP in Canada and 13% in the United States, substantial efforts are being made to identify the relationships between health care utilization and health itself. While it is suspected that the level of health perceived by the patient explains a sizable portion of consumption, many studies have focused on consumption by a specific client group, such as the elderly, or on consumption of health care in the years prior to death (Barer et al., 1987; Shapiro and Roos, 1984).

The newly constructed microdatabase opens the door to various studies that were previously not possible in Canada. For example, one of the projects proposed by the MCHPE consists of analysing morbidity with respect to an individual's occupation and by examining the extent to which the health care utilization for a particular class of illnesses is related to the basic occupational group. The census data can be used to classify individuals according to the reported occupation, or according to whether or not they are employed and whether or not they are in the labour force. Using the 9th revision of the International Classification of Diseases (ICD-9), the potential medical conditions to be studied will be musculo-skeletal disorders, cardio-vascular diseases, mental disorders, gastrointestinal illnesses and injuries.

From a general view, there are plans to study differences in the level of health care utilization by socio-economic status at different stages in life. On the one hand, it is well-documented that the greatest consumption of health services occurs toward the end of one's life (Barer et al., 1987; Roos et al., 1987).

On the other hand, a major decline in infant mortality between 1960 and 1990 has also been observed (Pappas et al., 1993; Marmot, 1986). These two phenomena alone are justification for undertaking more thorough comparisons at all age levels. Using data on visits to doctors, health care at home and hospital admissions, it will be possible to compare health care utilization by different age groups by socio-economic status for the classes of illnesses mentioned above. In addition, several studies (Ugnat and Mark, 1987; Williams, 1990; Wilkins et al., 1991) suggest that differences in health conditions according to socio-economic status are greater among persons between 35 and 64 years of age than for other age groups. Analyses by age group using this new database confirmed these hypotheses or shed new light on these matters (Mustard, 1995).

A study to examine the impact of parental socioeconomic status on the use of hospital and ambulatory medical care services during the first year of life has just been completed. It showed that after controlling for low birth weight, maternal age and the joint effects of education and income; for hospital care, education was significantly negatively associated and exhibited a threshold effect between the lowest quartile and all other quartiles; for ambulatory treatment care, income was significantly associated and exhibited a linear effect; for preventive care, both income and education were associated and exhibited a threshold effect between the lowest quartiles and all other quartiles. In the first year of life excluding the birth event, per person public health expenditures were more than twice as high in the lowest education or income quartile compared to the highest quartile (Knighton et al., 1997).

Although the links between socio-economic status and health are the object of intensive research, one of the most frequently encountered problems is that it is impossible to have precise information on socio-economic status at the individual level. Some researchers have no other choice but to use an indicator obtained through the aggregation of taxation or census data for an area of a given size, such as the census enumeration area or the postal code area (Wilkins, 1993). Little research has been done to verify the impact and the validity of this methodology. This tends to reduce the capacity of such models to detect more subtle but theoretically important determinants.

The HALS file, combined with administrative data from health care utilization records, opens the door to comparisons which, until now, have been difficult if not impossible to make. HALS offers us a clear and detailed image of individuals suffering from disability. Whether by age group or sex, by type of disability (mobility, sight, hearing, dexterity, cognition, etc.) or severity, the Manitoba population suffering from a disability can be compared to the general population by means of the census file. Specific analyses of these data focussed on mortality and on health care utilization (Tomiak et al., 1997).

Finally, as the population ages, a greater demand for long-term care services and in particular, nursing homes is expected. A study was initiated to assess the relative importance of predisposing, enabling and need characteristics in predicting nursing home entry.

The list presented here is not exhaustive and is provided only to demonstrate how the new database can be used to analyze health care utilization at different stages in life and for different topics. It is believed that the analytical benefits produced by the record linkage of the administrative sources are important in term of public interest.

Acknowledgments

The authors wish to thank the following persons for their significant and generous contribution to this study: Shelley Derksen, J. Patrick Nicol, and Leonard McWilliam, Manitoba Centre for Health Policy and Evaluation.

Footnote

[1] It should, however, be kept in mind that the match rate on Indian reserves was only 44.5%, considerably lower than the average rate of 74%. This could lead to a bias, since the matched individuals may have had very different characteristics from the reserve population as a whole.

Bibliography

Anderson, G.;Grumbach, K.; Lutt, H.; Roos, L.L.; and Mustard, C. (1993). Use of Coronary Artery Bypass Surgery in the United States and Canada: Influence of Age and Income, *Journal of the American Medical Association*, 269, 1661-1666.

Barer, M. L. et al. (1987). New Evidence on Old Fallacies, *Social Science Medicine* 24(10), 851-862.

David, P. et al. (1993). Linking Survey and Administrative Data to Study the Determinants of Health, *Proceedings of the American Statistical Association*, San Francisco.

Dougherty, G.; Pless, I.B.; and Wilkins, R. (1990). Social Class and the Occurrence of Traffic Injuries and Death in Urban Children, *Canadian Journal of Public Health*, 81, 204-209.

Dolson, D.; Maclean, K.; Morin, J. P.; and Théberge, A. (1987). Sample Design for the Health and Activity Limitation Survey, *Survey Methodology*, 13(1), 93-108.

Health and Welfare Canada (1981). *The Health of Canadians: Report of the Canada Health Survey*, Cat. No. 82-538E.

Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, 64, 1183-1210.

Knighton, T.; Houle C.; Berthelot J. M.; and Mustard C. (1997). The Impact of Socio-Economic Inequity on the Health Care Utilization Practices of Infants During the First Year of Life, Symposium on Intergenerational Equity in Canada, Statistics Canada, Ottawa. (For reprint contact Miles Corak at (613) 951-9047.)

Kovacevic, M. (1995). The Weight Adjustment for the Sample from the "Whole Population Database" (Private Household Component), Technical Note, Statistics Canada, March 24 (not published).

Maclean, D. R. (1993). Report of the Nova Scotia Nutrition Survey, Nova Scotia Heart Health Program, Department of Health, Government of Nova Scotia.

Marmot, M.G. (1986). Social Inequalities in Mortality: The Social Environment. In: *Class and Health, Research and Longitudinal Data*, (Ed. R.G. Wilkinson), London: Tavistock Publications.

Marmot, M.G. and Mcdowall, M.E. (1986). Mortality Decline and Widening Social Inequalities. *Lancet*, I, 274-276.

Mustard, C. (1995). Socioeconomic Gradients in Mortality and the Use of Health Care Services and Different Stages in the Life Course, Research Paper, Manitoba Centre for Health Policy and Evaluation.

Pappas, G. et al. (1993). The Increasing Disparity in Mortality Between Socioeconomic Groups in the United States, 1960 and 1986. *New England Journal of Medicine*, 329, 103-109.

Roos, L.L.; Nicol, J.P.; and Cageorge, S.M. (1987). Using Administrative Data for Longitudinal Research: Comparisons with Primary Data Collection, *Journal of Chronic Diseases*, 40(1), 41-49.

Shapiro, E. and Roos, L.L. (1984). Using Health Care: Rural/Urban Differences Among the Manitoba Elderly, *The Gerontologist*, 24(3), 270-274.

Smith, Martha E. (1981). *Generalized Iterative Record Linkage System*, Health Division, Statistics Canada.

Statistics Canada (1994a). *Health Status of Canadians: Report of the 1991 General Social Survey*, Cat. No. 11-612E, No. 8.

Statistics Canada (1994b). *Causes of Death 1992*, Cat. No. 84-208

Statistics Canada (1988*), The Health and Activity Limitation Survey, Selected Data for Canada, Provinces and Territories*, Cat. No. 41034.

Statistics Canada (1986b). *Report of the Canadian Health and Disability Survey 1983-1984*, Cat. No. 82-555E.

Statistics Canada (1986b). *Census Handbook*, Cat. No. 99-104.

Tomiak, M.; Berthelot, J.M; and Mustard, C. (1997). A Profile of Health Care Utilization of the Disabled Population in Manitoba, (submitted for publication).

Ugnat A M and Mark, E. (1987). Life Expectancy by Sex, Age and Income Level. *Chronic Disease in Canada*.

Young, T.K.; Gelskey, D.E.; Macdonald, S.M.; Hook, E.; and Hamilton, S. (1991). The Manitoba Heart Health Survey: Technical Report.

Williams, D. R. (1990). Socio-Economic Differentials in Health: A Review and Redirection,. *Social Psychology Quarterly* 53(2):81-99.

Wilkins, R.; Adams, O.; and Brancker, A. (1991). Changes in Mortality by Income in Urban Canada from 1971 to 1986, *Health Reports*, 1(2), 137-174.

Wilkins, R. (1993). Use of Postal Codes and Addresses in the Analysis of Health Data, *Health Reports*, 5(2): 157-177, Cat. No. 82-003.

Wolfson, M.C.; Rowe, G.; Gentleman, J.F.; and Tomiak, M. (1993). Career Earnings and Death: A Longitudinal Analysis of Older Canadian Men, *Journal of Gerontology: Social Sciences*.

Chapter 10

The Development of Record Linkage in Scotland: The Responsive Application of Probability Matching

Steve Kendrick, National Health Service, Scotland

Abstract

Since 1968, patient identifiable records of hospital discharges, cancer registrations and death records have been held centrally in Scotland in machine readable form. Patient details are held in order to enable record linkage using probability matching. In the 1970s and early 1980s over forty ad hoc linkages were carried out. Since the late 1980s, the records have been brought together into permanently linked data sets the largest of which now contains over 12 million records spanning the years 1981 to 1995.

These linked data sets have enabled a wide range of analyses to be carried out in response to demands from the health service and the medical research community. They have ranged from relatively simple aggregations of data at the patient level to complex studies of long term patient outcomes. Outcome indicators such as 30 day survival after acute myocardial infarction are now published at hospital level.

In addition to the main "internal" linkages over seventy linkages have been carried out between external data sets such as surveys (e.g., the West of Scotland Coronary Prevention Study), employee records and clinical audit records and the centrally held linked data sets.

The linkage techniques used have evolved to meet the challenges posed by a wide range of customer requirements and data sets. In particular there has been a shift from traditional sort-and-match methods to one pass techniques involving indexing in memory. This has been necessary to enable the linking of relatively small data sets to the main data set without multiple sorting of the data. The technique is currently being adapted to the main linkages to enable much more rapid incorporation of new data. Appropriate use of "best-link" principles has made possible either very high linkage accuracy (e.g., the linkage of Scotland's two main population registers) or reasonable accuracy in linking very poor quality data sets (e.g., linkage of records of victims of cardiac arrest to death records).

The paper will use the Scottish experience to illustrate how the application of probability matching needs to be closely attuned to the precise characteristics of and, in particular, the relationship between the data sets to be linked.

Introduction

Record linkage using probability matching, like many fields of human endeavour, has progressed as a highly fruitful interplay between theory and experiment, axioms and pragmatism. One viewpoint would see record linkage as primarily a highly practical enterprise based on common-sense and close attention to the empirical characteristics of the data sets involved in any linkage. Another would emphasize the rigorous grounding of record linkage practice in statistical theory and the theory of probability (Fellegi and Sunter, 1969; Newcombe et al., 1992; Arellano, 1992).

Howard Newcombe, pioneer and founder of probability matching techniques, recognises the value of both perspectives. His work has illustrated the continuing dialectic between the theory and the practical craft of linkage. From the point of view of the development of record linkage in Scotland however his most valuable contribution, beyond his initial formulation of the principles of probability matching, has been his emphasis on being guided by the characteristics and structure of the data sets in question and close empirical attention to the emergent qualities of each linkage (Newcombe et al., 1959; Newcombe, 1988). Particularly inspiring has been his insistence that probability matching is at heart a simple and intuitive process and should not be turned into a highly specialised procedure isolated from the day to day concerns of the organization in which it is carried out (Newcombe et al., 1986).

In this paper we wish to show how the development of the methods of record linkage used in the Scottish Health Service have been driven forward by concrete circumstances and in particular by the practical demands of our customers and the needs of the health service as a whole. Although almost no specifically "research and development" time has been devoted to the development of the Scottish system, our openness to the demands of customers and the sheer variety of linkages which this has engendered has in fact produced a rapid pace of development and change which shows no sign of abating.

Although we have pursued a highly pragmatic rather than a theoretical approach, the variety of linkages which have been undertaken has served to give shape to an overview of some of the main factors which need to be taken into account in designing linkages most effectively. The paper is thus in part the story of record linkage in Scotland, in part a concrete account of how our methods have evolved but also contains an overview of some the factors to do with data structure which may be relevant to linkage strategy. More than anything however the paper is an illustration of how the sensitive and flexible application of the very simple and basic principles outlined by Howard Newcombe can produce very powerful results.

The Context

The current system of medical record linkage in Scotland was made possible by an extremely far sighted decision made as long ago as 1967 by the predecessor organisation to the Information and Statistics Division of the Scottish Health Service and by the Registrar General for Scotland. The decision was taken that from 1968 all hospital discharge records, cancer registrations and death records would be held centrally in machine readable form and would contain patient identifying information (names, dates of birth, area of residence etc.).

The decision to hold patient identifying information was taken with probability matching in mind and reflected familiarity with the early work of Howard Newcombe in Canada and close contact between Scotland and the early stages of the Oxford record linkage initiative. (Heasman, 1968; Heasman and Clarke, 1979).

In what can now be regarded as the first phase of medical record linkage in Scotland, over 40 often sizeable linkages were carried out between the late 1960s and the mid-1980s (for example, Hole et al., 1981; Kendell et al., 1987). The linkages were primarily for epidemiological purposes and each involved the rather laborious specification and development of a bespoke computer program, the whole process often taking over a year to complete. Although the system represented a considerable achievement, by the mid-

1980s it was acknowledged that it would be increasingly inadequate for the perceived future needs of the Scottish Health Service especially in terms of management information.

In the late 1980s the decision was taken to reconstitute the linkage system. Increased computing power and data storage capacity enhanced the feasibility of linking once and for all the set of records pertaining to a given patient. New enquiries, whether epidemiological or relating to service management, would increasingly involve analysis of already linked data rather than requiring fresh linkages.

The years since 1989 have seen the creation of such permanently linked data sets of Scottish health related data. (Kendrick and Clarke, 1993). The largest currently contains all hospital discharge data, cancer registrations and Registrar General's death records from 1981 to 1995 (over 14 million records relating to just over 4 million individuals). A maternity/neonatal data set contains all maternity admissions, neonatal records, Registrar General's birth records and stillbirth and infant death records from 1980 to 1995. Finally, the data set with the longest time span contains linked psychiatric inpatient records and Registrar General's death records from 1970 onwards.

It was envisioned that the creation of the national linked data sets would be carried out purely by automated algorithms with no clerical checking or intervention involved. After linkage of five years of data in the main linked data set it was found that the false positive rate in the larger groups of records was beginning to creep up beyond the 1% level felt to be acceptable for the statistical and management purposes for which the data sets are used. Limited clerical checking has been subsequently used to break up falsely linked groups. This has served to keep both the false positive and false negative rates at below one per cent. More extensive clerical checking is used for specialised purposes such as the linking of death records to the records of the Scottish cancer registry to enable accurate survival analysis for example.

The existence of the linked data sets has generated a high level of demand for analysis. Approaching a thousand analyses have been carried out ranging from simple patient based counts to complex epidemiological analyses. Among the major projects based on the linked data sets have been clinical outcome indicators (published at hospital level on a national basis), analyses of patterns of psychiatric inpatient readmissions and post-discharge mortality and analyses of trends and fluctuations in emergency admissions and the contribution of multiply admitted patients.

However, far from reducing the requirement for specialised data linkage, the existence of permanently linked national data and facilities for linkage has served to fuel the demand for new linkages. Over a hundred and fifty separate probability matching exercises have been carried out over the last five years. These have consisted primarily of linking external data sets of various forms -- survey data, clinical audit data sets -- to the central holdings. A particularly important linkage in the context of a major trial of cholesterol lowering drugs enabled comparison of the accuracy of follow-up using probability matching with reporting based on direct contact with patients. Automated linkage was found to be just as accurate for tracking hospital admissions (West of Scotland Coronary Prevention Study Group, 1995). Other specialised linkages have involved extending the linkage of subsets of the ISD data holdings back to 1968 for epidemiological purposes. (for example, Gillespie et al., 1996). These exercises have varied enormously in scale and complexity, from following up the patients of a particular consultant to linking what are virtually two different registers of the population of Scotland. Linkage proposals are subjected to close scrutiny in terms of the ethics of privacy and confidentiality by a Privacy Advisory Committee which oversees these issues for ISD Scotland and the Registrar General for Scotland.

The Scottish linkage project has been funded primarily as part of the normal operating budget of ISD Scotland. Relatively little time or resources have been available for general research into linkage methodology. Instead the development and refinement of linkage methods has taken place as a response to a wide variety of immediate operational demands. We have become to all intents and purposes a general purpose linkage facility at the heart of the Scottish Health Service operating to very tight deadlines often set in terms

of weeks and in extreme cases, days. This has placed a high premium on developing quick, effective and accurate methods of linkage with an emphasis on fitness for purpose rather than straining for precision for its own sake.

Despite the lack of time and resources available for background research and development in linkage methodology, these conditions have in fact fostered, especially in recent years, a rapidly changing and developing approach to linkage.

Before describing the most significant developments involved, a brief overview of the main components will serve to set them in context.

The Elements of Linkage

For the purposes of this discussion, record linkage using probability matching can be regarded as having three phases or elements each involving a key question.

- **Bringing pairs of records together for comparison**. -- How do we bring the most effective subset of pairs of records together for comparison? It is usually impossible to carry out probability matching on all pairs of records involved in a linkage. Usually only a subset are compared, those which share a minimum level of identifying information. This has been traditionally achieved by sorting the files into "blocks" or "pockets" within which paired comparisons are carried out (Gill and Baldwin, 1987).

- **Calculating probability weights**. -- How do we assess the relative likelihood that pairs of records belong to the same person? This lies at the heart of probability matching and has probably been the main focus of much of the record linkage literature. (Newcombe, 1988).

- **Making the linkage decision**. -- How do we convert the probability weights representing relative odds into absolute odds which will support the linkage decision? The wide variety of linkages undertaken has been particularly important in moving forward understanding in this area.

It would probably be fair to say that of the three areas, it is the second, the calculation of probability weights which has received the most attention and is the best understood. Developments in Scotland over the last few years have occurred in the other two areas as the two subsequent sections will demonstrate.

Before moving on to these developments, our approach to the calculation of probability weights has been relatively conventional and can be quickly summarised. A concern has been to avoid overelaboration and over complexity in the algorithms which calculate the weights. Beyond a certain level increasing refinement of the weight calculation routines tends to involve diminishing returns. This relatively basic approach has been facilitated by the relative richness of the identifying information available on most health related records in Scotland. To take an example, for the internal linking of hospital discharge (SMR1) records across Scotland we have available the patient's surname (plus sometimes maiden name), first initial, sex and date of birth. We also have postcode of residence. For records within the same hospital (or sometimes the same Health Board) the hospital assigned case reference number can be used. In addition positive weights can be assigned for correspondence of the date of discharge on one record with the date of admission on another. Surnames are compressed using the Soundex/NYSIIS name compression algorithms (Newcombe, 1988) with additional scoring assigned for more detailed levels of agreement and disagreement. Wherever possible specific weights relating to degrees of agreement and disagreement are used.

Bringing the Pairs of Records Together: One Pass Linkage

The Limitations of Sort-and-Match

By the time the largest linked data set covered several years of data and consisted of several millions of records, a particular challenge emerged. The linkage team began to be asked to link data sets consisting of relatively small numbers of "external" or "newcomer" records to the central catalog of identifiable records. The external or newcomer records might consist of respondents to a survey, a specialised disease register or a particular group of employees.

In all cases the aim was to link the newcomer data set to the central catalog of records so that the experience of the individuals involved could be traced forward from the date of survey, the date last known to the disease register or the date of employment.

As we have seen, in record linkage it is impossible to bring together and compare all the pairs of records involved in the linkage. The number of pairs which are brought together for comparison is normally reduced to manageable proportions by some form of blocking by which only those pairs of records which share common sets of attributes are compared. For example a common strategy is to compare only those pairs of records which share either the same first initial and NYSIIS/Soundex code or the same date of birth. The normal method of achieving such blocking is to sort the two files concerned on the basis of the blocking criteria. Thus, for a first pass of linkage, the files would be sorted by first initial and NYSIIS/Soundex code to bring together into the same "pocket" or "block" all records sharing the same NYSIIS/Soundex code and first initial. Records would only be compared within this block. Because a number of truly linked pairs of records would not be brought together on this basis (for example, because of a misrecording of first initial), a second pass could be carried out which blocks by date of birth. This second pass involves resorting the files on the basis of date of birth to create a second set of pockets or blocks within which comparison takes place. The results of the first and second passes need to be reconciled and this involves sorting the file yet again.

The key point is that standard methods of blocking involve sorting all the records involved in the linkage at least twice and usually more often. When faced with the kind of linkage mentioned above, involving linking a small number of newcomer records to a central catalog holding several millions of records such a procedure is at best immensely wasteful and at worst impossible. No matter how few newcomer records are involved, it is still necessary to sort all the central catalog records for the years of interest. If only a few years are involved, and especially if linkage is restricted to a subset of the central records e.g., cancer registrations, the exercise is feasible but immensely inefficient. If it is desired to link newcomer records to the entire data set, the exercise becomes, in reality, impossible.

One Pass Linkage: Blocking Without Sorting

The question thus became: how can we link a relatively small number of newcomer records to the catalog without having to repeatedly sort the catalog? The solution adopted has been to store the newcomer records in memory and carry out blocking using indexes based on numerical elements of the blocking criteria. The catalog records can then be read in sequentially and compared with all the newcomer records which fit the chosen blocking criteria (Kendrick and McIlroy, 1996).

The linkage is thus carried out in the course of "one pass" through the catalog data set.

Before they are brought into contact with the catalog, all newcomer records are read into memory and stored in an array indexed by a unique numeric record identifier. Necessary pre-processing such as generation of NYSIIS/Soundex codes is also carried out.

The next step is the creation of blocking index arrays. In this description we assume that two sets of

blocking criteria are being used: first initial and NYSIIS/Soundex code on the one hand, and date of birth on the other.

The blocking index arrays are indexed by numeric elements of the blocking criteria. Thus the first blocking array uses the numeric element of the NYSIIS/Soundex code as its index. All NYSIIS/Soundex codes consist of a letter followed by three figures e.g., A536 or B625. The first blocking index array has a row for each number from 001 to 999 which covers all possible numeric elements of NYSIIS/Soundex codes. In each row are stored the numeric identifiers of the newcomer records whose NYSIIS/Soundex code has the relevant numeric element. For example, the identifiers of newcomer records with surname FRAME (NYSIIS/Soundex code F650) and BROWN (B650) would be stored in the same row.

The second blocking index array has three indices: year, month and day of birth. Along the fourth dimension of the array are stored the numeric identifiers of the newcomer records sharing that date of birth.

Catalog records are then read in one by one. Suppose the first catalog record is for someone named BROWN (NYSIIS/Soundex code B650) with date of birth 1st March, 1922.

- Row 650 of the first blocking index array is inspected to see whether any newcomer records share the numeric element of the Soundex code. If any are found, then the first newcomer record is accessed via its numeric identifier in the newcomer record array. An immediate comparison is made of the first letter of the Soundex code and the first initial. If both match then we proceed to full probability matching between the catalog and newcomer records. If neither or only one match then no further action is taken. We then look at the next newcomer record (if any) indexed on the relevant row of the blocking index array.

- Blocking by date of birth is even easier to simulate in that the blocking criteria are entirely numeric. The catalog record can be directed to all newcomer records which share the same day (in this case 1); month (in this case 3); and year (in this case 22) by directly accessing the relevant array.

How the results of the ensuing pair comparisons are stored and implemented depends upon the structure and purpose of the linkage. Whenever links above a certain weight occur they can be output and stored for implementation in a provisional linkage file. This provisional linkage file can itself be flexibly interrogated to implement a given structure of linkage e.g., we may be only interested in the best link (the link with the highest weight) achieved by each newcomer record (see below).

One Pass Linkage: Practical Considerations

The above strategy, whereby the newcomer records can be indexed only in terms of the numeric elements of any blocking criteria, is necessary when we are using a programming environment which only allows numeric indexing of arrays. If either the newcomer data or the catalog data is stored in a database which allows direct access by any type of key then the logic of the exercise would be simplified. The file could be flexibly indexed by whatever blocking keys are felt appropriate.

Our impression at present would be that using memory still has advantages in terms of speed of access. This of course is a practical issue and may well change quickly as relational databases and "search engines" improve in speed and efficiency.

The number of newcomer records which can be linked in one pass through the data is of course limited by the available memory. Memory is needed both for storing the elements of the newcomer records which are necessary for linkage and for storing the blocking index arrays. For most ad hoc linkages involving anything up to 15,000 newcomer records this has not tended to be a problem. Larger newcomer data sets

can often be linked in sections without affecting the logic of the linkage.

Of the three elements of linkage it is this which is most dependent upon the capabilities of available hardware and software. The implication is that as these capabilities develop, there will be immense potential for moving beyond current limitations.

The Linkage Decision: Relative and Absolute Odds, Structuring the Linkage and the Best Link Principle

At the heart of the record linkage enterprise is the decision as to whether two records are truly linked. Most often the question is one of whether the records involved relate to the same person. The calculation of probability weights aims to provide a mathematical grounding for this decision.

However, it is a fundamental characteristic of the odds represented by probability weights that they are relative odds rather than absolute odds. They only serve to rank the pairs of records involved in a given linkage in order of the probability that they are truly linked. The relative odds do not represent fixed absolute or betting odds such that a probability weight of 25, for example, would always representing absolute or betting odds of 50/50. The conversion factor will vary from linkage to linkage.

It is absolute odds which are needed to inform the linkage decision. In practical terms the issue of the determinants of the conversion from relative odds to absolute odds can often be bypassed in that the required threshold in terms of absolute odds can usually be identified empirically from inspection of a sample of pairs. However a broad understanding of the relationship between relative and absolute odds is useful in that it can help optimise the way a given linkage should be structured.

Not all linkages require the same absolute odds. The absolute odds required depend upon the purpose of the linkage. They depend upon the costs associated with missing a true link compared with making a false link. For statistical purposes, the required absolute odds may be 50/50. If the linkage is to be used for administrative or patient contact purposes where a false linkage may have extremely damaging consequences, very high absolute odds may be required.

Relative to Absolute Odds: A Priori Factors

Two of the factors involving in converting relative to absolute odds take the form of relatively straightforward numerical principles. Newcombe has stated them in the context of a search file and a file being searched (Newcombe, 1988; Newcombe, 1995). The first principle is that the higher the proportion of records in the search file for which there exists a linked record in the file being searched, the more favorable

will be the conversion factor between relative and absolute odds. The second proposition is that the larger the file being searched, the less favourable will be the conversion factor between relative and absolute odds.

These factors are given for any specific linkage.

Relative to Absolute Odds: Structural Factors

However the conversion factor between relative and absolute odds can be influenced by how the linkage is implemented. It is important to design the linkage in a way which takes maximum advantage of the structures of the files involved and the relationships between the records in the files. For example, are the relationships between the records in two files one-to-one, one-to-many or many-to-many? How much confidence do we have in previous linkages which may have been carried out on the files involved? How confident are we that a file to be linked already contains only one record per person?

For example, if we want to link to each other a set of hospital discharge records, we have no a priori knowledge of how many records belong to each person. Our best bet is to do a conventional internal linkage and inspect all resulting pairs in setting a threshold. In this case we have relatively little leverage to improve the terms of conversion between relative odds and absolute odds.

If however we are linking a file of hospital discharge records to a file of death records we can obtain some "structural leverage." Death only occurs once and assuming that this is reflected in there being only one death record per person in the file of death records, the linkage becomes many-to-one. Each hospital discharge record should link to only one death record. The terms of conversion from relative to absolute odds can be improved by only retaining, for each hospital discharge record, the best (highest weight) link which is achieved to a death record (see also Winkler, 1994).

Similarly, at the other end of the life cycle, if we are linking baby records to mothers records, assuming that the mothers records themselves have been correctly linked we should allow each baby to link to only one mother. This in fact was the first context in which the importance of structuring the linkage emerged in Scotland.

Best Link and Structural Leverage: An Example

The CHI/NHSCR Linkage

These rather abstract considerations can be best understood in the context of a particular example. In common with the rest of the United Kingdom, Scotland is committed to the development of a unique patient identifier to help streamline the management of all patient contacts with the health service. Historically, Scotland has possessed two health-related registers of the Scottish population. It was felt that the combined strengths of the two registers would provide a firm basis for a new patient identifier.

For the last twenty years the Community Health Index (CHI) has operated on a regional basis as a primary care patient register for such purposes as screening for breast and cervical cancer and childhood immunisation. It contains a wealth of operational information with high population coverage. However, the regional indexes were initially compiled on an opportunistic basis and there was a general perception that there were gaps in its coverage and that there was a high proportion of duplicate records for people who had moved from one area of Scotland to another.

Scotland also possesses a National Health Service Central Register (NHSCR) which has been carefully maintained to contain one record for each resident of Scotland. The NHSCR however contains relatively

little operational information.

The initial plan was to carry out an internal linkage of the aggregated regional CHI indexes in order to remove duplicate records and then to link the resulting aggregated CHI data set to the NHSCR to form the basis of a national index.

However, based on our early experience of structuring linkages to maximise the power of the linkage, it was felt that linking the CHI databases to each other via a linkage to the NHSCR would provide more "leverage" in the linkage. (Kendrick et al., 1997)

The data which was available on both data sets to enable linkage was reasonable but not excessively rich. We had forenames, surnames, sex and date of birth but the only residential data was Health Board of residence (average size 300,000). A major bonus was that National Health Service number was available in a well formatted form on all NHSCR records. Because of its irregular format, the British NHS number has been notoriously difficult to use and was available on only a proportion of CHI records and with wide variations in accuracy and formatting.

Although the linkage was primarily concerned with "current" CHI records, those reflecting the current residence and GP registration of the Scottish population, "redundant" CHI records for people who had died or moved to a new Health Board were included in the linkage as a possible basis for constructing historical traces. In order to find a correct NHSCR "home" for as many CHI records as possible the NHSCR file also contained deaths from 1981 as well as known emigrants.

Since we were confident that the NHSCR did contain one record for every Scottish resident but there were suspicions that the CHI data set contained duplicate records (as well as legitimately multiple historical records), it was decided to structure the link as many-to-one. Each CHI record was allowed to link to only one NHSCR record -- the one with which it achieved the highest probability weight. Each NHSCR record on the other hand was allowed to link to as many CHI records as necessary.

Relative to Absolute Odds: Conversion Factors

The linkage can be described in terms of the factors which were outlined in the previous section as determining the conversion factor between relative and absolute odds.

- **Purpose of the linkage**. -- Any links accepted from the linkage would form the basis of patient contact. A very high level of confidence in the validity of any links was required. Missed links were regarded as less of a problem in that they would normally be picked up in the course of the running of the new index. Thus very high absolute odds for linkage were required.

- **A priori probabilities**. -- Given that both sets of data represented a high level of coverage of the Scottish population, there was a very high probability that a person represented on the CHI file would also be represented on the NHSCR file. In terms of Newcombe's first rule, circumstances could not have been more favourable.

- **File sizes**. -- Reflecting as they did the entire Scottish population as well as deaths and transfers these were large files: approximately 6.3 million NHSCR records against 7.8 million CHI records. This gives a high coincidence factor and, according to Newcombe's second rule, would normally serve to push up the relative odds required for given absolute odds.

- **Structuring the file**. -- Given the knowledge that all Scottish residents were likely to be represented by one NHSCR record and one or more CHI records, it made sense to structure the linkage as a

best link many-to-one linkage i.e. allowing each CHI record to link only to the NHSCR record with which it achieved best link would be the most effective route and would maximise the conversion factor between relative and absolute odds.

In broad terms then the linkage faced two difficult circumstances: the requirement for very high absolute odds and the large file sizes. These were more than outweighed however by the massive leverage contributed by the use of the best link principle in the context of a very high a priori probability that people were represented in both files.

Linkage Results

Of approximately 5,360,000 current registered CHI records, 4,600,000 or 86% linked deterministically to an NHSCR record. There was a match between the Soundex/NYSIIS code of surname, first initial, date of birth, sex and NHS number. For the remaining 750,000 CHI records, probability matching was carried out.

Resources for clerical checking were limited and such checking was limited to a sample of best link pairs to determine a probability weight which would represent absolute odds for the correctness of a linkage which were sufficiently high for administrative purposes. Staff of Health Board Primary Care Teams and the National Health Service Central Register checked 2,500 pairs using existing search and confirmation systems. No incorrect links were found at a probability weight greater than 30 and this was chosen as the administratively acceptable threshold.

To put this outcome into a broad comparative perspective we can compare the CHI/NHSCR linkage with previous linkages in Scotland which did not use the best link principle but which linked similar types of record using virtually the same agreement and disagreement weights for the main identifying items such as name and date of birth.

In the linkage of the Scottish hospital discharge and death record data sets using probability matching, the fifty/fifty threshold (i.e., the weight at which it is equally likely that the two records belong or do not belong to the same person) has remained relatively constant at a probability weight of 25. The fifty/fifty threshold for the best links of CHI to NHSCR records is around 15. Similarly, the threshold below which links between Scottish Cancer Registrations and death records are clerically checked and above which they are accepted automatically is a weight of 40. In the CHI/NHSCR linkage as we have seen, this threshold is 30. In both cases the difference is ten units in the currency of binit weights or logs to the base 2. In terms of odds this is an improvement in the conversion factor from relative to absolute odds of 2^{10} or around a thousandfold.

Why the use of only best links in this context should contribute so much extra leverage compared with a pure threshold method is perhaps intuitively obvious but is much more difficult to explain in principle. The logic is perhaps best illustrated by a hypothetical example.

Let us suppose that a CHI record on which is recorded the name Angus MacAllan with date of birth 25/01/1952 has achieved its best link with an NHSCR record on which is recorded the name Angus McAllan born 24/01/1951. There is no NHS number on the CHI record and no other elements agree so that the link achieves what would be, in the context of an unstructured purely threshold linkage, only a moderate probability weight implying a less than fifty/fifty chance that the records belong to the same person. We can best assess the likelihood that these two records would belong to the same person in the CHI/NHSCR linkage context by an indirect route. Let us imagine what would have to be true for the two records not to belong to the same person. Either:

- there is no NHSCR record relating to the individual represented on the CHI record and in addition there exists on the NHSCR file a record relating to another Angus Mc/MacAllan with a highly similar date of birth; or

- there is an NHSCR record corresponding to the individual represented on the CHI record but there are sufficient discrepancies in the recording of the identifying information for this "true link" Angus MacAllan that an NHSCR record for another Angus MacAllan in fact achieves a higher probability weight with the CHI record.

Neither of these scenarios are impossible but they are highly improbable and it is much more likely that the two records really do belong to the same person.

The method used had two additional advantages. The file which was output from the linkage took the form of a copy of each CHI record to which was appended an extract from the NHSCR record to which it had achieved the best link and the weight at which the link was achieved. This file was used as a basis for generating pairs for inspection and links could be extracted at whatever weights were necessary. In essence this means that the threshold for linkage was set and could be varied retrospectively without having to rerun the linkage.

The problem of twins has always bedevilled record linkage. The CHI/NHSCR linkage was able to take advantage of the fact that the NHS numbers for most pairs of twins are consecutive and a high negative weight was given for pairs of records with consecutive NHS numbers. Linkages using best link are in normal circumstances better than linkages using only a numeric threshold. In the presence of consecutive NHS numbers for twins the linkage was very successful in correctly allocating the records for twins.

One Pass Linkage and the Structuring of Linkages

Although one pass linkage and the structuring of linkages in terms of the best link have developed as separate responses to different challenges, they are not entirely independent.

Given that one of the main aims of one pass linkage is to avoid having to repeatedly sort or restructure the larger or target file, it is natural to implement one pass linkage as a best link procedure i.e., each newcomer record is allowed to link only to the catalog or target record with which it achieves the highest probability weight. Thus, it is not possible for the linkage to bring together records in the target file by "bridging" between them -- this would involve restructuring or resorting. As we saw earlier, as patient record sets in the main linked database grew larger, the false positive rate crept upwards, often because of illegitimate bridging by new records. As the main production linkages are adapted to one pass linkage, this problem will be minimised.

Although the affinity between one pass linkage and the best link principle is one of practical convenience, as we have seen, depending upon the circumstances of the linkage, the best link principle often has highly beneficial effects. Practicality and best practice often go hand in hand.

Linkage in Scotland: A Possible Future

Another way of looking at the CHI/NHSCR linkage is to see the NHSCR file as a target file at which the regional CHI files were aimed for linkage. As we have seen finding the best link record in the target file for each CHI record proved to have a dramatic effect on the accuracy of the linkage.

The much richer "national CHI" file which has resulted from the linkage and the introduction of national search and enquiry facilities provides an even better target for the linkage of other data sets.

For example, in November 1996 Scotland experienced a severe outbreak of infection by the E-coli 0157 bacterium. Several different sets of records were generated in the course of the outbreak: a case register, community clinic contacts, laboratory records, known exposed cohorts and hospital patients. The quality of identifying information on many of these records was rather poor reflecting the circumstances in which they were collected. ISD Scotland was asked to link these records so that the records for each individual involved in the outbreak could be gathered together. Rather than attempt to link the different sets of records directly to each other, the records were "aimed" at the local Community Health Index and linked to it. Again this method paid off in terms of much more accurate linkage.

It is likely that more and more linkages in Scotland will take the form of aiming data sets at the target of the national CHI. Ultimately the objective is to use such linkages, whereby for example laboratory data sets or hospital Master Patient Indexes are linked to the national CHI , to populate an increasing proportion of Scotland's health records with a unique patient identifier. It is intended that this will eventually reduce the need to record patient identification details such as names and dates of birth on operational records and communications. Instead identification will be via the national CHI number. Such a system is already in place in Tayside Health Board where the CHI number is implemented on a wide range of primary and acute health care records.

In this context the role of probability matching in the Scottish Health Service and the methods used to carry it out are likely to change even more rapidly over the next few years than they have over the last ten years.

As we have emphasised it has been the openness of record linkage in the Scottish Health Service to the demands of a wide range of customers which has driven the rapid development in our methods and this is likely to continue.

In this context the common sense and pragmatic approach to record linkage championed by Howard Newcombe has been especially useful and appropriate as guidance. Working as we are in his footsteps we can summarise some of the most salient emphases.

Record linkage is about being guided by the data and staying as close to the data as possible at all stages. The people who know the data best must be involved. Linkage is an evolutionary and recursive process at all levels. Linkage is a continual learning process and linkage is about what works, not what ought to work.

Finally, record linkage is not about the mechanical application of complex and abstract rules. As circumstances change and data sets vary there is unlikely ever to be one definitive best method of carrying out record linkage using probability matching. Progress will come rather from the flexible and responsive application of what are, at heart, very simple principles.

Acknowledgments

The following people have contributed over the years to the collective enterprise which is described. In ISD Scotland, James Boyd, Dorothy Gardner, Lena Henderson, Kevin McInneny, Margaret MacLeod, Fiona O'Brien, Chris Povey, Jack Vize, David Walsh, and Bruce Whyte have contributed their insight and programming skills. John Clarke and May Sleigh provided invaluable continuity and guidance. The expertise of Angela Bailey, Eileen Carmichael, Janey Read, and Maggi Reid in checking output has helped keep the system on course. In the Data Centre of the Scottish Health Service's Common Services Agency Gary Donaldson, Alison Jones, Ruth McIlroy, and Debbie McKenzie-Betts have laboured to put the linkage systems on a production basis.

References

Arellano, M.G. (1992). Comment on Newcombe et al.1(992). *Journal of the American Statistical Association*, 87, 1204-1206.

Fellegi, I.P. and Sunter, A.B., (1969). A Theory of Record Linkage, *Journal of the American Statistical Association*, 40, 1183-1210.

Gill, L.E. and Baldwin, J.A. (1987). Methods and Technology of Record Linkage: Some Practical Considerations, in *Textbook of Medical Record Linkage*, Baldwin J.A. et al. (eds), Oxford: Oxford University Press.

Gillespie, W.J.; Henry, D.A.; O'Connell, D.L.; Kendrick, S.W.; Juszczak, E.; McInneny, K.; and Derby, L. (1996). Development of Hematopoietic Cancers after Implantation of Total Joint Replacement, *Clinical Orthopaedics and Related Research*, 329S, S290-296.

Heasman, M.A. (1968). The Use of Record Linkage in Long-term Prospective Studies, in *Record Linkage in Medicine: Proceedings of the International Symposium, Oxford, July 1967,* Oxford: Oxford University Press.

Heasman, M.A. and Clarke, J.A. (1979). Medical Record Linkage in Scotland, *Health Bulletin (Edinburgh)*, 37: 97-103.

Hole, D.J.; Clarke, J.A.; Hawthorne, V.M.; and Murdoch, R.M. (1981). Cohort Follow-Up Using Computer Linkage with Routinely Collected Data, *Journal of Chronic Disease*, 34, 291-297.

Kendell, R.E.; Rennie, D.; Clarke, J.A.; and Dean, C. (1987). The Social and Obstetric Correlates of Psychiatric Admission in the Puerperium., in *Textbook of Medical Record Linkage*, Baldwin J.A. et al. (eds), Oxford: Oxford University Press.

Kendrick, S.W. and Clarke, J.A. (1993). The Scottish Medical Record Linkage System., *Health Bulletin (Edinburgh)*, 51, 72-79.

Kendrick, S.W. and McIlroy, R. (1996). One Pass Linkage: The Rapid Creation of Patient-Based Data, in *Proceedings of Healthcare Computing 1996: Current Perspectives in Healthcare Computing 1996*, Weybridge, Surrey: British Journal of Healthcare Computing Books.

Kendrick, S.W.; Douglas, M.M.; Gardner, D.; and Hucker, D. (1997). The Best-Link Principle in the Probability Matching of Population Data Sets: The Scottish Experience in Linking the Community Health Index to the National Health Service Central Register, *Methods of Information in Medicine* (in press).

Newcombe, H.B. (1988). *Handbook of Record Linkage*, Oxford: Oxford University Press.

Newcombe, H.B. (1995). Age-Related Bias in Probabilistic Death Searches Due to Neglect of the Prior Likelihoods, *Computers and Biomedical Research*, 28, 87-99.

Newcombe, H.B.; Kennedy, J.M.; Axford, S.J.; and James, A.P. (1959). Automatic Linkage of Vital Records, *Science*, 130, 954-959.

Newcombe, H.B.; Smith, M.E.; and Lalonde, P. (1986). Computerised Record Linkage in Health Research: An Overview, in *Proceedings of the Workshop on Computerised Linkage in Health Research (Ottawa, Ontario, May 21-23, 1986)*, Howe, G.R. and Spasoff, R.A. (eds), Toronto: University of Toronto Express.

Newcombe, H.B.; Fair, M.E.; and Lalonde, P. (1992). The Use of Names for Linking Personal Records, *Journal of the American Statistical Association*, 87, 1193-1204.

West of Scotland Coronary Prevention Study Group (1995). Computerised Record Linkage Compared with Traditional Patient Follow-up Methods in Clinical Trials and Illustrated in a Prospective Epidemiological Study, *Journal of Clinical Epidemiology*, 48, 1441-1452.

Winkler, W.E. (1994). *Advanced Methods for Record Linkage*, Statistical Research Division, Statistical Research Report Series No. RR94/05, Washington D.C.: U.S. Bureau of the Census.

Chapter 11

The Use of Names for Linking Personal Records

Howard B. Newcombe, Consultant
Martha E. Fair and Pierre Lalonde, Statistics Canada

The skill of a human who searches large files of personal records depends much on prior knowledge of how the names vary in successive documents pertaining to the same individuals (e.g., as with ANTHONY–TONY, JOSEPH–JOE, WILLIAM–BILL). Now, an essentially exact procedure enables computers to make similar use of an accumulated memory of their own past experiences when searching for, and linking, records that relate to particular persons. This knowledge is further applied to quantify the benefits from various refinements of the rules by which the discriminating powers of names are calculated when they do not precisely agree or are substantially dissimilar. Of the six refinements tested, by far the most important is the recently developed exact approach for calculating the ODDS associated with comparisons of names that are possible synonyms.

KEY WORDS: Data base maintenance; File searching; Probabilistic linkage; Quantitative judgment; Record linkage.

Personal documentation in machine-readable form has become so extensive in any advanced society as to constitute, collectively, a detailed but highly fragmented life history for virtually all its members. The files exist to serve the needs of people and of society as a whole, and frequent access is involved. Much of the searching is necessarily based on names and personal particulars that are apt to be reported differently on successive documents for the same individuals. The problems are familiar to clerks, but now access by computer is becoming the norm.

With automated searching, many choices are possible between *refinements* and *simplifications* in the way that names get compared. Rarely, however, have the merits of alternative approaches been quantified in terms of gains or losses of discriminating power, so as to reduce the guesswork when designing a system. The potential for sophistication in automated comparisons of names is substantial. Humans develop special skills in recognizing nicknames, ethnic variants, diminutives, and corrupted forms due to truncations, misspellings, and typographical errors. This is known to be based on a relatively simple rationale, supported by remembered data. If a machine is to acquire similar ability, it too must rely on past experience (Newcombe, Fair, and Lalonde 1989; Newcombe, Kennedy, Axford, and James 1959). Although there is now an essentially exact way of measuring the discriminating powers of comparison pairs like CARL–KARL, GEORGE–GYORGY, JACOB–JAKE, JOHN–JACK, and WILLIAM–BILL, much clerical labor and large amounts of data are needed to set it up (Fair, Lalonde, and Newcombe 1990, 1991; Newcombe et al. 1989). Simpler comparisons are, therefore, likely to remain popular in many procedures that use names to access files.

Whether or not this exact approach becomes widely applied, its existence now provides a convenient standard against which to judge the performance of other treatments of names. So we have used the approach in this article to quantify the gains and losses of discriminating power due to various *refinements* and *shortcuts* commonly used in automated searching and linkage.

* Howard B. Newcombe is a consultant, P.O. Box 135, Deep River, Ontario K0J 1P0, Canada. Martha E. Fair is Chief and Pierre Lalonde is Project Manager, Occupational and Environmental Health Research Section, Canadian Centre for Health Information, Statistics Canada, Ottawa, Ontario K1A 0T6, Canada. The authors thank John Armstrong, Michael Eagen, and William E. Winkler for helpful critical comments on an early version of this article, and also the associate editors and referees who substantially influenced its final form.

The test is special to names as identifiers; is suitable for fine-tuning this component of a record linkage system; and is uninfluenced by the adequacy of the rest of the identifiers. It differs from, but is complementary to, more direct tests of overall performance.

1. COMPUTER LINKAGE

Where a computer is used to search large files of personal records and bring together the records for particular individuals, it may emulate with varying degrees of success the strategies of a human clerk who does the same job. To determine whether a pair of records is correctly matched, the names are compared along with other identifiers (e.g., year, month, and day of birth; sex and marital status; and various geographic particulars such as place of birth, residence, work, or death). Sometimes, however, these comparisons point in different directions.

The problem then is to determine, as in a court of law, where the preponderance of the evidence lies. The comparisons must be considered not only separately but also in combination. A particular comparison outcome (e.g., JOHN–JOHN or JOHN–JACK) will argue *for* linkage when it is more common among correctly matched pairs than among random false matches. Conversely (as with JOHN–JOE), an outcome will argue *against* linkage when the opposite is the case. These *likelihood ratios* (or individual ODDS in favor of linkage) may be combined to assess the collective evidence from the full set. But this is not the whole of the relevant information.

In addition, a human clerk may recognize two further factors: the *size* of the file being searched and the *likelihood* that the individual is represented in it. Thus, when looking for a particular JOHN BROWN in the telephone directory for a small town where he is thought to reside, finding the name suggests that it may well belong to the right person. This would definitely *not* be so when searching a large national death register, especially if this JOHN BROWN were unlikely to have died.

Automated searches have from the outset used much the same reasoning as does a human clerk; this provides numerous options when calculating the ODDS for particular

Reprinted from: © 1992 American Statistical Association
Journal of the American Statistical Association
December 1992, Vol. 87, No. 420, Review Paper
Reprinted with permission.

theory alone (Newcombe and Kennedy 1962; Newcombe et al. 1987). (The need for practice and theory to complement each other is discussed elsewhere; see Scheuren, Alvey, and Kilss 1986; Winkler 1989b.)

When Statistics Canada first actually used probabilistic linkage in the early 1980s, based on the Fellegi–Sunter theory, it was to search the newly established Canadian mortality data base, which extended back to 1950. Their linkage system, known as CANLINK or GIRLS (for Generalized Iterative Record Linkage System) included innovations described by Howe and Lindsay (1981), Hill (1981), Hill and Pring-Mill (1985). In particular, a *preliminary linkage* step was introduced that temporarily ignored specific values of names, thereby eliminating in a simple fashion many unpromising record pairs, and an *iterative update* of the outcome frequencies from LINKED pairs of records was used. The preliminary step was needed because the death files were now blocked by just a single surname (as a NYSIIS phonetic code; see Appendix H of Newcombe 1988), and the blocks were larger than those based on pairs of surnames for family linkage. The iterative updates were required because to get new linkage jobs started, outcome frequencies from earlier linkages were often used initially and replaced later with increasingly appropriate data as the new files of LINKS were progressively improved. (The effect of omitting this update is considered in Sec. 3.4.) A further intended refinement, recognition of *partial agreements* of names (like THOMAS–TOM), was less successful; as a result, modified procedures had to be devised (Eagen and Hill 1987; Fair et al. 1990, 1991; Newcombe 1988; Newcombe et al. 1987, 1989; Winkler 1985, 1989a.) The matter is referred to again in Section 3.5.

Howe and Lindsay (1981) also recognized explicitly, for the first time, the concept of the *prior odds* or *prior likelihood* but failed to apply it to create a scale of *absolute* ODDS that might be used for setting thresholds. Earlier, two thresholds had been proposed as part of the Fellegi–Sunter theory to distinguish *positive links* and *positive nonlinks,* plus an intermediate category of ambiguous matches called *possible links.* The thresholds were to be calculated in advance as "error bounds" that would limit the numbers of false-positive and false-negative links and would identify pairs in need of special assessment. But when the ODDS from the full sets of identifiers were combined, it was found that the resulting overall ODDS served only to array the record pairs, *relative to one another,* in descending order of the likelihood of a correct match. Thus, in practice, the two thresholds got assigned subjectively. On the scale of *relative* ODDS available at the time, they fell high above the crossover or 50/50 odds point (e.g., in the case of the death searches by a factor of well over 1 million, and greater than the size of the file being searched).

An empirical conversion to a scale of presumed *absolute* ODDS indicated why. When allowance was made for the size of the death file, $1/N$(File B), and for the proportion of search records that find a matching death record in it, $N(A \mid \text{LINK})/N$(File A), the new scale brought the subjective thresholds close to the crossover or 50/50 odds point. Together, these two factors were taken to represent the prior likelihood of a correct match on a single random pairing (i.e., before examining any identifier or blocking information).

The new scale of absolute ODDS was controversial at first, although the results were consistently believable over many empirical tests, whereas those from the alternative were not. Later, it was shown to use just a variant of the prior odds, $P(\text{LINK})/P(\text{NONLINK})$, already recognized by Howe and Lindsay (1981). The implications are substantial but were not explored by those authors (see Secs. 2.3 and 3.1 and Fig. 1). In practice, however, it was soon found that the concept of the prior likelihood could be applied with great flexibility in many ways. For example, as a refinement it was calculated separately for subsets with differing prior likelihoods (see Newcombe 1988, chap. 28 and apps. B and D.3).

What refining the practice achieved, as distinct from formal theory, was enhanced flexibility in the access to discriminating power. *Individual identifiers* were compared freely, just as a human might do when seeking clues to the true linkage status of a record pair; and the prior likelihood of a correct match, in the case of a death search, was exploited to take into account the age of the individual in a given year, and the actuarial likelihood that he or she might have died in that year. For linkages of cancer records with death files, the approach even used survival curves appropriate to particular diagnoses. The practices are fully described, but in nontechnical language for those working close to the files, who design, implement, and test the detailed procedures (see, for example, Newcombe 1988, sec. 28.2 and apps. D.2 and D.3).

This is the technological setting within which the current study has been carried out.

1.2 General Method

Any formal statement of the comparison procedure for individual identifiers should allow for the flexibility that exists in practice. This is especially true of names when they do not precisely agree (e.g., as allowing recognition of the comparison DANIEL–DANNY). Moreover, because some kind of grouping of possible synonyms is inevitable, this too must be exceedingly flexible if discriminating power is not to be wasted (Scheuren 1985). We will deal first with formal expressions that permit flexibility when estimating likelihood ratios (or ODDS in favor of linkage as indicated by particular comparisons), and second with grouping under conditions of minimum constraints. (Other accounts use logarithms of the likelihood ratios and refer to them as "weights." The ratios may also be viewed as factors by which comparisons of particular identifiers raise or lower the overall "betting odds" in favor of linkage.)

Conceptually, each first given name on one file is compared with every first given name on the other file, and second given names are likewise compared. Generally, LINKED pairs (of names or records) are vastly outnumbered by *possible* NONLINKED pairs, i.e., actual plus potential. (This concept is fundamental and is not altered by "blocking" that reduces the *actual* numbers of comparison pairs; see Fellegi 1985.) Although LINKS and NONLINKS are thought of as uncon-

taminated with pairs of the opposite kind, modest admixtures have only slight effects on the ODDS.

When comparing value A_x from a Record A (which is used to initiate a search) with value B_y from a Record B (which is in the file being searched), the ODDS in favor of a correct LINK associated with outcome $A_x \cdot B_y$ (i.e., the comparison pair of values) may be written in terms of the relative probability of occurrence of the particular outcome in LINKS as compared with NONLINKS; that is,

$$\text{ODDS} = P(A_x \cdot B_y \mid \text{LINK})/P(A_x \cdot B_y \mid \text{NONLINK}). \quad (1.1)$$

But except where files A and B are both very small, the denominator in this expression will be closely approximated by $P(A_x) \cdot P(B_y)$, because any fortuitous LINKS in the random pairs will be vastly outnumbered by the NONLINKS. Thus the expression may be converted to

$$\text{ODDS} = P(A_x \cdot B_y \mid \text{LINK})/P(A_x) \cdot P(B_y). \quad (1.2)$$

This implies that we need to know in advance the number of LINKS with values A_x and B_y. In practice crude approximations are estimated initially from sample linkages carried out manually or from previous linkage studies and are revised iteratively as the current LINKS are progressively refined.

An expanded form of this procedure is sometimes used to support an existing practice in the case of death searches. This involves ignoring the frequency of value A_x, both in File A and in the LINKS, on the grounds that names are unlikely to be strongly correlated with the probability of death and with whether a Record A is LINKED to a Record B. Justification depends on the magnitude of the error introduced by the assumption. The expanded version has two parts:

$$\text{ODDS} = \underbrace{\frac{P(B_y \mid A_x \cdot \text{LINK})}{P(B_y)}}_{\substack{\text{SIMPLIFIED} \\ \text{FORMULA}}} \cdot \underbrace{\frac{P(A_x \mid \text{LINK})}{P(A_x)}}_{\substack{\text{CORRECTION} \\ \text{FACTOR}}}. \quad (1.3)$$

Current practice views the second part (the "correction factor") as approximating unity, so it can be ignored, except where the assumption is thought to be seriously misleading (as it might be if ethnicity and ethnic names were correlated with mortality).

What the relative probabilities fail to do is indicate explicitly how the ODDS should be calculated using data that are in short supply. Examples include outcome values $A_x \cdot B_y$ that are represented only once or twice in an available real file of LINKS and, especially, numerous other outcome values representing pairs of *possible synonyms* that have not actually occurred in the available LINKS but probably would occur if that file were larger. Because crucial steps in the reasoning have to do with numbers of outcome values, as distinct from their likelihoods, it is helpful to convert the last two expressions to a form actually used to obtain *estimated* relative probabilities, as

$$\text{ODDS} = \frac{N(A_x \cdot B_y \mid \text{LINK})/N(\text{LINKS})}{N(A_x \cdot B_y \mid \text{NONLINK})/N(\text{NONLINKS})} \quad (1.4)$$

and

$$\text{ODDS} = \underbrace{\frac{N(A_x \cdot B_y \mid \text{LINK})/N(A_x \mid \text{LINK})}{N(B_y)/N(B)}}_{\substack{\text{SIMPLIFIED} \\ \text{FORMULA}}} \times \underbrace{\frac{N(A_x \mid \text{LINK})/N(\text{LINKS})}{N(A_x)/N(A)}}_{\substack{\text{CORRECTION} \\ \text{FACTOR}}}, \quad (1.5)$$

where the general term $N(* \mid \text{LINK})$ represents the number of records among LINKED pairs that have attribute (*), $N(\text{LINKS})$ = number of linked pairs, $N(A)$ = number of records in File A, $N(B)$ = number of records in File B, $N(A_x)$ = number of records in File A with value x, and $N(B_y)$ = number of records in File B with value y. (For the origins of this version, see Newcombe et al. 1989.)

It is convenient to retain the distinction between a search file (File A) and a file being searched (File B), even though conceptually the roles could be reversed. For one thing, the search file usually is smaller than the file being searched. Also, the distinction has special significance for the death searches, because informal versions of a given name (e.g., nicknames) are more commonly used by employers and others while one is alive rather than by undertakers after one has died.

Here we need to introduce two concepts related to the ways in which the range of possible outcomes may be handled:

1. Grouping or "pooling" of similar values of $A_x \cdot B_y$, which individually are represented poorly or not at all in the available LINKS (the "quantity" problem)
2. Increasing sacrifice of discrimination as the within-group heterogeneity grows when its definition is broadened to ensure representation in the LINKS (the "quality" problem).

A tradeoff between "quantity" and "quality" is unavoidable. The definition of an outcome group needs to be broad enough so that $N(A_x \cdot B_y \mid \text{LINK})$ is represented by at least one comparison pair. Otherwise, no ODDS can be calculated. But because the definition is widened to increase the representation, it will also let more heterogeneity into the group. (Thus as the error due to statistical fluctuation diminishes, so the error due to lessened specificity increases.)

The earliest linkage operations simplified matters by recognizing just two categories of outcome—*agreements* and *disagreements*—and by attributing specificity for value only to the former category. But major errors arose from an unsuccessful attempt to adapt the earlier procedures, to recognize "partial agreements" such as JOSEPH–JOE (Newcombe et al. 1987). (The term "partial agreement" is commonly applied, for reasons of convenience, to any possible synonyms regardless of similarity, as with ELIZABETH–BETTY.)

The problem posed by the value-specific partial agreements of names may be handled in various ways, but only one of

these appears to be precise. A compromise solution, now in routine use, is based on the numbers of early characters that agree. ODDS are first calculated for different levels of agreement (i.e., one, two, three, four or more agree); actual values are ignored at this stage. Such "global ODDS" are later adjusted upward or downward, depending on whether the particular values of the *agreement portions* are rare or common (Eagen and Hill 1987; Newcombe 1988; Newcombe et al. 1987), but this neglects the values of the *disagreement portions* (e.g., it wrongly treats diverse name pairs like JOHN–JONATHAN and JOHN–JOSEPH as equally likely to be synonyms). An alternative approach that recognizes phonetic components common to the two names has also been developed (Winkler 1985, 1989a).

A precise treatment of partial agreements of names recognizes both values in a comparison pair and avoids resorting to globally defined (i.e., value-nonspecific) levels of agreement. This permits it to deal with outwardly dissimilar comparison pairs (e.g., EDWARD–TED, MARGARET–PEGGY). Any necessary groupings must be defined in value-specific ways. The frequency with which the two values are related by actual usage then determines the magnitude of the precise ODDS. A modest manual test showed that the approach worked where sufficient data from LINKED pairs of records could be made available (Newcombe et al. 1989). That was followed by an expanded application based on an accumulated composite file of LINKS from many past searches of the Canadian mortality data base (Fair et al. 1990, 1991). This refinement will be considered further in Section 3.5.

(The current emphasis on flexibility also extends to other identifiers that are apt to be reported differently on separate occasions or that may change over time, as with MARITAL STATUS, OCCUPATION, INDUSTRY, and PLACES OF RESIDENCE, WORK, and DEATH. For these, there likewise is no need to prejudge in which direction the comparisons will argue. "Agreement" and "disagreement" are often poor indicators, but the ODDS—when they have been calculated—will decide.)

1.3 Combining the ODDS

When the likelihood ratios or ODDS for particular identifiers are combined over the full set in a record pair, it is usual to assume as a tolerable approximation that the identifiers are independent of one another. The overall *absolute* ODDS (in the sense of "betting odds" in favor of linkage) may then be represented by

$$\text{Absolute ODDS} = R_1 \cdot R_2 \cdot \cdots \cdot R_n \cdot P(\text{LINK}), \quad (1.6)$$

where R_1 to R_n are the likelihood ratios (ODDS) for identifiers 1 to n (including any used for blocking) and are independent of each other, and $P(\text{LINK})$ is the prior likelihood of a correct match on a singly random pairing. The latter term is similar to the *prior odds*, $P(\text{LINK})/P(\text{NONLINK})$, recognized but not used by Howe and Lindsay (1981). Confusion remains concerning the implications, and is not explicitly addressed by existing formal theory (see Sec. 2.1).

The version of this expression used to calculate *estimated* absolute ODDS from actual counts is unfamiliar to many, so it is necessary to be explicit: R_1 to R_n become frequency ratios, and $P(\text{LINK})$ becomes $N(\text{LINKS})/N(\text{LINKS} + \text{NONLINKS})$. Because each linked pair contains one record from File A and one from File B, $N(\text{LINKS}) = N(A|\text{LINK}) = N(B|\text{LINK})$. Also, where each record on File A is compared in succession with every record on File B, the total number of comparison pairs, regardless of their linkage status, will together equal the product of the two file sizes; that is, $N(\text{LINKS} + \text{NONLINKS}) = N(\text{File A}) \cdot N(\text{File B})$. The concept is valid even where, in practice, only the pairings that occur within blocks are actually seen; but this implies that likelihood ratios for blocking identifiers will be taken into account. Thus by substitution we may obtain

$$\text{Absolute ODDS} = R_1 \cdot R_2 \cdot \cdots \cdot R_n \cdot \frac{N(A \mid \text{LINK})}{N(\text{File A})} \cdot \frac{1}{N(\text{File B})}. \quad (1.7)$$

Howe and Lindsay (1981) had felt that their prior odds, $P(\text{LINK})/P(\text{NONLINK})$, could not be readily estimated. The solution came to us by observing *human stratagems* and through reasoning based on *counts* rather than on *probabilities*. At first, it was hard to persuade others that this practice is valid, perhaps because our way of thinking was unconventional (David Binder and Geoffrey Howe, personal communication, November 10 to December 11, 1982). A further possible reason might be the common custom of *not* calculating frequency ratios for blocking identifiers; but then *NA* and *NB* would represent the sizes of Files A and B *within* the particular block, and the prior likelihoods would differ from block to block.

Calculation (1.7) has been used over the past decade for searches of Canadian death files. The application is exceedingly flexible and allows refinement through redefinition of Files A and B to represent, separately, a multiplicity of subsets (based on age, death year, selected diagnoses, and so on) of populations that are internally heterogeneous. (For details, see Newcombe 1988 chap. 28 and apps. B and D.2.)

2. EMPIRICAL DISTRIBUTIONS OF LINKS AND NONLINKS

A feedback of empirical data from the LINKS and NONLINKS is the most basic requirement of a linkage system. For example, the expressions by which the ODDS for the individual identifiers are calculated require these data as input. Also, such data are needed when assessing errors due to assumptions that are not strictly correct.

Above all, direct observation of individual record pairs often yields clues to more suitable comparison steps. These clues are most likely to become apparent to humans when resolving difficult matches manually. An experienced person can be less bound by artificial constraints than the automated system, and he or she is still, given existing linkage systems, in a better position to be guided by memories of past encounters with similar problems.

Theoretical papers on linkage make strong assumptions to get results, and linkage practice does the same to simplify

procedures. Examples include the use of artificially simplified ways of comparing names, which may not adequately exploit their true discriminating power, and the practice of simply multiplying the ODDS for individual identifiers to combine them for a whole set, which would be strictly proper only if they were independent of each other (Fellegi and Sunter 1969; Howe and Lindsay 1981).

Only with better data from LINKS and NONLINKS can many of the uncertainties be resolved. Recognition of this has led, in part, to the idea of accumulating large files of LINKS and creating even larger files of NONLINKS (see, for example, Fair et al. 1990, 1991; Lalonde 1989; Newcombe et al. 1989). It has also emphasized the use of additional evidence on the *true linkage status* of record pairs assigned borderline absolute ODDS in an automated operation (Fair, Newcombe, and Lalonde 1988a; Fair, Newcombe, Lalonde, and Poliquin 1988b).

We will deal first with the latter point.

2.1 The Assumption of Independence

Calculated overall "absolute ODDS" usually assume that the components in the identifier sets are independent of each other. Rarely is this assumption strictly correct. It can be seen to be misleading when scanning visually for record pairs that were wrongly classed as positive LINKS and positive NONLINKS. Our unpublished observations include examples of *multiple agreements* (e.g., of rare ethnic names and related places of birth) that have spuriously raised the ODDS to create false positives. Conversely, there are examples of *multiple disagreements* (especially on year, month, and day of birth—perhaps due to multiple wrong guesses by an informant at the time of a death), which have spuriously lowered the ODDS to create false negatives.

The effects of these and other such biases are best visualized in the overlap between the numbers of verified LINKS and NONLINKS, when distributed along a scale of absolute ODDS that assumes independence, as in Figure 1 (data of Fair et al. 1988a, 1988b; and Lalonde 1986). We will refer to points on this scale as "theoretical" ODDS to distinguish them from the "empirical" ODDS, which are the ratios of observed counts of LINKS/NONLINKS at various points on the same scale. (Total LINKS and NONLINKS are not shown in the Figure; but conceptually the latter vastly outnumber the former.)

In practice there is no need to actually create the bulk of the possible NONLINKS, because most would fall so very low on the scale. Major misunderstanding arises, however, when the enormous preponderance of actual plus potential NONLINKS over LINKS is not kept in mind. Thus the distributions and their crossover points serve little purpose if

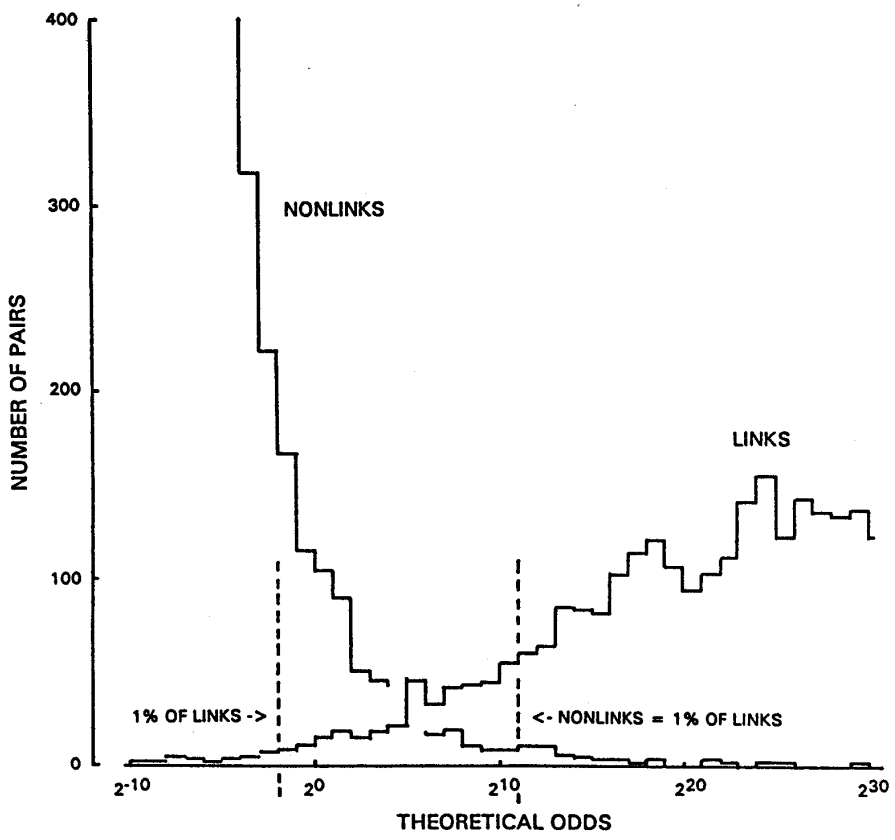


Figure 1. Overlapping Parts of the Distributions of LINKS and NONLINKS, on a Scale of Theoretical ODDS (Lalonde 1986). Note that empirical error bounds (broken lines), set at the 1% levels, are displaced upward on the theoretical scale.

plotted as proportions of LINKS compared with proportions of NONLINKS. Likewise, upper and lower "error bounds," when expressed in such terms, make nonsense of the concept. (The data in Figure 1 are from searches of 1,300,000 death records, initiated by 30,000 work records, yielding 2254 LINKS; the vital status of doubtful pairs was confirmed using taxation files. Because the number of possible pairings, i.e., actual plus potential, is the product of the two file sizes, NONLINKS outnumber LINKS by 17,000,000 to 1.)

Marked discrepancies are revealed in Figure 1 between the theoretical ODDS scale, based on the assumption of independence, and the corresponding observed ratios of LINKS versus NONLINKS. For example, where the theory indicates that the ODDS in favor of linkage are 1/1, in reality they are only 1/6; and where the observed ODDS are 1/1, the theory says that they should be 16/1. Moreover, if one wants to set lower and upper thresholds to limit the number of LINKS wrongly classed as "positive nonlinks" to 1% of all LINKS, and to likewise limit the NONLINKS wrongly classed as "positive links" to a similar *number* (i.e., 1% of the LINKS), the correct thresholds would be represented by theoretical ODDS of approximately 1/4 and 2,000/1. Thus the true error bounds are displaced upward on a scale of ODDS that assumes independence.

There has been confusion in the past, which is best avoided by thinking in terms of numbers (i.e., *counts*) as distinct from proportions. One does *not* limit false positives to 1% of NONLINKS because, in our example, that would create 17,000,000 times as many false positives as false negatives. Indeed, the Fellegi-Sunter theory emphasizes that NONLINKS typically will greatly outnumber LINKS; for example, see slides #9 and #10 of Fellegi 1985. More explicitly, where this is the case "no one could possibly conclude" that the two error bounds would be properly set at equal proportions (i.e., 1%) of the LINKS and of the NONLINKS (I. P. Fellegi, personal communication July 8, 1987).

2.2 Data on Name Comparisons Involving Synonyms

Value-specific information to do with $N(A_x \cdot B_y \mid \text{LINK})$, heretofore lacking in quantity, is contained in a composite file of 64,937 LINKED pairs of male given names derived from 26 linkage projects. All of the projects involved searches of the Canadian Mortality Data Base (File B, containing 3,397,860 male given names), initiated by records of various study cohorts, including employment records, survey responses, cancer registrations, birth records, and entries in a national radiation dose register (composite File A, containing

Table 1. Common Male Given Names From the Canadian Death File, 1950–1977

		Total observed				Total observed	
Rank	Name*	Number	Percent	Rank	Name*	Number	Percent
Formal Names							
1	JOHN	187,486	5.30	26	JEAN (male)	22,661	.64
2	WILLIAM	170,669	4.83	27	FRANCIS	21,596	.61
3	JAMES	111,513	3.16	28	HAROLD	21,588	.61
4	JOSEPH	104,767	2.96	29	GORDON	19,158	.54
5	GEORGE	95,188	2.69	30	HERBERT	19,133	.54
6	CHARLES	70,040	1.98	31	SAMUEL	18,927	.54
7	ROBERT	66,575	1.88	32	ANDREW	18,440	.52
8	THOMAS	64,182	1.82	33	DONALD	17,416	.49
9	HENRY	55,718	1.61	34	DANIEL	16,076	.46
10	EDWARD	55,837	1.58	35	STANLEY	14,575	.41
11	ARTHUR	52,221	1.48	36	PATRICK	13,402	.38
12	ALBERT	47,660	1.35	37	NORMAN	13,270	.38
13	ALEXAND(ER)	38,343	1.09	38	ROY	12,943	.37
14	FREDERI(CK)	36,864	1.04	39	RAYMOND	12,338	.35
15	DAVID	33,530	.95	40	EMILE	12,261	.35
16	ERNEST	32,041	.91	41	HENRI	12,107	.34
17	ALFRED	30,902	.87	42	KENNETH	12,076	.34
18	FRANK	29,376	.83	43	DOUGLAS	11,843	.34
19	PAUL	26,919	.76	44	LEONARD	10,978	.31
20	PETER	26,889	.76	45	EUGENE	10,968	.31
21	WALTER	26,718	.76	46	VICTOR	10,797	.31
22	HARRY	24,830	.70	47	GEORGES	10,446	.30
23	MICHAEL	24,645	.70	48	ALLAN	10,384	.29
24	RICHARD	24,070	.68	49	LEO	10,200	.30
25	LOUIS	23,860	.68	50	EDWIN	10,156	.29
				51	CLARENC(E)	9,974	.28
Informal Variants							
1	FRED	7,947	.23	8	JOE	866	.025
2	JACK	5,575	.16	9	DAN	781	.023
3	ALEX	3,550	.10	10	BILL	314	.009
4	MIKE	3,267	.10	11	PETE	265	.008
5	SAM	2,014	.06	12	DON	240	.007
6	RAY	1,911	.056	13	ANDY	220	.006
7	TOM	990	.029	14	DAVE	179	.005
				15	ED	43	.001

* Truncated at seven characters in the records of the Canadian mortality data base.

Table 2. Pooling of Synonyms in Value-Specific Groups: Example Based on CHARLES Compared with KARL and Related Variants

Value of name	Numbers in File B*	Value of name	Numbers in File B*	Value of name	Numbers in File B*
KARL	3,002	KARLIOU	1	KARLS	2
KARLA	1	KARLIS	82	KARLSEN	2
KARLDON	1	KARLMER	1	KARLSON	2
KARLE	6	KARLO	36	KARLSSO	1
KARLEY	1	KARLOFF	1	KARLTON	1
KARLHEI	2	KARLOL	1	KARLY	2
KARLIE	1	KARLOS	1		

* Based on an alphabetic listing from the death file. Of these names, only KARL was actually interchanged with CHARLES in the linked pairs of records. However, the other potential combinations with CHARLES cannot be classed as full disagreements.

Table 4. Examples of Partial Agreements That Are Not Well Represented

Values*		Total observed	Values*		Total observed
x	y		x	y	
ALBERT	-ALBERTO	1	ALBERT	-ALBERTS	0
ARTHUR	-ARTIMUS	1	ARTHUR	-ARTIMON	0
DOUGLAS	-DOUGLES	1	DOUGLAS	-DOUGLIS	0
ERNEST	-ERNES	1	ERNEST	-ERNE	0
HAROLD	-HARLOD	1	HAROLD	-HARLOE	0
LEO	-LEODA	1	LEO	-LEODAS	0
PETER	-PEDER	1	PETER	-PEDAR	0
VICTOR	-VIATEUR	1	VICTOR	-VIATIAR	0

* Truncated at seven characters in the files of Fair et al. (Fair, Lalonde, and Newcombe (1991)). The synonyms are all represented in the parent files A and B.

1,574,661 male given names). (For details, see Fair et al. 1990, 1991.)

The data used in the current study are from the LINKED pairs of names containing any of the 51 most common given names in the death file or any of the 15 most common informal variants. These names are listed in Table 1, together with their counts and percentage frequencies in the death file.

The 51 common names account for more than half (1,842,327/3,397,860) of all given names in the death records of males. Among 64,937 LINKED pairs of male given names, they were present 33,183 times on the Records A (25,673 as first names and 7,510 as second names) and 33,988 times on the Records B (26,536 as first names and 7,452 as second names), for a total of 67,171 times. A name pair that partially agrees may occur in either of two configurations, e.g., as FRANK-FRANCIS or as FRANCIS-FRANK, depending on which value comes from File A and which value comes from File B. Where two or more of the 51 names get interchanged with each other (as happens with HARRY, HENRI, and HENRY), some of the same information may be duplicated in a slightly different form within the tables.

The 15 common informal variants represent less than 1% (28,164/3,397,860) of all given names in the death records of males. Among the 64,937 LINKED pairs of male given names, these were present 1,554 times on the Records A (1,426 as first names and 128 as second names) and 701 times on the Records B (633 as first names and 68 as second names), for a total of 2,255 times.

Table 3. Examples of Partial Agreements That Are Well Represented

	Values*		Numbers observed		
Rank	x	y	Total	$N(A_x \cdot B_y \mid LINK)$	$N(B_x \cdot A_y \mid LINK)$
1.	MICHAEL	-MIKE	173	12	161
2.	FREDERI	-FRED	169	12	157
3.	ALEXAND	-ALEX	152	11	141
4.	JOHN	-JACK	90	23	67
5.	FRANCIS	-FRANK	73	19	54
6.	JOSEPH	-JOE	62	2	60
7.	FREDERI	-FREDRIC	52	28	24
8.	ALLAN	-ALLEN	47	28	19
9.	HENRY	-HENRI	44	40	4
10.	SAMUEL	-SAM	37	3	34
11.	PETER	-PETE	33	3	30
12.	THOMAS	-TOM	33	7	26
13.	WILLIAM	-WILLI	20	18	2

* Truncated at seven characters in the LINKS of Fair et al. (1991).

Application of the linkage rationale to outcomes defined in wholly value-specific ways depends on more than just the ODDS formula for its success. The chief obstacle is created by the many value pairs that are rare in the available LINKS, plus the even more numerous possible ones that have not been observed at all. Grouping is necessary, but must be based on wholly value-specific group definitions. The roles played in the process by Files A and B and the LINKS are illustrated in Tables 2–5. Group definitions are based on selected blocks of names in alphabetic listings, chosen to bring rare synonyms into the same groups with common forms (Table 2). Comparison pairs that are common in the LINKS present no special problem (Table 3). However, possible pairs that are rare or absent in the available LINKS need to be grouped with others that are more common (Table 4). ODDS are calculated for specific name pairs and for specific groups as a whole, using expression 1.4 (Table 5). (For details see Fair et al. 1990, 1991.)

There are no rules explicitly stating how the boundaries of the groups should be determined, except that variants known to yield widely different ODDS on their own should not be put into the same group. Apart from this, the process is unavoidably subjective—but it is far from entirely arbitrary. In particular, it is greatly aided by strong impressions gained while perusing alphabetical listings of names from Files A and B.

3. APPLICATION: REFINEMENTS AND SHORTCUTS

Many choices have had to be made in the past between shortcuts in the way the ODDS are calculated versus corresponding refinements in which the shortcuts are not used. Such choices are inescapable, but only rarely have their effects on the calculated ODDS been quantified. Indeed, where data to support the more refined alternative were lacking, the comparison often was not possible. But now the extensive data from large files of LINKS accumulated at Statistics Canada make it attractive to assess the effects on discriminating power when people's names are compared in alternative ways.

Table 5. Comparison Outcomes for the Given Name GEORGE, With Examples of Possible Groupings

Values*		Total outcomes	ODDS
x	y		
Full Agreement			
GEORGE–GEORGE		3,130	89.7/1
Partial Agreement			
GEORGE–GEO		6	87.9/1
GEORGE–GEOR to GEORGDZ (including GEORDIE)		11	14.9/1
GEORGE–GEORGES		28	12.1/1
GEORGE–GEORGET to GEORGZ (including GEORGIO)		3	21.6/1
Other (including disagreements)			
GEORGE–G* (* = other; few synonyms)		16	1/5.6
GEORGE–non-G (full disagreements)		175	1/13.2

* Data for $A_x \cdot B_y$ and $B_x \cdot A_y$ are pooled.

We consider here six shortcuts (and their corresponding refinements):

1. Use of the simplified formula (see expression 1.5)
2. Pooling of first and second given names, to reduce the number of look-up tables of the value-specific frequencies, $N(B_y)/N(B)$, when using the simplified formula
3. Use of a wholly versus a partially global term in the numerator of the simplified formula when calculating ODDS for the various levels of outcome (i.e., both A_x and B_y being nonspecific in the LINKS, versus A_x being specified as equal, successively, to each of the 51 common names)
4. Not updating the global ODDS
5. Recognizing the specificities of just the agreement portions of names that only partially agree
6. Pooling complementary partial agreements (e.g., $A_x \cdot B_y$ = MICHAEL–MIKE, plus $A_y \cdot B_x$ = MIKE–MICHAEL).

Past and current practices with regard to these shortcuts are reviewed elsewhere (Hill 1981; Howe and Lindsay 1981; Newcombe 1988).

The importance of a given refinement as compared with its corresponding shortcut is assessed by comparing the ODDS when calculated in the two ways. The ratios of the two ODDS will be termed "error factors" or "correction factors." These factors vary for different names as represented in File A (e.g., the given name JOHN) and for different comparison outcomes (e.g., JOHN–JACK). One such type of "correction factor" is defined in the second part of expression 1.5. Its use as part of the full expression constitutes a refinement, its omission constitutes a shortcut, and its use on its own reveals the factor difference between the ODDS as obtained in the two ways.

Comparisons between different refinement/shortcut choices may be based either on the frequency distributions of the error levels, as defined earlier, or on the median and maximum error factors. Sometimes a combination of the two may be appropriate. Data from the six types of comparisons are presented in Figure 2 (parts a to f) and Table 6 (lines 1 to 6). The histograms in Figure 2 are appropriately weighted throughout; for example, in part a of Figure 2 by the frequencies of the names in File A.

The magnitudes of such error factors may vary with the particular name or linkage project; that is, forming a distribution of error factors as shown in Figure 2. The log error factor approach, with base 2, is used in this Figure. (Log error factor = 1 indicates a difference by a factor of 2, log error factor = 2 indicates a difference by a factor of 4, and so on.) Because we are dealing with a spectrum of error factors and need to divide it into discrete levels, we have recognized central values of 1, 2, 4, 8, 16, and so on (equivalent to logs to the base 2 = 0, 1, 2, 3, 4, and so on). Standard rounding of the logs is used to assign the appropriate central values.

3.1 Ranking the Choices

The effect of choosing a shortcut, or its corresponding refinement, is best seen in a listing of the associated error factors in descending order. These create in the mind a compelling picture. What they teach us is that the feedback of actual data does away with the need for guesswork. For our current purposes it is sufficient that the results of the tests be summarized (Fig. 2, Table 6) and that examples be given.

Use of the simplified formula, for example, results in error factors as high as 6.4, with 13% of the 34,737 comparisons associated with the four-fold level of error. Nine of the 51 common names and 5 of the 15 informal names are involved (i.e., DOUGLAS, ERNEST, EMILE, FRANK, HAROLD, CLARENCE, ALFRED, HERBERT, HARRY, FRED, PETE, MIKE, SAM, ALEX). Similarly modest error factors result from pooling of first plus second names, use of a wholly global numerator, and pooling complementary partial agreements. In these examples the magnitudes of the error factors vary with the values of the given names.

The effects of not updating the ODDS differ in that the error factors vary with the quality of the files used to initiate the death searches and, therefore, with the particular linkage study. Error factors are greater for the partial agreements than for the full agreements and disagreements, independent of the actual values of the names; for this reason, only the partial agreements are considered here. Again, the effects of the shortcut are modest. The largest are associated with search files (Files A) in which the quality of the identifiers differed most widely from the average; that is, were either much better

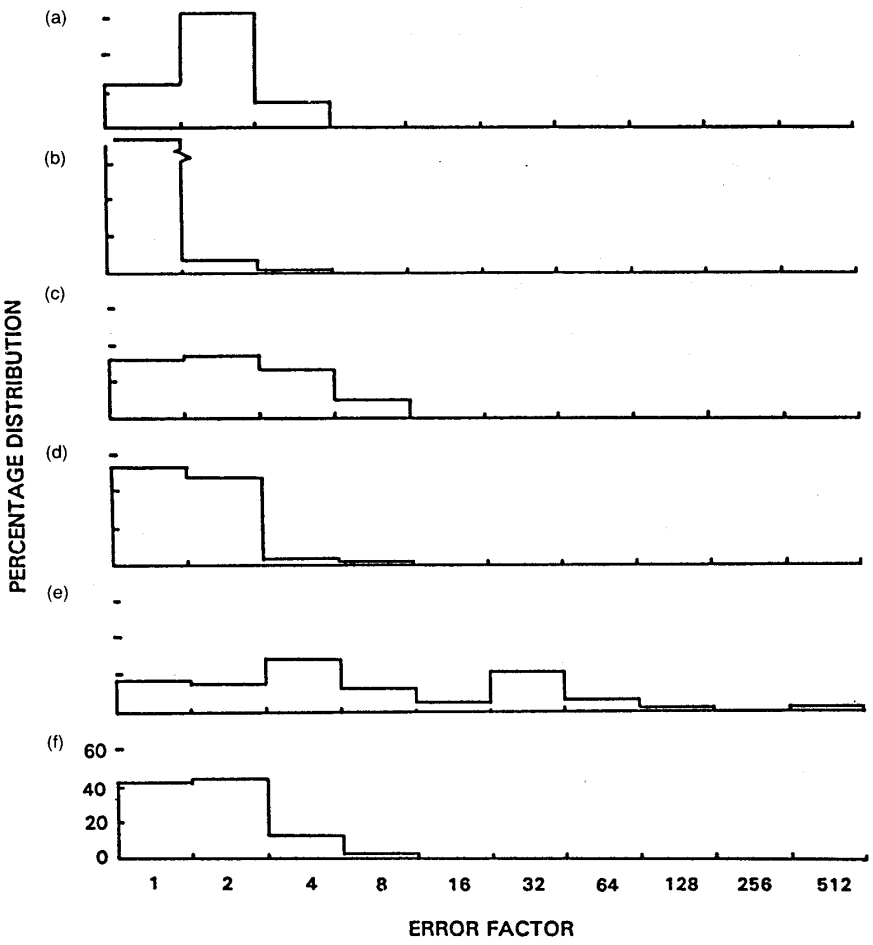


Figure 2. Frequency Distributions of Error Factors Resulting From Shortcuts in the Comparison Procedures for Male Given Names. (a) Simplified formula; (b) Pooled first plus second names; (c) Wholly global numerator; (d) Not updating the ODDS; (e) Recognizing the specificities of just the agreement portions; (f) Pooling complementary partial agreements.

(as with infant death-to-birth linkages) or much worse (as with certain employment records). This is because in such cases the composite ODDS most poorly represent the ODDS appropriate for the particular project.

Only for one kind of choice are the error factors truly large. This has to do with the practice of recognizing the specificities of only the agreement portions of names that do not fully agree (versus recognizing the full specificities of both members). The most extreme examples (with their error factors) include WALTER-WLADYSL (686.7), ERNEST-EARNEST (412.7), PETER-PIO (190.2), WILLIAM-BILLY (160.6), ROY-LEROY (155.8), JOHN-JUHO (82.6), LEONARD-LENARD (82.4), RAYMOND-RAIMOND (77.6), LOUIS-LOIS (72.7), and JOHN-JAN (57.7). Only when the full specificities are taken into account does the discriminating power get efficiently exploited.

Table 6. Ranking the Choices Between Refinements Versus Shortcuts: Partial Agreements Only

Shortcut	Median error	Maximum error	Rank*
1. Simple Formula	1.7	6.4	(3)
2. Pooling First and Second	1.1	14.2	(6)
3. Global Numerator	2.1	12.2	(2)
4. Update Omitted	1.4	6.4	(5)
5. Partial Specificity	4.9	686.7	(1)
6. Complementary Partials	1.4	11.2	(4)

* Rank based on median error factor, followed by maximum.

4. CONCLUSIONS AND RECOMMENDATIONS

Current tests assess the degree to which inherent discriminating power is exploited where names are used to bring together records of the same persons, especially when alternative forms of a name are compared. The emphasis differs from that of procedures based on degrees of phonetic similarity plus lists of exceptions, in that both values get recognized and necessary data are drawn from large accumulations of linked pairs of records.

Motivation to achieve maximum refinement in record linkage comes from the social trend towards larger and more

numerous personal data banks. Complex influences govern the trend. Records proliferate because people rely on governments and the commercial sector for increased security and benefits of many sorts, plus conveniences and luxuries where possible. The process is slowed by fears that the right to privacy might suffer, but it is also accelerated by public insistence on a right to know whether perceived threats to health and well-being are real, because the best answers often come only through increased access to personal data banks (in Canada, see Bouchard, Roy, and Casgrain 1985; Fair 1989; Jordan-Simpson, Fair, and Poliquin 1988; Leyes 1990; Medical Research Council of Canada 1968; Newcombe et al. 1983; Roos, Wajda, and Nichol 1986; Smith and Newcombe 1980, 1982; elsewhere, see Arellano, Petersen, Pettiti, and Smith 1984; Baldwin, Acheson, and Graham 1987; Copas and Hilton 1990; Jaro 1989; Kilss and Alvey 1985; Patterson 1980; Rogot, Sorlie, Johnson, Glover, and Treasure 1988; Winkler 1989a,b,c,d; also see early reviews by Acheson 1967 and Farr 1875). A logical step in this evolution is the automation of registers embracing whole populations (Dunn 1946; Leyes 1990; Marshall 1947; Redfern 1990; Scheuren 1990).

The current approach follows a general trend in statistics, which is to develop empirical reference distributions using computers, rather than to rely mainly on theoretical distributions. Here, we use large composite files of `LINKS` (Fair et al. 1990, 1991) and even larger files of random pairs to serve as `NONLINKS` (Lalonde 1989). Examples as applied to other statistical problems include uses of the "bootstrap" method (Efron and Tibshirani 1986, 1992). Moreover, those involved with linkage technology stress the need to archive empirical data from past linkage studies, and use it to compare the performances of different systems (see, for example, Howe 1986; Howe and Spasoff 1986a,b; Jabine and Scheuren 1986; Scheuren et al. 1986; Science Council of Canada 1986; Smith 1986).

In a sense, we emphasize here a role for semiautomated "learning," from past experience. Complexity need not be a serious barrier, because complex procedures, once developed, may be used repeatedly and can evolve through successive refinements.

[*Received October 1989. Revised May 1991.*]

REFERENCES

Acheson, E. D. (1967), *Medical Record Linkage,* Oxford, U.K.: Oxford University Press.

Arellano, M. G., Petersen, G. R., Petitti, D. B., and Smith, R. E. (1984), "The California Automated Mortality Linkage System," *American Journal of Public Health,* 74, 1324–1330.

Baldwin, J. A., Acheson, E. D., and Graham, W. J. (eds.) (1987), *Textbook of Medical Record Linkage,* Oxford, U.K.: Oxford University Press.

Bouchard, G., Roy, R., and Casgrain, B. (1985), *Reconstitution Automatiques des Familles, le Système SOREP* (Vols. I and II), Chicoutimi, Quebec: Centre Interuniversitaire de Reserches sur les Populations (SOREP).

Copas, J. B., and Hilton, F. J. (1990), "Record Linkage: Statistical Models for Matching Computer Records," *Journal of the Royal Statistical Society,* Ser. A, 153 (Part 3), 287–320.

Dunn, H. L. (1946), "Record Linkage," *American Journal of Public Health,* 36, 1412–1416.

Eagen, M., and Hill, T. (1987), "Record Linkage Methodology and its Application," in *Statistical Uses of Administrative Data, Proceedings of an International Symposium,* eds. J. W. Coombs and M. P. Singh, Ottawa: Statistics Canada, pp. 139–150.

Efron, B., and Tibshirani, R. (1986), "The Bootstrap Method for Assessing Statistical Accuracy" (with discussion), *Statistical Science,* 1, 54–77.

——— (1992), "Statistical Data Analysis in the Computer Age," *Science,* in press.

Fair, M. E. (1989), *Studies and References Relating to Uses of the Canadian Mortality Data Base,* Ottawa: Statistics Canada, August 1989.

Fair, M. E., Lalonde, P., and Newcombe, H. B. (1990), *Tables of ODDS For Partial Agreements of Male Given Names in Linking Records,* Report OEHRS No. 9, Ottawa: Statistics Canada.

——— (1991), "Application of Exact ODDS for Partial Agreements of Names in Record Linkage," *Computers and Biomedical Research,* 24, 58–71.

Fair, M. E., Newcombe, H. B., and Lalonde, P. (1988a), *Improved Mortality Searches for Ontario Miners Using Social Insurance Index Identifiers,* Report No. INFO-0264, Ottawa: Atomic Energy Control Board.

Fair, M. E., Newcombe, H. B., Lalonde, P., and Poliquin, C. (1988b), *"Alive" Searches as Complementing Death Searches in the Epidemiological Follow-Up of Ontario Miners,* Report No. INFO-0266, Ottawa: Atomic Energy Control Board.

Farr, W. (1875), in *Supplement to the 35th Annual Report of the Registrar General,* London: Her Majesty's Stationery Office, p. 110.

Fellegi, I. P. (1985), "Tutorial on the Fellegi-Sunter Model for Record Linkage," in *Record Linkage Techniques—1985 (Proceedings of the Workshop on Exact Matching Methodologies, Arlington, Virginia, May 9–10, 1985),* eds. B. Kilss and W. Alvey, Washington, DC: Department of the Treasury, Internal Revenue Service, pp. 127–138.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory of Record Linkage," *Journal of the American Statistical Association,* 40, 1183–1210.

Hill, T. (1981), *Generalized Iterative Record Linkage System: GIRLS,* Ottawa: Statistics Canada.

Hill, T., and Pring-Mill, F. (1985), "Generalized Iterative Record Linkage System," in *Record Linkage Techniques—1985, (Proceedings of the Workshop in Exact Matching Methodologies Arlington, Virginia, May 9–10, 1985),* eds. B. Kilss and W. Alvey, Washington, DC: Department of the Treasury, Internal Revenue Service, pp. 327–333.

Howe, G. R. (1986), "Possible Future Directions in Record Linkage," in *Proceedings of the Workshop in Computerized Record Linkage in Health Research (Ottawa, Ontario, May 21–23, 1986),* eds. G. R. Howe and R. A. Spasoff, Toronto: University of Toronto Press, pp. 231–233.

Howe, G. R., and Lindsay, J. (1981), "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-Up Studies," *Computers and Biomedical Research,* 14, 327–340.

Howe, G. R., and Spasoff, R. A. (eds.) (1986a), *Proceedings of the Workshop on Computerized Record Linkage in Health Research (Ottawa, Ontario, May 21–23, 1986),* Toronto: University of Toronto Press.

——— (1986b), "Recommendations of the Workshop on Computerized Linkage in Health Research," in *Proceedings of the Workshop on Computerized Record Linkage in Health Research (Ottawa, Ontario, May 21–23, 1986),* eds. G. R. Howe and R. A. Spasoff, Toronto: University of Toronto Press, pp. 18–23.

Jabine, T. B., and Scheuren, F. (1986), "Record Linkages for Statistical Purposes: Methodological Issues," *Journal of Official Statistics,* 2, 255–277.

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association,* 84, 414–420.

Jordan-Simpson, D., Fair, M. E., and Poliquin, C. (1988), "Canadian Farm Operator Study: Methodology," *Health Reports (Statistics Canada),* 2, 141–155.

Kilss, B., and Alvey, W. (eds.) (1985), *Record Linkage Techniques—1985, (Proceedings of the Workshop on Exact Matching Methodologies, Arlington, Virginia, May 9–10, 1985),* Washington, DC: Department of the Treasury, Internal Revenue Service.

Lalonde, P. (1989), "Deriving Accurate Weights Using Non-Links," in *Proceedings of the Record Linkage Sessions and Workshop, Canadian Epidemiology Research Conference—1989,* eds. M. Carpenter and M. E. Fair, Ottawa: Statistics Canada, pp. 149–157.

Leyes, J. (1990), "Release of a Pilot Longitudinal Administrative Database," *The Daily (Statistics Canada),* Monday, October 22, 1990, p. 6.

Marshall, J. T. (1947), "Canada's National Vital Statistics Index," *Population Studies,* 1, 204–211.

Medical Research Council of Canada (1968), *Health Research Uses of Record Linkage in Canada,* Report No. 3, Ottawa: Author.

Newcombe, H. B. (1967), "Record Linking: The Design of Efficient Systems for Linking Records into Individual and Family Histories," *American Journal of Human Genetics,* 19, 335–359.

——— (1988), *Handbook of Record Linkage: Methods for Health and Sta-*

tistical Studies, Administration and Business, Oxford, U.K.: Oxford University Press.
Newcombe, H. B., Fair, M. E., and Lalonde, P. (1987), "Concepts and Practices that Improve Probabilistic Record Linkage," in *Statistical Uses of Administrative Data, Proceedings of an International Symposium (Ottawa, Ontario, November 23–25, 1987),* eds. J. W. Coombs and M. P. Singh, Ottawa: Statistics Canada, pp. 127–138.
——— (1989), "Discriminating Powers of Partial Agreements of Names for Linking Personal Records, Part I: The Logical Basis, and Part II: The Empirical Test," *Methods of Information in Medicine,* 28, 86–91, 92–96.
Newcombe, H. B., and Kennedy, J. M. (1962), "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information," *Communications of the Association for Computing Machinery,* 5, 563–566.
Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science,* 130, 954–959.
Newcombe, H. B., Smith, M. E., Howe, G. R., Mingay, J., Strugnell, A., and Abbatt, J. D. (1983), "Reliability of Computer versus Manual Death Searches in a Study of Eldorado Uranium Workers," *Computers in Biology and Medicine,* 13, 157–169.
Patterson, J. E. (1980), "The Establishment of a National Death Index in the United States," in *Cancer Incidence in Defined Populations (Banbury Report No. 4),* eds. J. Cairns, J. L. Lyon, and M. Skolnick, Cold Spring Harbor, Long Island, New York, Cold Spring Harbor Laboratory, pp. 443–451.
Redfern, P. (1990), "Sources of Population Statistics: An International Perspective," in *Population Projections: Trends, Methods and Uses,* OPCS Occasional Paper 38, London: Office of Population Censuses and Surveys, Her Majesty's Stationery Office.
Rogot, E., Sorlie, P. D., Johnson, N. J., Glover, C. S., and Treasure, D. W. (1988), *A Mortality Study of One Million Persons: First Data Book,* NIH Publication No. 88-2896, Bethesda, MD: Public Health Service, National Institutes of Health.
Roos, L. L., Wajda, A., and Nicol, J. P. (1986), "The Art and Science of Record Linkage: Methods that Work with Few Identifiers," *Computers in Biology and Medicine,* 16, 45–57.
Scheuren, F. (1985), "Methodological Issues in Linkage of Multiple Data Bases," *Record Linkage Techniques—1985,* Washington, DC: Department of the Treasury, Internal Revenue Service, pp. 155–178.
——— (1990), Discussion of "Rolling Samples and Censuses," by L. Kish, *Survey Methodology,* 16, 72–79.
Scheuren, F., Alvey, W., and Kilss, B. (1986), "Record Linkage for Statistical Purposes in the United States," in *Proceedings of the Workshop in Computerized Record Linkage in Health Research (Ottawa, Ontario, May 21–23, 1986),* eds. G. R. Howe and R. A. Spasoff, Toronto: University of Toronto Press, pp. 198–210.
Science Council of Canada (1986), *Proceedings: A National Workshop on the Role of Epidemiology in the Risk Assessment Process in Canada,* Catalogue No. SS24-23/1985, Ottawa: Author.
Smith, M. E. (1986), "Future Needs and Directions for Computerized Record Linkage in Health Research in Canada: Future Study Plans," in *Proceedings of the Workshop in Computerized Record Linkage in Health Research (Ottawa, Ontario, May 21–23, 1986),* eds. G. R. Howe and R. A. Spasoff, Toronto: University of Toronto Press, pp. 211–230.
Smith, M. E., and Newcombe, H. B. (1975), "Methods for Computer Linkage of Hospital Admission-Separation Records into Cumulative Health Histories," *Methods of Information in Medicine,* 14, 118–125.
——— (1979), "Accuracies of Computer Versus Manual Linkages of Routine Health Records," *Methods of Information in Medicine,* 18, 89–97.
——— (1980), "Automated Follow-up Facilities in Canada for Monitoring Delayed Health Effects," *American Journal of Public Health,* 73, 39–46.
——— (1982), "Use of the Canadian Mortality Data Base for Epidemiological Follow-up," *Canadian Journal of Public Health,* 73, 39–46.
Sunter, A. B. (1968), "A Statistical Approach to Record Linkage," in *Record Linkage in Medicine (Proceedings of the International Symposium, Oxford, July 1967),* ed. E. D. Acheson, London: E & S Livingstone, pp. 89–109.
Winkler, W. E. (1985), "Preprocessing of Lists and String Comparison," in *Record Linkage Techniques—1985,* eds. W. Alvey and B. Kilss, Washington, DC: Department of the Treasury, U.S. Internal Revenue Service, pp. 181–187.
——— (1989a), *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage (Technical Report),* (paper presented at the Annual ASA Meeting in Anaheim, CA) Washington, DC: Statistical Research Division, U.S. Bureau of the Census.
——— (1989b), "The Interaction of Record Linkage Practice and Theory," in *Proceedings of the Record Linkage Sessions and Workshop, Canadian Epidemiology Research Conference—1989,* eds. M. Carpenter and M. E. Fair, Ottawa: Statistics Canada, pp. 139–148.
——— (1989c), "Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage," in *Proceedings of the Fifth Census Bureau Annual Research Conference,* Washington, DC: U.S. Bureau of the Census, pp. 145–155.
——— (1989d), "Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage," *Survey Methodology,* 15, 101–117.

Comment

MAX G. ARELLANO*

Because the discussion is focused primarily on first name variants, the title perhaps should more appropriately be "The Use of Given Names for Linking Personal Records." While not stated, the implication is that the "special skill" developed by humans is of considerable value in the decision making process. The fact is that the "special skills" vary considerably from person to person and that the biases that they bring to the evaluation may hinder rather than assist in the record linkage process.

The "past experience" argument is spurious. There is no reason to believe that the lessons learned from a Canadian mortality study will be of any benefit to an evaluation of a Cuban expatriate population or that experience gained in a study of mortality among Chicago nurses will be of any benefit to a study of child abuse in Seattle.

It does not follow at all that "if a machine is to acquire similar ability, it too must rely on past experience." For instance, an analysis of the decisions made by the operators may well reveal that their judgments are based primarily on their perceptions of probability of occurrence and the reliability of the data. These factors are quantifiable and not dependent on past experience.

* Max G. Arellano is Chief Scientist, Advanced Linkage Technologies of America, Inc., Berkeley, CA 94707.

Journal of the American Statistical Association
December 1992, Vol. 87, No. 420, Review Paper

1. COMPUTER LINKAGE

The presentation is much too informal. It is difficult enough to figure out how the authors derive their likelihood ratio without trying to deduce how they arrived at the "two additional factors." What are the consequences of failure to recognize these two factors? The authors' arguments would be much easier to follow if they were presented in mathematical terms.

I fail to see the relevance of "describing the insights on which their success depends" in "plain language . . . such as they would understand" to the clerical searchers?

Probabilistic linkage procedures must be based on a probability model with definable probability distribution or density functions. The Fellegi-Sunter model is a probability model; I see no evidence of a probability model in this article. This is not to say that there is no merit in the approach presented in this article; however, it should properly be presented as a subjective probability or expert system.

The concept of falsely matched random pairs is a fascinating topic. But if, as the authors state in the first paragraph of page 7, "it was not difficult to determine the corresponding likelihoods for a control group of falsely matched random pairs (NONLINKS)," then why didn't they present the procedures that they used to obtain this value? I believe that this would have contributed immensely to their presentation.

I understand the context within which the historical development is being presented, and I am in complete sympathy with the authors' objectives. However, the point must be made that the validity of the linkage rationale that they describe is a function of the correctness of the linkage decisions that were made. A statement is badly needed regarding whether it was possible to confirm their decisions or how they were able to establish a level of confidence in them. After all, this is the central issue in record linkage.

The authors seem to feel that refinements in linkage decision criteria only proceeded "independent of the formal theory" (p. 1194). There is no reason, however, to believe that these or similar developments could not have or did not proceed within the context of the formal theory, perhaps without the knowledge of the authors.

Routine cross-comparisons can also be extremely wasteful of available resources if they are not called for by the nature of the data. In most linkage evaluations, 85–90% of the correct linkages can readily be detected as exact matches on name and birthdate.

The authors state that "frequent close scrutiny of difficult matches provided insights that would have been missed had refinement been sought through theory alone" (pp. 1194–1195). In view of the fact that, as the authors readily admit, their procedures are not based on the formal theory, the validity of this statement is doubtful. How can they be sure of the correct direction of these "difficult matches" without reference to the subjects whose records are being linked?

In the development of their decision criteria, Fellegi and Sunter stated very clearly that the effect of their weight computation is "to array the record pairs, relative to one another, in descending order of the likelihood of a correct match."

If the authors had observed the strict requirements of the Fellegi-Sunter model, they would have realized that the restriction of the comparison-space to linkages with identical surname phonetic codes requires an adjustment to the computation of the surname weights. This adjustment would have compensated for the distortion that they observed in the "crossover" point.

The discussion of prior likelihood is unnecessarily vague. What are prior likelihoods? How are they estimated? It is not sufficient to simply show these as $P(\text{LINK})/P(\text{NONLINK})$.

The authors would do better to present their derivation in terms of the Fellegi-Sunter model. Within the context of the Fellegi-Sunter model, there is no need for concern about "fortuitous LINKS in the random pairs."

"ODDS" should be expanded on. ODDS of what?

One cannot have conditional probabilities without either a probability distribution or density function. I don't see any evidence of either.

The derivation leads to the conclusion that we need to know the number of links with value A_x and B_y (p. 1196). But this is exactly what we are trying to accomplish with the linkage; that is, this information is not known. The authors gloss over this point without explaining how they intend to fill in the blanks.

The "tradeoff" argument (p. 1196) is completely spurious. The categories are determined by the characteristics of the data. It is not reasonable to assume that the operators of linkage software can be expected to ensure that every outcome group is broad enough so that "$N(A_x \cdot B_y | \text{LINK})$ is represented by at least one comparison pair. Otherwise no ODDS can be calculated" (p. 1196). This sounds as though the procedure is controlling the application. Linkage software can readily be designed so that empty categories are either assigned zero values or some predetermined default value.

Partial given name agreements can be easily handled by phonetically encoding the name and constructing an exception list. This procedure has been in use by most organizations with which I am familiar for at least the past 16 years.

The authors state that "confusion still remains concerning the implications, and is not explicitly addressed by existing formal theory" (p. 1197). The authors are obviously privy to some controversy to which I am not.

We keep coming back to the fact that $N(A|\text{LINK})$ is unknown. The authors should have expanded on how they obtain this value.

2. EMPIRICAL DISTRIBUTIONS OF LINKS AND NONLINKS

The authors apparently believe that the results of particular linkage evaluations can be extrapolated to other linkage evaluations. Although this may be true in general, it cannot be relied on as a matter of policy. For instance, the reporting of demographic information by psychiatric patients may be much less reliable than information gathered for epidemiologic research purposes, the point being that "memories of past encounters with similar problems" may well lead you astray.

Although the authors criticize the practice of simply multiplying the ODDS for individual identifiers to combine them

for a whole set "which would only be proper if they were independent of each other," (p. 1198) this appears to be exactly what they do—or do they believe that the $P(\texttt{LINK})$ term corrects for the dependence among the identifiers?

The authors appear obsessed by the presence of false-positive links and false-negative links. The purpose of a record linkage, however, is not to eliminate these links, but rather to minimize them. There is a point beyond which the cost of refining the rules outweighs the advantages of applying them, particularly if the refinement requires an extensive amount of manual review.

The authors state that the number of possible pairings is the product of the two file sizes. This is true, however, only if all possible pairwise comparisons are actually formed between the two files—a practice that would be prohibitively expensive. The actual number of pairings is a function of the blocking strategy that was used. The difference is not at all trivial.

The problem to which the authors allude beginning on page 1199, of establishing upper and lower threshold values is not related to the independence problem. It is a function of the far greater size of NONLINKS relative to the LINKS—a fact, by the way, that is well known to persons involved in probability linkage, despite the concerns expressed by the authors. The threshold problem would exist even if a correction for the dependence of the identifiers could be incorporated into the computation of the total odds.

The "strong impressions gained while perusing alphabetic listings of names from Files A and B" (p. 1200) are of value only if their validity can be established by reference to the truely valid linkages. Under any circumstances, however, unless these "impressions" can be translated into formal rules, these procedures are obviously not suitable for mass production purposes.

3. CONCLUSIONS AND RECOMMENDATIONS

Rarely, if ever, does an experienced human clerk obtain feedback regarding the validity of a difficult linkage decision. Without this information, the clerk cannot possibly know whether his intuition was correct or not. If the clerk is not routinely receiving this feedback, the rules he has been developing may well lead to the systematic introduction of error into the decision criteria he is applying to the linkages.

The authors contend that the thought patterns (of the "experienced human clerk") clearly differ from those of a skilled mathematician. However, the consensus among most persons involved in probability linkage with whom I am familiar is that subjective judgment is based on perceptions of probabilities of occurrence, a feel for the reliability of the data, and a familiarity with the various ways in which the same item of information can be recorded. There is no mystery; all of these factors are readily quantifiable.

Before one can "learn" from past experiences (p. 1203), two elements are necessary:

1. One must rigorously define how to measure a "success." The authors have failed to do so.
2. One must demonstrate that the lessons learned from a particular linkage evaluation have relevance to the new linkage evaluations that are under active consideration. Personally, I would hesitate to apply the lessons which the authors have learned from their Canadian experience to our ongoing linkage evaluations in California.

4. REVIEWER'S SUMMARY

The authors' bias toward an informal approach to the development of linkage decision criteria is obvious, as is their sentiment that no real value can come from pursuing formal probability linkage models such as the Fellegi-Sunter model. One must ask, however, if the authors are aware of any objective basis for their assertion that an informal approach is superior to an approach based on a formal mathematical model.

Organizations with which I have been affiliated have used various versions of the Fellegi-Sunter probability linkage model for the past 17 years, with a great deal of success. Our linkage evaluations have included files with over one million records. Although manual review of the borderline linkages is an essential element of our linkage processing, because of the very large number of linkages identified it would be impractical for us to become overly involved in resolution of the difficult matches. Although we routinely observe the instances in which there is a substantial amount of conflict among the identifiers, I would question the wisdom of applying the lessons learned from the outcome of one difficult match to another difficult match.

Newcombe would do well to explore the operation of systems that use a formal probability linkage model; perhaps he would then gain a greater appreciation of them. We welcome his call for greater mutual cooperation. If there is sufficient interest, we would be glad to participate in a comparative linkage methodology evaluation study.

REFERENCE

Fellegi, I., and Sunter, A. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association,* 64, 1183–1210.

Rejoinder

HOWARD B. NEWCOMBE, MARTHA E. FAIR, and PIERRE LALONDE

Arellano has provided a detailed critique of our article, much of which does not actually contradict what we have said or conflict with our own understanding, even though the language may differ. Any rejoinder, therefore, should confine itself to major points of difference on matters of emphasis or fact.

We do *not,* for example, believe that added refinement is always cost-effective in all situations. But by exploring the ways in which the comparison space may be more finely partitioned, we hope to expand both the present and the future *potential* for improved linkage performance at acceptable cost. It was the crudity of the popular agreement-disagreement distinction that provided the initial major motivating force. What impressed us as a source of innovation was the wealth of alternative comparison procedures and of multiple alternative outcome definitions, that got applied freely by a person's mind. Many of these proved highly effective in the case of difficult links, once the true status of the record pairs was confirmed later by independent means.

The emphasis we have placed on multiple partitioning of the comparison space has applications that are not confined to any particular identifier field. For example, colleagues at one time were concerned that our recognition of multiple outcomes from comparisons of place of work with place of death, when doing death searches, might be contrary to linkage theory. The observation was that workers at an Ontario uranium refinery who migrated before dying tended to die more often, either in the home province or in western Canada, but less often in eastern Canada and only rarely in the most easterly province, as compared with the random expectation. A somewhat different pattern (i.e., empirical distribution) was observed for workers in the uranium mines of Saskatchewan and the Northwest Territories; so there was no question of extrapolating from one subset of the cohort to the other. Here, final verification of the linkage status of the record pairs was not in doubt. Even before that verification, however, approximate likelihood ratios contributed to the linkage process and to the updating and iterative refinement, both of the linked files and of the likelihood ratios together. To establish useful comparison rules, we first needed to "learn" what only the linked files could "teach" concerning the empirical distributions and the outcome definitions most likely to exploit their discriminating power to good advantage. Earlier objections to the approach were later withdrawn. But if this broad emphasis on added partitioning of the comparison space to reveal a greater diversity of usable differences in observed versus random distributions is indeed fundamentally flawed, as Arellano seems to believe, we would welcome from him a concrete example to that effect.

We also appreciate Arellano's stated interest in "comparative linkage methodology evaluation studies," especially if this interest encompasses the current focus on given names. Thus he could readily compare his own practice of recognizing phonetic similarity plus an exception list with our wholly value-specific approach, using Canadian data that have been published in great detail for just such a purpose (Fair, Lalonde, and Newcombe 1990). Moreover, Figure 2 of our article indicates a convenient way to display the results. Indeed, the two approaches need not be mutually exclusive, since ours provides what might be viewed as just a very long "exception list" based on the most appropriate data for searches of the particular File B.

We are aware that in principle any use of data from old linkages when starting a new linkage operation must involve some degree of extrapolation, at least initially. But this is not necessarily so for the later stages, after there has been opportunity for iterative adjustments based on the new links.

Arellano has alluded to a number of exceedingly simple concepts which appear to him to give rise to logical difficulties. For example:

- "We keep coming back to the fact that $N(A|\texttt{LINK})$ is unknown."
- "The concept of falsely matched random pairs is a fascinating topic. But, . . . why didn't they present the procedures that they used to obtain this value?"
- "The problem . . . of establishing an upper and lower threshold value is not related to the independence problem."

At the risk of repeating what is in the article, we will consider these together here:

- $N(A|\texttt{LINK})$: The simple answer is that one may do a small preliminary linkage, perhaps manually, to arrive at the approximate proportion of records in File A that will find a correct match in File B. There is no serious obstacle to this because, as Arellano points out, often 85 to 90% of the linkages are easy anyway. What is curious about the question itself is that this first step is the same as is routinely employed to obtain preliminary estimates of the likelihood ratios. The process thereafter, of iteratively refining early crude estimates, has been repeatedly emphasized in the literature (e.g., see Howe and Lindsay 1981).
- Random Pairs: Again, only modest ingenuity is needed to solve the problem. Where the outcomes of interest are defined in complicated ways, there is no need to resort to theory to determine their frequencies of occurrence in random pairs. Instead, one uses the computer to put together large numbers of random pairs, among which the proportions of the outcomes of special interest may be determined by tabulation (Lalonde

Journal of the American Statistical Association
December 1992, Vol. 87, No. 420, Review Paper

1989). Alternatively, for simple value specific outcomes such as ROBERT compared with BOB, the random expectation is just the product of the proportions of these two values in Files A and B (or Files B and A) prior to linking.

- Thresholds and Independence: The statement that lack of independence has no effect on the placing of the upper and lower thresholds is too sweeping to be correct. Where *correlated disagreements* (e.g., due to multiple wrong guesses on the part of an informant) have spuriously moved true links downward below the lower threshold, or where *correlated agreements* of rare specific values (e.g., of ethnic surnames and forenames, plus places of birth) have spuriously moved false matches upward above the upper threshold, preset thresholds will no longer accurately perform their intended function. Such effects are often too large to be ignored when setting the thresholds.

Initially it had *not* been our intention to raise in this article the contentious matter of the "prior likelihood" of a correct match on a single random pairing. Indeed, we did not invent the concept—but we did devise the procedure for estimating the magnitude. For all practical purposes, prior likelihoods are essentially similar to the "prior odds" that appear explicitly in the weight formula of Howe and Lindsay. The idea is also implicit in the Fellegi-Sunter theory, where two conditional probabilities (i.e., of a link and of a nonlink) are described (Fellegi and Sunter 1969, exps. 6 and 7, pp. 1185–1186). Each contains a term for a prior probability (of a match and of a non-match, respectively) before the comparison of any identifiers. These terms are $P[(a, b) \mid M]$ and $P[(a, b) \mid U]$, and their ratio represents the prior odds contained in the Howe-Lindsay weight formula. In an early version of our article, Figure 1 drew criticism from reviewers as being unsupported and incorrect. This is why details of our use and derivation of an estimated "prior likelihood" are included here together with the related idea that *blocking* be treated as *not* altering, either the total number of *possible* record pairings (actual plus potential), or the use of likelihood ratios derived from the blocking identifiers. Indeed, unless valid links are known to be lost due to blocking and their numbers can be estimated, there is no special reason why blocking need make any difference at all to the calculation of total weights or absolute odds in favor of a correct match.

Alternatively, of course, one may legitimately view each block as containing its own Files A and B; then, likelihood ratios for blocking identifiers are ignored, but a separate prior likelihood is required for every block, which may be cumbersome. Falling in between these two legitimate alternatives is a common practice that recognizes blocks and ignores likelihood ratios based on blocking identifiers, but *omits* the prior likelihood. Test results from this might seem satisfactory where the blocks happen to be small and most search records find a correct match, but it is hardly justified on logical grounds. As well, for searches of an accumulated national death file, with large blocks based only on a single surname code and with most cohort members still alive, the scale of odds that this incomplete treatment yields does not even remotely approximate the absolute scale needed for predefined error bounds.

Finally, although we are mindful of major differences of emphasis in various workers, we are unaware of any fundamental conflict between our approach and existing theory. If Arellano believes that there is such a conflict, we hope that its nature will get spelled out clearly in the future. Because much of record linkage development and application is of necessity in the hands of people trained in disciplines other than mathematics, any such clarifications ought to be in a form understandable by all who are engaged in implementing the linkage rationale.

Chapter 11

Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida

Matthew A. Jaro, System Automation Corporation

A test census of Tampa, Florida and an independent postenumeration survey (PES) were conducted by the U.S. Census Bureau in 1985. The PES was a stratified block sample with heavy emphasis placed on hard-to-count population groups. Matching the individuals in the census to the individuals in the PES is an important aspect of census coverage evaluation and consequently a very important process for any census adjustment operations that might be planned. For such an adjustment to be feasible, record-linkage software had to be developed that could perform matches with a high degree of accuracy and that was based on an underlying mathematical theory. A principal purpose of the PES was to provide an opportunity to evaluate the newly implemented record-linkage system and associated methodology. This article discusses the theoretical and practical issues encountered in conducting the matching operation and presents the results of that operation. A review of the theoretical background of the record-linkage problem provides a framework for discussions of the decision procedure, file blocking, and the independence assumption. The estimation of the parameters required by the decision procedure is an important aspect of the methodology, and the techniques presented provide a practical system that is easily implemented. The matching algorithm (discussed in detail) uses the linear sum assignment model to "pair" the records. The Tampa, Florida, matching methodology is described in the final sections of the article. Included in the discussion are the results of the matching itself, an independent clerical review of the matches and nonmatches, conclusions, problem areas, and future work required.

KEY WORDS: Census adjustment; Census coverage evaluation; EM algorithm; Postenumeration survey.

1. INTRODUCTION

Record-linkage methodology and software were developed at the U.S. Bureau of the Census during the past several years primarily to support census coverage evaluation efforts. By matching individuals counted in a census to those counted in an independent postenumeration (or pre-enumeration) survey, estimates of the quality of the enumeration can be produced. An important use of matching is to support an adjustment operation if it is decided to adjust the 1990 decennial census.

Clerical procedures typically used for such evaluations are too costly, unreproducible, error-prone, and time-consuming to be a viable alternative for such an adjustment (especially in view of the fact that state-level tabulations are due to the U.S. president by December 31, 1990). Therefore, the technical success of any adjustment procedure rests primarily on the ability to match a large number of records quickly, economically, and accurately. Even a few matching errors may be of critical importance, since population adjustments can be less than 1% in some instances. A complete discussion of the adjustment and census methodology issues can be found in Citro and Cohen (1985), Ericksen and Kadane (1985), and Wolter (1986).

Record linkage has numerous applications in both the private and public sectors. Examples include purging a list of duplicates, determining multiple-frame survey overlap, and geographic coding.

The Record Linkage Staff of the Statistical Research Division was established to implement a statistically justifiable, economical, and accurate record-linkage system to replace previous ad hoc systems and to reduce the number of cases that must be manually matched (see Jaro 1985).

Generalized computer programs have been written to implement the methodology discussed in this article. The first test of this software was the 1985 census of Tampa, Florida. The actual matching was conducted using a personal computer, although a mainframe version of the software also exists. Generalization is achieved through a program that automatically "writes" a customized program that will perform the matching for a particular application. The user specifies the fields to be matched, the record formats, parameters, blocking variables, etcetera, and the generation program creates a matcher that can be run with the desired files. This software generation technique results in a program that executes efficiently—a requirement for matching very large files.

This article presents the theoretical background necessary to understand the statistical basis of record linkage in general, the methodology developed for the estimation of parameters required by any record-linkage activity, the basic algorithmic approach used by the matcher, the specific methodology used for matching the 1985 census of Tampa to the postenumeration survey (PES), and the results of this process.

2. THEORETICAL CONCEPTS

2.1 Background

Consider two computer files, A and B, consisting of records taken from a population. Each file consists of a

* Matthew A. Jaro is Director of Research and Development, System Automation Corporation, Silver Spring, MD 20910. This work was accomplished while he was a Principal Researcher, Statistical Research Division, U.S. Bureau of the Census. The author acknowledges the contributions of R. P. Kelley on the parameter estimation methodology; Danny R. Childers, who designed and tabulated the PES; and Sue Finnegan, who directed the manual matching activities.

Journal of the American Statistical Association
June 1989, Vol. 84, No. 406, Applications and Case Studies

number of fields, or "components," and a number of records, or "observations." Typically, each observation corresponds to a member of the population and the fields are attributes identifying the individual observation, such as name, address, age, and sex. The objective of the record linkage or matching process is to identify and link the observations on each file that correspond to the same individual. The records are taken to contain no unique identifiers that would make the matching operation trivial. That is, the individual fields are all subject to error.

We can define two disjoint sets **M** and **U** formed from the cross-product of **A** with **B**, the set **A** × **B**. A record pair is a member of set **M**, if that pair represents a true match. Otherwise, it is a member of **U**. The record-linkage process attempts to classify each record pair as belonging to either **M** or **U**.

2.2 Weights

The components (fields) in common between the two files are useful for matching. Not all components, however, contain an equal amount of information, and error rates vary. For example, a field such as sex only has two value states and consequently could not impart enough information to identify a match uniquely. Conversely, a field such as surname imparts much more information, but it may frequently be reported or transcribed (keyed) incorrectly.

Weights are used to measure the contribution of each field to the probability of making an accurate classification. Newcombe and Kennedy (1962) discussed the concept of weights based on probabilities of chance agreement of component value states. Fellegi and Sunter (1969) extended these concepts into a more rigorous mathematical treatment of the record-linkage process. Their definition of weights takes into account the error probabilities for each field by using a log-likelihood ratio. Let m_i = Pr{component i agrees $\mid r \in \mathbf{M}$} and u_i = Pr{component i agrees $\mid r \in \mathbf{U}$} for all record pairs r. If, for a given record pair, component i agrees (matches), then the weight for component i, $w_i = \log_2(m_i/u_i)$. If component i disagrees, then the weight $w_i = \log_2((1 - m_i)/(1 - u_i))$.

2.3 Decision Procedure

For any record pair, a composite weight can be computed by summing the individual component weights. Since $m_i > u_i$ in most cases, fields that agree make a positive contribution to this sum, whereas fields that disagree make a negative contribution. A most significant concept advanced by Fellegi and Sunter (1969) is an optimal decision procedure for record linkage. For this procedure, three states are defined. A record pair is classified as a match if the composite weight is above a threshold value, a nonmatch if the composite weight is below another threshold value, and an undecided situation if the composite weight is between these two thresholds.

The threshold values can be calculated (see Sec. 3.4) given the acceptable probability of false matches (the probability that a record pair is classified as a match when the records do not represent the same individual) and the probability of false nonmatches.

2.4 Estimation of the u_i

Values for the m_i and the u_i probabilities must be estimated for each pair of files to be matched. Estimating the u_i (the probability that a component agrees given **U**) is simplified by the fact that the cardinality of the set **U** (denoted by $|\mathbf{U}|$) is generally much greater than that of **M**. For two files, both of equal size, F, $|\mathbf{M}| = pF$, where p is the proportion of matched pairs, and $|\mathbf{U}| = F^2 - pF$. Consequently, estimates for the u probabilities can be obtained by ignoring the contribution from **M** and considering only the probability of chance agreement of the component i. Usually this can be estimated from a sample of pairs rather than from all pairs.

Estimating the m_i probabilities (the probability that a component agrees given **M**) is more difficult. Conditioning on **M** presupposes an a priori knowledge of F correctly matched pairs. This could be obtained by a prelinked sample of the population. If such a sample were obtained clerically, much expense would be involved and the error rates for the clerical operation might be too high to permit accurate parameter estimation. One solution is blocking and using a latent trait model.

2.5 Blocking

For files of average size $|\mathbf{A} \times \mathbf{B}|$ is too great to consider all possible record pairs. Since there are many more record pairs in **U** than in **M** and 2^n possible comparison configurations involving n fields, drawing record pairs at random would require a sample size approaching all record pairs (for typical applications) to obtain sufficient information about the relatively rare **M** cases.

The two files can be partitioned into mutually exclusive and exhaustive blocks designed to increase the proportion of matched pairs observed while decreasing the number of record pairs to compare. Comparisons are restricted to record pairs within each block. Consequently, blocking is important for the actual matching and for parameter estimation activities. Blocking is generally implemented by means of sorting the two files on one or more variables. For example, if both files were sorted by zip code, the pairs to be compared would only be drawn from those records where zip codes agree. Record pairs disagreeing on zip code would not be considered and hence would be automatically classified as nonmatches (elements of **U**).

To be effective at enriching the **M** cases, such blocking variables must contain a large number of value states that are fairly uniformly distributed and such variables must have a low probability of reporting error (i.e., a high weight). Blocking is a trade-off between computation cost (examining too many record pairs) and false nonmatch rates (classifying record pairs as nonmatches because the records are not members of the same block). Multiple-pass matching techniques using independent blocking variables for each run can minimize the effect of errors in a set of blocking variables. R. P. Kelley has developed an algo-

rithm that may assist in choosing the best blocking scheme in light of these trade-offs (see Kelley 1984).

3. PARAMETER ESTIMATION METHODOLOGY

3.1 Comparison Configuration Frequencies

This section discusses the methodology used to estimate the m_i probabilities. Information about the comparison configurations observed is provided by the matching software itself, which tabulates frequencies for all 2^n possible patterns of agreement and disagreement on n fields. To increase the proportion of matched pairs examined, these tabulations are performed using just those record pairs in which both observations come from the same block. Fortunately, the m_i probabilities may reasonably be expected to be independent of the blocking schemes chosen as long as errors in the blocking variables do not exclude an excessive number of matched records from the tabulations. This is consistent, however, with the goal of choosing blocking variables with low reporting error rates. The independence of the m_i probabilities to choices in blocking and the effect of errors in the blocking variables to the final estimates are yet to be determined.

Given frequencies for all possible agreements and disagreements, the m_i probabilities can be estimated using any of several procedures. The EM algorithm described here is the most effective of those developed and tested.

3.2 The EM Algorithm

Given n fields and a sample of N record pairs drawn from $A \times B$, let $\gamma_i^j = 1$ if field i agrees for record pair j, let $\gamma_i^j = 0$ if field i disagrees for record pair j, for $i = 1, \ldots, n$ and $j = 1, \ldots, N$. Further, let γ^j be the vector of ones and zeros showing field agreements and disagreements for the jth pair in the sample, and let γ be the vector containing all of the γ^j.

The m_i and u_i probabilities can be defined as $m_i = \Pr\{\gamma_i^j = 1 \mid r_j \in \mathbf{M}\}$ and $u_i = \Pr\{\gamma_i^j = 1 \mid r_j \in \mathbf{U}\}$ for a randomly selected record pair r_j and $i = 1, 2, \ldots, n$. Define p as the proportion of matched pairs equal to $|\mathbf{M}|/|\mathbf{M} \cup \mathbf{U}|$. The elements of $\mathbf{M} \cup \mathbf{U}$ (i.e., all record pairs r_j) are distributed according to a finite mixture with the unknown parameters $\Phi = (\mathbf{m}, \mathbf{u}, p)$. We will use an EM algorithm to estimate these parameters; in particular, the $\mathbf{m}$ vector is of the greatest interest. Notation is consistent with that used in Dempster, Laird, and Rubin (1977).

Let $\mathbf{x}$ be the complete data vector equal to $\langle \gamma, g \rangle$, where $g_j = (1, 0)$ iff $r_j \in \mathbf{M}$ and $g_j = (0, 1)$ iff $r_j \in \mathbf{U}$. Then, the complete data log-likelihood is

$$\ln f(\mathbf{x} \mid \Phi) = \sum_{j=1}^{N} g_j \cdot (\ln \Pr\{\gamma^j \mid \mathbf{M}\}, \ln \Pr\{\gamma^j \mid \mathbf{U}\})^T + \sum_{j=1}^{N} g_j \cdot (\ln p, \ln(1 - p))^T.$$

Although it is quite reasonable to expect the empirical frequencies to belie independence, since errors in one component will often induce errors in another, such departures will not likely disturb the ordering by composite weights of the 2^n configurations. Consequently, we will assume an independence model:

$$\Pr(\gamma^j \mid \mathbf{M}) = \prod_{i=1}^{n} m_i^{\gamma_i^j}(1 - m_i)^{1-\gamma_i^j} \quad (1)$$

and

$$\Pr(\gamma^j \mid \mathbf{U}) = \prod_{i=1}^{n} u_i^{\gamma_i^j}(1 - u_i)^{1-\gamma_i^j}. \quad (2)$$

Our application of the EM algorithm begins with estimates of the unknown parameters ($\hat{\mathbf{m}}$, $\hat{\mathbf{u}}$, $\hat{p}$) and consists of iterative applications of the expectation (E) and maximization (M) steps until the desired precision is obtained. The algorithm is not particularly sensitive to starting values, and the initial $\hat{m}$ values can be guessed. It is important, however, that the $\hat{m}$ values be greater than their corresponding $\hat{u}$ values. For Tampa, .9 was used for all of the initial $\hat{m}$. Estimation of $\hat{u}$ is discussed in Section 2.4.

For the E step, replace g_j with $(\hat{g}_m(\gamma^j), \hat{g}_u(\gamma^j))$, where

$$\hat{g}_m(\gamma^j) = \frac{\hat{p} \prod_{i=1}^{n} \hat{m}_i^{\gamma_i^j}(1 - \hat{m}_i)^{1-\gamma_i^j}}{\hat{p} \prod_{i=1}^{n} \hat{m}_i^{\gamma_i^j}(1 - \hat{m}_i)^{1-\gamma_i^j} + (1 - \hat{p}) \prod_{i=1}^{n} \hat{u}_i^{\gamma_i^j}(1 - \hat{u}_i)^{1-\gamma_i^j}}.$$

$\hat{g}_u(\gamma^j)$ can be derived similarly.

For the M step, the complete data log-likelihood can be separated into three maximization problems. Setting the partial derivatives equal to 0 and solving for $\hat{m}_i$ yields

$$\hat{m}_i = \sum_{j=1}^{N} [\hat{g}_m(\gamma^j)\gamma_i^j] \Big/ \sum_{j=1}^{N} [\hat{g}_m(\gamma^j)].$$

Further, the matrix of second partial derivatives can be shown to be negative-definite.

In practice, we store frequency counts, $f(\gamma^j)$ for each of the possible 2^n patterns of γ^j. These counts are obtained as follows: each file is partitioned into blocks by means of the blocking variables. For each block, all record pairs in the block are examined. For each record pair, the comparison vector γ^j is computed and 1 is added to the frequency count for that particular configuration. There are 2^n such counters. The counters are not reset after a block is processed, but represent the number of observations of each configuration over all blocks. Both estimation and actual matching are accomplished using the same blocks. The EM algorithm is run once using these frequencies.

The E step computes $\hat{g}_m(\gamma^j)$ for each of the 2^n patterns. This can be done without examining the individual observations, since the frequencies are a sufficient statistic for the M step. By replacing the individual observations with the frequencies, we obtain

$$\hat{m}_i = \sum_{j=1}^{2^n} [\hat{g}_m(\gamma^j)\gamma_i^j f(\gamma^j)] \Big/ \sum_{j=1}^{2^n} [\hat{g}_m(\gamma^j) f(\gamma^j)].$$

The arguments for the $\mathbf{u}$ probabilities are similar.

Finally, the proportion of matched pairs p can be esti-

mated by

$$\hat{p} = \sum_{j=1}^{N} [\hat{g}_m(\gamma^j)]/N = \sum_{j=1}^{2^n} [\hat{g}_m(\gamma^j) f(\gamma^j)] \Big/ \sum_{i=1}^{2^n} f(\gamma^j).$$

It must be remembered that the frequencies used for the p, **m**, and **u** estimates were obtained from record pairs within blocks and represent an accumulation over all blocks. Since blocking greatly reduces the number of nonmatched pairs observed and because blocking selects record pairs that are likely to match, the **u** probability estimates obtained using blocked data will be biased. Consequently, the **u** probabilities must be computed directly on unblocked data, as explained in Section 2.4, and the EM algorithm is only used to compute the **m** probabilities, where blocking enriches the number of matched pairs observed while avoiding comparisons on relatively large numbers of unmatched pairs.

The EM algorithm is highly stable and the least sensitive to the starting values of any of the methods studied. The algorithm is very simple to implement, and the probabilities will always be within bounds. The other methods are based on numerical analysis techniques, and it is possible for probabilities to exceed 1. The greater stability of the EM algorithm comes from the fact that logarithms lower the degree of the equations, whereas the method-of-moments techniques described subsequently use squared products of probabilities that are close to 1 and 0.

Comparison of the convergence criteria, rapidity of convergence, and sensitivity to independence for these methods are currently being studied.

3.3 Other Estimation Methods

The second estimation technique studied involves minimizing a system of 2^n equations (one for each comparison vector configuration) using the IMSL routine ZXSSQ [minimum of the sum of squares of m functions in n variables using a finite difference Levenberg–Marquardt algorithm (see IMSL 1984)]. The system was more sensitive to initial values, and it was possible for solution sets to have probabilities out of bounds.

The third method examined was due to Fellegi and Sunter (1969, app. II). The authors presented an algebraic solution for three fields, but it is easy to generalize the equations and solve the system of nonlinear simultaneous equations using numerical methods. The results agree with the other two methods used. However, the system is rather sensitive to the starting values, and in one case a penalty function had to be introduced to keep the probabilities within bounds.

3.4 Calculation of Threshold Values

An algorithm in our matcher determines the threshold weights as follows. There are 2^n possible configurations of agreement and disagreement of n components. These configurations can be ordered by the composite weight (the sum of the individual weights, w_i, for each component). After ordering the composite weights, the sum of Pr(• | **M**) and 1 minus the sum of Pr(• | **U**) can be calculated [see Eqs. (1) and (2), Sec. 3.2]. The maximum weight for a nonmatch decision is the weight of the configuration where the sum of Pr(• | **M**) does not exceed the desired probability that a matched pair should be classified as unmatched. The minimum weight for a match decision is the weight of the configuration where 1 minus the sum of Pr(• | **U**) does not exceed the desired probability that an unmatched pair should be classified as matched. Weights between these two thresholds are undecided cases.

For applications such as census matching, with approximately 10 components, this technique is computationally feasible. Unpublished experimentation has been performed by sampling component configurations for problems having many components, but the large number of cells makes it difficult to obtain a sufficient number of observations in each cell, so sampling error is not an overpowering factor.

4. MATCHING ALGORITHM

This section describes the basic operation of the matcher. Before matching, fields such as house address should be separated into components and spellings should be standardized. Both files must be sorted by the blocking variables.

4.1 Composite Weight Calculation

The matcher processes one block at a time, building a matrix (C) containing the composite weights for all pairs within the block being processed. The composite weights are computed by summing the individual weights for agreement or disagreement on each field (see Sec. 2.2). The simple agreement/disagreement dichotomy modeled by the theory is too simplistic for noncategorical fields. For example, character strings are compared using an information–theoretic character comparison algorithm that provides for random insertion, deletion, replacement, and transposition of characters. The weight assigned for such comparisons is prorated according to a measure of similarity between character fields (see Jaro 1978, pp. 106–108). If two character fields match exactly, the full weight for agreement is assigned to the comparison. If they disagree slightly, however, it would be wrong to assign the disagreement weight. Consequently, the weight assigned for the comparison will be somewhat less than the full agreement weight.

Similarly, weights for integer or continuous variables such as age can be prorated proportionally to the ratio of the difference and the minimum of the two values being compared (delta percent). For example, if age disagrees by one year in an 80-year-old man, it is less serious a mismatch than for a 1-year-old baby.

4.2 Assignment

After the matrix containing the composite weights for all pairs within the block is constructed (C_{ij} in the following), the records can be paired up (assigned). One record on file A can be assigned to one and only one record on file B, and vice versa. We wish to choose an assignment

scheme that maximizes the sum of the composite weights of the assigned record pairs. This is a degenerate transportation problem known as the linear sum assignment problem, which can be solved by a simple method requiring only addition and subtraction. The use of such a linear programming model to provide the assignments represents an advance over previous ad hoc assignment methods. The problem can be formulated as follows: Maximize

$$Z = \sum_{i=1}^{k} \sum_{j=1}^{k} C_{ij} X_{ij}$$

subject to

$$\sum_{j=1}^{k} X_{ij} = 1, \qquad i = 1, 2, \ldots, k,$$

and

$$\sum_{i=1}^{k} X_{ij} = 1, \qquad j = 1, 2, \ldots, k,$$

where C_{ij} is the cost (weight) of matching record i on file A with record j on file B, X_{ij} is an indicator variable that is 1 if record i is assigned to record j and 0 if i is not assigned to j, k_a is the number of records in the block being processed from file A, k_b is the number of records in the block being processed from file B, and k = maximum(k_a, k_b). If $k_a \neq k_b$, the matrix is made square (with dimension k) by inserting entries whose values are large negative numbers (less than any possible composite weight). This prevents these entries from being assigned.

An excellent discussion of the theory of assignment problems can be found in Cooper and Steinberg (1974, chap. 11). The computer algorithm was obtained from Burkard and Derigs (1981) and is highly efficient and economical of storage since the original matrix elements remain unaltered (the computations are performed on vectors, since all operations apply to entire rows or columns).

Once an optimal assignment vector is obtained, an assigned pair can be classified as a match if the composite weight is greater than the Fellegi–Sunter threshold value. After all assigned pairs in the block are processed, the records for the next block can be read.

4.3 Duplicates

Duplicates can be detected by examining each row or column of the assignment matrix. If more than one entry is above the cutoff threshold, then there is a possibility of a duplicate. Two similar records on both files would probably be two separate individuals (a father and son, for example), but two similar records on only one file would probably be a duplicate.

4.4 File Preparation

To match any file, free-form information must be standardized. This is especially true of fields such as street address and person name. The components of the name should be separated into individual fields (given name, middle initial, and surname). This is much more effective and accurate than trying to match an entire name as a single character string. For street address, the various components of the address should be placed in individual fields and the spellings of common abbreviations (such as BD, BLVD) should be standardized. Punctuation should be removed from the fields.

The technique of SOUNDEX encoding (Knuth 1973, pp. 391–392) is a method of transforming a person's name into some code that tends to bring together all variants of the same name. For example, *Smith* and *Smythe* would both be coded as S530. Surname is often an important blocking variable. To maximize the chance that similarly spelled surnames reside in the same block, the SOUNDEX system can be used to code the names, and the SOUNDEX code can be used as a blocking variable. There are better encoding schemes than SOUNDEX, but SOUNDEX with relatively few states and poor discrimination helps ensure that misspelled names receive the same code.

SOUNDEX is not recommended for matching nonblocking variables, since nonphonetic errors result in different codes and different names may receive the same code.

5. TAMPA MATCHING METHODOLOGY

This section describes the computer match of the 1985 census of Tampa, Florida, to the PES. The object of the matching study was to identify all individuals who responded to both the PES and to the census. The records consisted of individual data and contained name, address, and demographic characteristics. The primary goal of the computer matcher was to eliminate the first-level clerical match (where matches could be determined with relatively unsophisticated personnel). The system exceeded this goal. A multiple blocking strategy was used to increase the numbers of matched records given errors in the blocking variables. The strategies are called Pass I and Pass II, respectively.

5.1 Pass I Match

The following variables were used for matching:

1. Census block numbering area (CBNA) (blocking variable)
2. Census block number (blocking variable)
3. Surname (SOUNDEX) (blocking variable)
4. Given name (m = .98, u = .09)
5. Middle initial (m = .35, u = .03)
6. Relation to head of household (m = .39, u = .20)
7. Sex and marital status (combined) (m = .82, u = .21)
8. Birthdate (m = .94, u = .04)
9. Race and Hispanic origin (combined) (m = .90, u = .67)
10. Street name (m = .96, u = .03)
11. House number (m = .99, u = .01)
12. Apartment number (m = .35, u = .26)

The blocking variables for Pass I were census block numbering area (CBNA), census block number, and SOUN-

DEX code of surname. CBNA and census block number were used as blocking variables, since only census data for PES sample blocks were keyed and, consequently, it would be unlikely that data would be available for units geocoded to incorrect blocks. All records failing to match in Pass I would participate in the Pass II match, which used different variables for blocking.

The results of the Pass I match were as follows: 7,358 PES records read, 8,798 census records read, 4,375 matched pairs, 165 nonclassified pairs, 628 unmatched PES records, 702 unmatched census records, 2,190 skipped PES records, and 3,556 skipped census records. Records are said to be *skipped* when one or more of the blocking variables do not match. The nonclassified, unmatched, and skipped records are input to the Pass II process.

5.2 Pass II Match

In an attempt to match records that failed to match in Pass I, an independent blocking scheme was chosen for Pass II. The blocking variables were CBNA, census block number, SOUNDEX of street name, house number, and apartment number. CBNA and block number were reused, since no records exist outside of the sample area and detecting geocoding errors would be unlikely.

Apartment number had to be used, since some high-rise developments contained more than 500 units at a single address and the matcher has a maximum block size that cannot be exceeded. Subsequently, the matcher was modified to correct this problem by flagging such "overflow" blocks, which could be processed in a separate subsequent run.

A blank apartment number may mean either that the apartment number is not appropriate or that the value was not reported. Since apartment numbers are sometimes not appropriate, blank apartment numbers would be accepted as a valid value.

The Pass II match was useful, since it displayed record pairs and groups in household sequence.

The results from Pass II were as follows: 2,983 PES records read, 4,423 census records read, 212 matched pairs, 885 nonclassified pairs, 1,114 unmatched PES records, 1,321 unmatched census records, 772 skipped PES records, and 2,005 skipped census records. The matching decisions were made very conservatively to limit the number of false matches. This is important where estimation of relatively rare events is required (such as for undercount estimation). The tight error tolerances account for the high number of nonclassified cases. Most of these could be resolved quickly, since records are already paired. All nonclassified pairs, unmatched records, and skipped records would be processed clerically. Many records were unmatched because of construction and demolition in the PES area, vacancies, noninterviews, proxy data, geocoding errors, etcetera.

The total number of records matched automatically from both passes was 4,587, with 885 nonclassified pairs that could be rapidly resolved. Approximately 40 minutes (wall-clock time) were required to conduct the Pass I match on an IBM PC-AT, with 20 minutes required for Pass II. Sorts required about 15 minutes each.

6. CLERICAL REVIEW AND FINAL MATCHING RESULTS

After the computer matching was completed, the records were grouped by household and printed on a computer-generated matching form for clerical review. Many of the nonmatches were easily converted to matches by reviewing the persons in the household together.

A total of 5,343 persons were matched in the entire process, with 4,587 matched by the computer (85.9%). Of the 885 nonclassified persons, 225 were at vacant addresses or were noninterviews, leaving 660 persons that could be resolved clerically. Of these 660 persons, 83.39% were determined to be actual matches. The number of persons who were either matched automatically or with a quick verification of the computer-assigned possible match is 5,177 (computed by 4,587 persons matched automatically plus 83.39% of 660 persons). A total of 5,177 out of the 5,343 cases yield an effective match rate of 96.89% for the automated system, leaving only about 3% of the cases for extensive clerical intervention. The clerical and professional review staffs were able to match only 19.47% of the residual nonmatched records from the computer operation.

6.1 Review of Computer Matches

All of the matches assigned by the computer were reviewed to evaluate the computer matching. Eight persons assigned by the computer matcher were actual errors, yielding an error rate of .174% (8 of 4,587). Inferences regarding the matcher's accuracy, however, should not be made from this one case study, as results can vary with the accuracy of the geographic reference material, data entry procedures, and numbers of hard-to-count population groups in a specific area.

6.2 Matching in Neighboring Blocks and Duplicates

The census questionnaire information was only entered for the PES sample blocks. Therefore, it was impossible to detect geographic coding errors automatically by the computer (since no machine-readable data were available). The blocks that bordered the sample blocks, however, were searched clerically to attempt to reduce the number of nonmatches.

Of 1,692 persons not matched in the sample blocks, 726 were matched to neighboring blocks, resulting in a 42.9% reduction in the nonmatch rate. The largest reduction was in blocks that were predominantly black and Hispanic and contained multi-unit structures.

The matching processes (both automatic and clerical) include the detection of duplicate records for an individual. Searching on neighboring blocks uncovered 32 census duplicates. The total number of duplicates within the sample and surrounding blocks was 145 (2.6% of the matched records).

7. CONCLUSIONS AND FUTURE WORK

The automated matching system greatly exceeded expectations in terms of both match rate and accuracy: 96.89% of the records that were matched were matched either automatically or could be quickly verified. The error rate was .174% (again, no inferences should be drawn from this). The 1986 test census of Los Angeles, California, includes an automated extended search to detect movers and geocoding errors. Several matching errors were due to problems in high-rise multi-unit structures. A new methodology that should eliminate these problems has been developed.

The use of the EM algorithm for parameter estimation appears to be the most promising of all techniques attempted in terms of insensitivity to choice of starting values and ease of implementation.

Additional work is required in a number of areas. Questions to be answered include: What is the sensitivity of the final classification to parameter estimation error and statistical dependence of reporting errors and/or value states of fields? Can successful models for multiple state comparison vectors be developed? For example, instead of agreement and disagreement states, the vector could be augmented to include cases where values are missing (currently, these receive zero weight). Can cutoff thresholds be computed by means of a closed-form equation without enumerating 2^n configurations, or can such a form be developed for changes in only one weight? If so, then weights could be adjusted by means of a Bayesian procedure to account for changes in the distribution of the value states of a field in different geographic areas, and, further, weights could be adjusted where value state distributions are likely to be skewed. For example, an agreement on a name like Humperdinck should carry more weight than an agreement on Smith, but strictly speaking, the cutoff thresholds would change if the weights for a field were changed, and a closed-form equation for the thresholds would permit changing the cutoff values for particular cases during the matching process. Can the errors in both the automated and manual phases of matching be properly modeled, and can variances be computed?

A calibration data set is being developed from the Tampa, Florida, experience. A linked file such as this can be used to measure sensitivity and the relative merits of various matching schemes. I will attempt to answer systematically most of the questions posed in this article and to improve the mathematical underpinning of record-linkage methodology.

[*Received January 1987. Revised October 1988.*]

REFERENCES

Burkard, R. E., and Derigs, U. (1981), "Assignment and Matching Problems: Solution Methods With FORTRAN-Programs," in *Lecture Notes in Economics and Mathematical Systems* (No. 184), New York: Springer-Verlag, pp. 1–11.

Citro, C. F., and Cohen, M. L. (1985), *The Bicentennial Census, New Directions for Methodology in 1990*, Washington, DC: National Academy Press.

Cooper, L., and Steinberg, D. (1974), *Methods and Applications of Linear Programming*, Philadelphia: W. B. Saunders.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society*, 39, 1–38.

Ericksen, E. P., and Kadane, J. P. (1985), "Estimating the Population in a Census Year: 1980 and Beyond" (with discussion), *Journal of the American Statistical Association*, 80, 98–131.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183–1210.

International Mathematical and Statistical Libraries, Inc. (1984), *User's Manual*, Houston: Author.

Jaro, M. A. (1978), "UNIMATCH: A Record Linkage System, User's Manual," Washington, DC: U.S. Bureau of the Census.

——— (1985), "Current Record Linkage Research," in *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 140–143.

Kelley, R. P. (1984), "Blocking Considerations for Record Linkage Under Conditions of Uncertainty," in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 602–605.

Knuth, D. E. (1973), *The Art of Computer Programming, Volume 3: Sorting and Searching*, Reading, MA: Addison-Wesley.

Newcombe, H. B., and Kennedy, J. M. (1962), "Record Linkage," *Communications of the Association for Computing Machinery*, 5, 563–566.

Wolter, K. M. (1986), "Some Coverage Error Models for Census Data," *Journal of the American Statistical Association*, 81, 338–346.

Matt Jaro, Founder, President & CEO:

MatchWare Technologies Inc.'s founder, CEO and director of technology, Matt Jaro, enjoys 30+ years of experience in the science of probabilistic record linkage and information technology. Jaro has led MatchWare to a position of global leadership in the practical implementation and use of this methodology.

Matt Jaro conceived, designed, and authored MatchWare's proprietary software. Well known as a speaker and author in the fields of address matching and geographic information systems, Matt is regarded as an international authority on probabilistic record linkage methodology.

Jaro holds degrees in mathematics from California State, and computer science from George Washington. Prior to founding MatchWare, he held a variety of information technology positions with public and private sector organizations including: Booz-Allen Applied Research, the U.S. Census Bureau, The Corporation for Applied Systems, Public Technology, Inc., and System Automation.

In the mid-80's, Jaro was a principal researcher at the U.S. Census Bureau where he developed the mathematical methodology and software to perform statistically valid matching procedures in support of estimating census coverage. Although application specific, the census estimation methodology Matt developed was precedent-setting in the field of probabilistic record linkage.

Chapter
11

Record Linkage and Genealogical Files

Nancy P. NeSmith
The Church of Jesus Christ of Latter-Day Saints

Whenever there are large computerized genealogical files the problem of duplication of records for the same individual or family within the file always exist. Many indexing schemes can be used which allow some matching of entries being added to the files with slight variations, but computer technology in the past has been limited in matching entries in which more than one field such as surname, given names, date or locality have disagreement.

Genealogists know that disagreement comes through different record sources used for identification or through transcription errors. Bringing together records with discrepancies has always been a genealogical nightmare. If the records don't even come together using various sorting schemes, how can the records be analyzed for matching or merging decisions? In other words, how does one know if two different records refer to the same individual or family?

One solution to this dilemma is the use of Record Linkage theory. Record linkage refers to a computer program which uses a detailed algorithm based on probability to determine if two records being compared represent the same individual. This technique, developed in Canada by Howard B. Newcombe (1988), has been used in the statistical, demographic, and medical disciplines to identify and link two or more records representing the same person or entity.

The theory underlying record linkage was developed around the need for an algorithm which would mimic human decision making in comparing a record from one file with a record from a second file. To do this, two records which represent the same person are studied and field comparisons made. Fields are items of information in the record, such as given name, surname, birthdate, birthplace, etc. The outcome of this comparison (agreement, disagreement or partial agreement) that is common in linked records is noted. If there is enough agreement, the probability is high the records being compared from the two different files represent the same individual. If the comparison outcome is more common to unlinked records, the probability is high the records being compared represent two different individuals.

Using the comparison statistics, a record linkage system computes the odds in favor of a match or against a match for any two records selected for comparison. For example, if a surname in two records matches, the computer calculates the odds of the two names matching by chance and how often the surname field agrees in the truly linked records contained in the comparison file. From these two statistics the program determines a score which represents the odds above chance the two surnames matching are for the same person.

Each algorithm may be tailored to the uniqueness of the genealogical data elements in its geographic area. This eliminates applying an "English" standard to all geographic areas of the world. Another advantage is the algorithm may be refined to specific cultural or variable record types.

Reprinted with permission from the *Utah Genealogical Journal* (1992), 20, 3 and 4, 113-119.

To develop the algorithm, samples of files which need to be matched or merged are examined by

specialists to locate duplicates for statistical analysis. Based on their analysis and the purpose of the linkage for the file, the specialists choose blocking, weighting, and threshold parameters which will be used by the computer for each geographic area to determine if the records being compared are a match.

Searching the File -- Blocking

When searching a file to see if there is a record which matches the request, it would be ideal to compare every record in the file with the request. However, this is not practical in most data bases so an indexing scheme is used to retrieve only the entries in the file which are most likely to match the request. This is called blocking or retrieval. The intent is to reduce the number of comparisons the computer must make. The implicit assumption is that only records with a reasonable chance of being linked are retrieved.

Blocking effectiveness can be described in terms of "recall" and "precision." Recall is a measure of how many relevant records in the file are included by the blocking scheme. Precision is a measure of how many of the total records retrieved by the blocking scheme are relevant.

For example, if you're looking for a record of Joseph Jones, and the file contained two records for him, one in which he is identified as Joseph Jones and the other as J. Jones, the system would exhibit good recall if it retrieved both records. However, the system tuned to recall near matches such as J. Jones may retrieve irrelevant entries where the letter J. stood for John or James rather than Joseph. These irrelevant entries are known as "noise."

Precision measures the amount of noise. A problem with a system tuned for precision over recall is relevant entries can be missed because the narrower search parameters used to limit the noise also limit the recall. Whenever you tune for recall you increase the noise; when you tune for precision you decrease recall. The two concepts have been found to be in opposition. The goal is to find an acceptable balance between the two which suits each specific application.

It is possible to enhance the recall of a system without greatly reducing the precision by using some form of authority control to bring together the equivalent names of people which are spelled differently and locality names which are different but refer to the same locality. Blocking schemes are tuned to the specific file or part of the file being searched. Those fields which are accurate, discriminating, and most often present in the records are chosen because they help give a balance between precision and recall.

Weight Calculations

Using the blocking parameters the computer retrieves a set of records which can now be compared in detail with the query to determine their similarity to it. As fields in the query and candidate record are compared, a statistical score or weight is computed which reflects agreement, partial agreement, or disagreement of the two fields being compared. A positive weight is calculated for agreement, a smaller positive weight is calculated for partial agreement, and a negative weight is calculated for non-agreement on that field. If either record has missing information in the field being compared, a weight of zero is assigned. The weights are added to each other to obtain a total weight which reflects the similarity of the pair of records being compared.

The weights are tailored to the locality or record source. For example, surnames for England agree more often than surnames in Denmark and other countries which have patronymic surnames. This affects the weights. England would have a smaller positive weight for agreement on surname than Denmark but would have a higher negative weight for disagreement on surname than Denmark because the surnames seldom disagrees for England. Also taken into consideration for the calculation of the weight is the relative size of the name pool. For example, there are fewer surnames in England than there are in United States records. The

fewer the names, the less significant is the agreement. These types of calculations and comparisons are done on the fields for gender, names, localities, and dates.

It is not necessary to weight all of the fields in a record. Generally fields which were used as blocking parameters are not weighted. The fields are not weighted because the records retrieved have already matched on these fields and weighting them only increases the overall score of each record by the same amount. Other fields may not be weighted because they are not statistically discriminating and don't contribute significantly to the equation.

Threshold Determination

Once the file records have been retrieved using the blocking scheme, compared field by field to the query, each field weighted, and each total record's weight determined; then a decision can be made about whether or not a duplicate was found. The total weight which is used to decide whether a record should be considered a match or non-match with the retrievals from the file is called the threshold.

Generally, scores above a certain threshold indicate a match and those below it indicate a non-match. For example, if the weight of 40 is considered the threshold, then all retrievals scoring less than 40 are considered non-matches. All retrievals with scores of 40 and above are considered matches. The threshold decision is based on the total weights of truly matched records for that specific locality (truly matched means the two records are known to refer to the same person).

There is often a small range of scores which includes intermingled matches and non-matches. This is called the gray area. For example in a study of criminals and law abiding citizens, the range of scores could be from -150 to 150. Criminals have scores ranging from -150 to +50. Citizens had scores ranging from +30 to +150. Everyone with a score below 30 is a criminal, everyone with a score above 50 is a citizen. Those with a score between 30 and 50 could be either a citizen or a criminal. This is the gray area for this study, meaning if 40 is picked as the threshold score there is the possibility of a non-match scoring high enough to appear to be a match (false positive) or there is a possibility of a match scoring so low it appears to be a non-match (false negative). Which number to pick is called the threshold decision.

In making a threshold decision it is important to decide the purpose of the links. For optimal linkage it is important to follow the rule which states that before a threshold is picked, decide the purpose of the links. If the goal is to arrest all the criminals in a town and not let any of them go free, then a threshold of 50 would be picked. But as a result, some law abiding citizens would be arrested because their score would be similar to criminals. If the goal is to arrest as many criminals in a town but to not falsely arrest any citizens, then a threshold of 30 would be picked. As a result, some criminals would go free, but no citizens would be arrested.

The Family History Department and Record Linkage

The theory underlying this technology is the best approach known to the scientific community. For this reason, the Family History Department of The Church of Jesus Christ of Latter-day Saints has chosen to implement its usage in their genealogical systems and databases. It is currently being used to retrieve entries within the department's genealogical files, and will be used in match and merge decisions. The results have been very satisfying and its efficiency has been improved by taking full advantage of name and locality authority systems. The use of record linkage for the massive files and record linking needs of Family History makes the most efficient use of the Department's computer resources in eliminating or matching duplicates in their files. This technology has not been employed in the Personal Ancestral File7 as that program was developed before the implementation of Record Linkage in 1988.

References

Newcombe, Howard (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business*, Oxford: Oxford University Press.

Wrigley, E. A. (ed). (1973). *Identifying People in the Past*. London: Edward Arnold.

* Nancy P. NeSmith, 5440 South Lighthouse Road, Salt Lake City, Utah 84123. Miss NeSmith received a BS in Genealogy and undertook graduate studies in Family History at Brigham Young University. She is currently a Systems User Specialist in the LDS Family History Department.

Chapter 11

A Review of the Statistics of Record Linkage for Genealogical Research

As Used for the Family History Library, Church of Jesus Christ of Latter-Day Saints

David White, Utah State University and
Church of Jesus Christ of Latter-Day Saints

Introduction

The Church of Jesus Christ of Latter Day Saints maintains massive genealogical files, which consist of millions of names. The two largest files are the International Genealogical Index (IGI) and the Ancestral File. The IGI contains over 200 million individual vital records and, because of its size, is divided into geographical subfiles. The Ancestral File contains over 21 million names arranged in family groups and pedigrees. These files are growing, and one of the major challenges is to be able to query these files in such a way that correct records are retrieved for various genealogical purposes and adding duplicates to these files is avoided. Record linkage is used for this purpose.

For this paper, a *record* will be defined as the collection of items that refer to a specific event, such as a birth or christening. Each item for the event, such as the day**,** month, year of birth, surname, and given names of the father and mother, is stored in what are called *fields* in computing terminology. Two records are defined as "linked" if the odds are high that they represent the same person. One of the first challenges for record linkage is finding those fields that are useful for calculating these odds. Although all the records in the IGI are birth and marriage events, they come from various sources. The same is true for the Ancestral File. For example, a birth record for the same individual may come from a civil record, an ecclesiastical record, or a family source. This may result in multiple records in the file for the same person when the available information is sparse or varies.

Comparing a Pair of Records (Calculating the Odds)

We begin with statements about the comparison of field entries coming from two records when it is known that these records refer to the same person. Such records are termed "matched" or "duplicates" by researchers. The records may or may not be from the same source. An example of different sources containing records about the same person would be births coming from civil records and ecclesiastical records. An example where only one source is involved would be ecclesiastical records about the same person who has moved from one jurisdiction to another within the same denomination, and the record keeping agency includes both jurisdictions.

Next, consider birth records and, within a birth record, the field containing the given name of the mother. We consider a pair of birth records and desire evidence to either confirm or deny that these records represent the same person. Suppose the given name of the mother shows up in both records, and we have n pairs of such records, which are matched (i.e., each pair is known to refer to the same person). Further, suppose that in k instances, the mother's given name for one record of the pair is the same as that for the other record. Then, the probability that the given names are the same when the records are matched is estimated by k/n, and we use the equation

$$P(S|M) \cong k/n \qquad (1)$$

where ≅ means that the two sides of the equation are "close," although they may not be exactly equal. We read P(S|M) as the probability that an entry in a specific field is the same for both members of a pair, given that we

have a matched pair.

Next, consider the probability that the given names of the mother are the same when the two records are randomly paired. Such pairs of records are termed "unlinkable" by Newcombe. As an example, suppose we have a file with a total of m different given names appearing for the mothers of the child. Then, a typical event describing a pair with the same given names is:

Both given names are Dorothy, or
Both given names are Phyllis, or
.....
.....
Both given names are Agnes,

where Agnes completes the total of the m given names appearing in the records.

First, assume that the records come from the same file and that there are N_1 records with the name Dorothy, N_2 records with the name Phyllis, and so on, to N_m for Agnes. If $N_1 + N_2 + \ldots + N_m = N$, then the probability that one element of a pair is Dorothy will be estimated by N_1/N and that both elements are Dorothy will be estimated by $(N_1/N)^2$ -- we multiply the probabilities for the two elements together since the events are independent (any two records were randomly paired). Since we allow that both elements having the same given name can happen with any one of the m alternatives, we add the probabilities for the m possible given names together, to get

$$P(S) \cong (N_1/N)^2 + (N_2/N)^2 + \ldots + (N_m/N)^2 \cong \Sigma_{j=1}^{m} (N_j/N)^2 . \quad (2)$$

We read P(S) as the probability that two corresponding elements of a pair are the same when the records have been selected at random.

Now, assume that the records come from different files, and that there are m first given names for mothers in common between the two files. Assume that there are L_1 records in the first file with the name Dorothy, and N_1 corresponding records in the second file, or in general, L_j and N_j records with the j^{th} given name.

Then, the probability that the given name of the mother will be the same is estimated by

$$P(S) \cong (L_1N_1)/LN + (L_2N_2)/LN + \ldots + (L_mN_m)/LN \cong \Sigma_{j=1}^{m}(L_jN_j)/LN, \quad (3)$$

where now,

$$L = \Sigma_j L_j \quad \text{and} \quad N = \Sigma_j N_j \quad (4)$$

and the summation of the subscript j is not limited to the m alternatives in common, since each file may have alternatives not in common with the other file.

We now have P(S|M), the probability that two elements in a pair are the same, given that the pair is matched, and P(S), the probability that they are the same when they have been paired randomly. For the rest of the paper, we will deal only with the case where the two elements come from the same file; the other case

corresponds in the same way as described above.

Another way to estimate P(S) is to actually create a set of randomly matched pairs and calculate the proportion of matches obtained. This way is computationally less intensive and may be a practical alternative for people with more meager computational resources.

We next consider the probability law:

$$P(M|S)P(M) = P(S|M)P(S). \quad (5)$$

What we want is P(M|S), which is the probability that two records of a pair do, in fact, represent the same person when the first given names of the mothers are the same. P(S|M), which we have, is the probability that the elements of a pair are the same when they are matched. Using equation (5), we get

$$P(M|S) = [P(S|M)P(M)]/P(S). \quad (6)$$

This is an application of what is sometimes called "Bayes' Rule," being used more often in recent years and which has caused a good deal of controversy in the statistical community; it has been used successfully in Record Linkage.

Perhaps the pair of records does, in fact, represent the same person, even though the records of birth give a different given name for the mother. We then want $P(M|S^c)$, where S^c means the two elements in a pair do not agree. (S^c is read as the "complement" of S). Then, the analogue of equation (6) gives

$$P(M|S^c) = [P(S^c|M)P(M)]/P(S^c). \quad \text{Further,} \quad (7)$$

$$P(S^c) = 1 - P(S) = 1 - [(N_1/N)^2 + (N_2/N)^2 + ... + (N_m/N)^2], \text{ and} \quad (8)$$

$$P(S^c|M) = 1 - P(S|M) = 1 - k/n, \quad (9)$$

so that we get for the probability of a match when the elements are not the same,

$$P(M|S^c) = [(1 - k/n)P(M)]/\{1 - [(N_1/N)^2 + (N_2/N)^2 + \ldots\ldots + (N_m/N)^2]\}. \quad (10)$$

Note that k and n are different from N_1, N_2, . . ., N_m or N. This is because they come from a sample of duplicates of size n, whereas N_1, N_2, . . . , N_m are the total numbers of records in the file for each of the names. Recall that k is the number of pairs of records in the set of duplicates (or matches) for which the given name of the mother is the same. P(M) in equations (5), (6), (7), and (10) is the probability that two records "match" (represent the same person) when they have been paired at random. It will be very small.

Next, let E be the event describing whether the mothers'given names are the same, not the same, or missing. Then,

$$P(M|E) \cong P(M) \text{ times } (k/n)/ [(N_1/N)^2 + (N_2/N)^2 + \ldots\ldots + (N_m/N)^2] \quad (11)$$

if the names are the same

and $P(M|E) \cong P(M)$ times $(1 - k/n)/\{1 - [(N_1/N)^2 + (N_2/N)^2 + \ldots\ldots + (N_m/N)^2]\}$

if they are different.

We further define P(M|E) = P(M) if one or both elements are missing in the record pair. This makes sense, since E tells us nothing new about the match when the information is missing.

Since virtually all records contain more than one element or "field," we must allow for this in our formulas. We let Q be the number of elements or fields common to both records and consider the i^{th} field, where i ranges from 1 to Q. Then, let n_i be the number of elements with both entries present for the i^{th} field in the sample of duplicates, and k_i be the number of element pairs in the i^{th} field which are the same. Letting E_i be the event for the i^{th} field (same, not same, or missing), we get the following:

$$P(M|E_i) \cong P(M)P(E_i/M)/P(E_i) = P(M) \text{ times } (k_i/n_i)/[\Sigma_j(N_{ij}/N_i)^2], \quad (12)$$

when the i^{th} elements are the same and where the summation Σ_j is over all possible values of j in N_{ij}/N_i for the i^{th} element or field. Note that N_{ij} now has two subscripts, the first subscript (i) to account for the field and the second (j) to account for the altemative values for the i^{th} field. The range of j is from 1 to J_i because there are a different number of alternatives in each field. For example, there are two alternatives for gender and, in our case, m altematives for the given name of the mother.

If the elements for the i^{th} field are not the same, the formula is:

$$P(M|E_i) \cong P(M)P(E_i/M)/P(E_i) = P(M) \text{ times } (1 - k_i/n_i)/\{1 - [\Sigma_j(N_{ij}/N_i)^2]\} \text{ and} \quad (13)$$

$P(M|E_i)$ = P(M) when one or both of the i^{th} elements are missing.

$$P(E_i|M)/P(E_i) \quad (14)$$

can be referred to as the "odds" in favor of a match, given the event E_i with respect to the i^{th} field. Note that P(M) does not appear in (14). It does appear in (12) and (13), however, which are the probabilities that two records refer to the same person, given the event E_i for the i^{th} field. There are Q events for each pair of records (an event for each of the fields). Next, we consider

$$P(M|E_1, E_2, \ldots, E_Q), \quad (15)$$

which is the probability that both members of the pair represent the same person when events E_1 and E_2 and ... E_Q have occurred. If most of the paired fields are the same, this probability will be close to 1, and we should conclude the pair is "linked," as distinguished from the cases where they are known to have been matched by prior identification of duplicates. If most of the paired fields are not the same, (15) will be close to zero, and we conclude the records are not a match. There is a gray area in between where the evidence is not conclusive. We assume that the events $E_1, \ldots, E_Q$ are independent (that is, that one pair of fields being the same tells us nothing about the Asameness@ of any other pair).

If we assume this, we get the formula:

$$P(M|E_1, E_2, \ldots, E_Q) = P(M|E_1)P(M|E_2) \ldots P(M|E_Q) = \Pi_{i=1}^{Q}P(M|E_i) \quad (16)$$

$$= \Pi_{i=1}^{Q} P(E_i|M)P(M)/P(E_i)$$

$$= P(M)[\Pi_{i=1}{}^{Q} P(E_i|M)]/[\Pi_{j=1}{}^{Q} P(E_i)].$$

(Note that $\Pi_{i=1}{}^{Q} P(E_i|M)$ means to take the product of the $P(E_i|M)$ as the subscript j ranges from 1 to Q; similarly for $\Pi_{j=1}{}^{Q} P(E_i)$ in the above expression.)

The "odds" in favor of the records representing the same person are calculated as the probability of the above events when the records are matched, divided by the same probabilities when the records are randomly paired -- which is the last expression of (16), except that P(M) would be dropped. Referring to the second page of NeSmith's paper, the probability of the two names "matching by chance" is

$$\Pi_{j=1}{}^{Q} P(E_i), \tag{17}$$

while the probability of the names being the same in the "truly linked records" is

$$\Pi_{i=1}{}^{Q} P(E_i|M). \tag{18}$$

These are the two statistics used to calculate the odds. As before stated, this is (18) divided by (17).

Blocking

Finding the matches or duplicates (those pairs which are known to represent the same person) involves the time of experienced researchers who must consider a large sample of record pairs and find those which will be identified as duplicates. If all possible pairs are to be considered for the cases of interest, we will have an impossible task before us, with literally billions of pairs to evaluate. To cut down on the enormity of this task, we attempt to gather together records, which are likely to be matches, by sorting on fields, which will put potential matches close to each other in a listing of available records. Such fields usually include a surname code (such as Soundex), a given name code, and possibly a range for birthdates, and a county identification of some kind. If these four fields were used, a listing of records would put people together if they had the same surname code, given name within the surname code, birthdate range within the names, and the same county.

A block is defined as the set of records whose pairs are the same with respect to a set of fields, such as the above four fields. Each distinct set of fields used for this purpose is called a *blocking scheme.* The records whose blocking fields match will be adjacent to each other in a file, which has been indexed on the basis of these fields. Such a list can be constructed with any good data base management system. A block may, and probably will, contain a number of records, which are not duplicates; but a qualified researcher can browse the list and determine which of the pairs within the block should be considered as representing the same person. The size of the block should be modest -- not more than 10 to 20 records, so that the worker can compare them on a monitor screen. In order to find as many duplicates as possible, this process must be repeated for several blocking schemes. Even then, the number of blocks for a data bank may be too numerous to make searching all of them for duplicates feasible. Then, a subset is used, such as some representative date ranges. One of the problems of interest is how large the sample of duplicates obtained by the workers should be. Current practice is to find about 1,000 to 1,500 duplicates -- a substantial amount of work.

One blocking scheme will often have better properties than another. Measures of how good a scheme is include:

- **Blocking Recall**. -- It often happens that when a second blocking scheme is used, there will be a few of the duplicate pairs found in the first scheme, which will now be separated; that is, the two members of the pair will not show up in the same block. Since our searching procedures only look for record matches within the same block, such duplicates will not be detected using the second scheme. If we

now consider several blocking schemes, the percentage of known duplicate pairs, which are picked up with any one of the schemes, may well be less than 100%. Hopefully, we will find one of them, which picks up a higher percentage of duplicate pairs than do the others. *Blocking recall* is defined as the percentage of known duplicates, which are identified with a particular blocking scheme.

- **Block Noise**. -- This is the number of non-duplicates in the blocks divided by the total of the block sizes in the blocking scheme. Greater block noise requires more computing time for a search, but recall for the scheme is usually better.

- **Block Precision**. -- For a blocking scheme, this is the ratio of the number of duplicates to the number of non-duplicates in the blocks, multiplied by 100. A blocking scheme with high precision has mostly duplicate pairs within the block. There are not many non-duplicates.

The greater the precision, the less recall, as a general rule, as indicated on the third page of NeSmith (1994). Her comment about increasing recall without seriously reducing precision relates to the use of a name code, such as Soundex, and a place code, which different versions of place names are tied to. This can be considered as a partial agreement for the fields concerned, and the blocks that use these fields will be somewhat larger, including proper and/or place names, which are "close" to each other. We decrease the block noise and increase the block precision by increasing the number of fields used for blocking. Fewer fields, conversely, increase both noise and recall.

Calculating the Weights

The weights are obtained from the odds by taking logarithms. Using equation (16), we take logarithms, to get

$$\log P(M|E_1,E_2,..,E_Q) = \log P(M) + \Sigma_{i=1}^{Q} \log\{P(E_i|M)/P(E_i)\}. \quad (19)$$

Note that P(M) is a constant term, which factors out of (16), and is simply an additive constant in (19). Such a constant does not influence the results. We can drop this constant and simply consider the term

$$L = \Sigma_{i=1}^{Q} \log\{P(E_i|M)/P(E_i)\}. \quad (20)$$

The Aweights@ referred to in the NeSmith paper are the individual terms

$$w_i = \log\{P(E_i|M)/P(E_i)\}. \quad (21)$$

For each field, there are three weights -- one for when the two field entries are the same, one for when they are not, and zero for when one or both entries are missing. This weight will be a positive value when the two fields are the same; it will tend to be negative if the entries for the two fields are different.

If the probability is high that the records are for the same person, then most of the E_i will be in agreement (the i^{th} elements are the same for most of the i), and the sum of the weights (sum of the "log-odds") will be high, usually positive. If the probability is low, then most of the E_i will not be in agreement, and the sum of the weights will be low, usually negative. If data are missing in the i^{th} field, so that $P(M|E_i) = P(M)$, then from (12), $P(E_i|M)/P(E_i)=l$ and $\log[P(E_i|M)/P(E_i)]=O$. That is, the weight is zero if one or both field elements are missing. Fields with missing elements, therefore, neither add to nor subtract from the evidence we are interested in.

The researcher must identify each of the pairs in the subset of blocks as either a duplicate or a non-duplicate, and the weights are calculated from the duplicate pairs in the blocks, plus a set of counts (the

N_{ij}) for each field. The counts do not come from the blocks but from the complete set of records to be linked --that is, from the entire file. Note that as per the comment in NeSmith on the fifth page, **the weights are not calculated for the fields used as blocks because those fields always are the same for the records in the blocks, whether duplicates or not, and thus have little or no discriminating power**. For the same kind of reason, some fields have poor discriminating power because they do not change a great deal in some files – such as a geographical area in which a few family names predominate. A small amount of variability in a field reduces its usefulness for linkage algorithms. Algebraically, this shows up in the denominator of (18)/(17), above, because the chance of the fields being the same with random pairing becomes larger. But this is (17), which thus decreases the odds for a link when the field entries are the same.

Thresholds

Many genealogical tasks involve a search for someone in a large data bank. This search is usually termed a "query," and the framework for this is the set of linkage algorithms described above. One uses the fields available for the person to be searched for, chooses a blocking scheme employing some of those fields, and then searches the block into which the person being searched for fits. If a record in the block, when paired with the query, has the sum of the weights higher than a value called the *threshold*, a link has been found for the query. High recall for a blocking scheme means that there is a good chance of finding this link if it exists---but since high recall goes with more "noise," more computing time is involved. If a large number of queries are involved, the computing time may become an important issue.

The *threshold* is simply a constant value, C, which is a cutoff point for L, the sum of the log-odds for a pair. We consider the pair as representing the same person if L is greater than or equal to C; that is, the pair is "linked." The pair is not linked (i.e., considered as representing different people) if L is less than C. As an illustration for thresholds, we consider five sets of date ranges, which were used with Norway data (1736- 1755), (1781-1794), (1805-1814), (1836-1845), (1866-1875). These subsets of the complete set of data were used because finding duplicates for the complete set would have been too time-consuming. Several blocking schemes were used for identifying duplicates. For each blocking scheme and for each block in the scheme, all possible pairs were obtained, and the worker identified each pair as either a match (i.e., a duplicate) or a non-match. The scheme chosen as best for linking on the basis of precision and recall used the fields: birth year, birth county code (to standardize county names), the given name code for the principal (whose birth is recorded), and the father=s given name code. A block consisted of all records, which were the same for all four of the above fields. The fields used for weighting purposes were:

- the latitude minutes of the birth town
- the birth day
- the birth month
- the death day
- the death month
- the death year
- the mother's given name code, and
- the mother's surname code.

The N_{ij} were obtained with the computer from all records in the Norway File. The weights were then calculated with computing facilities for each of the eight fields, according to the formulas described in the preceding section and using the duplicates identified by the researchers. Next**, for each pair of records within each block (duplicates or not) and for each field in the pair**, the weights are then used. The total of the weights is obtained to get the value for the sum of the log-odds (see equation (17)). We now have a set of weight totals for the duplicates (matched pairs) and another set for the non-duplicates (unmatched pairs). A frequency histogram was obtained for both the matched and the unmatched pairs; these appear in Table 1.

Table 1. -- Frequency Distributions for Matched and Unmatched Pairs

Class Limits for Weight Totals	Unmatched Pairs	Matched Pairs
-34.35 to 0	-27.56	0
-27.55 to 0	-20.76	6
-20.75 to -13.96	257	0
-13.95 to -7.16	602	1
-7.15 to -0.36	381	33
-0.35 to 6.44	58	255
6.45 to 13.24	2	540
13.25 to 20.04	0	344
20.05 to 26.84	0	15
26.85 to 33.64	0	19
33.65 to 40.44	0	13
40.45 to 47.24	0	0

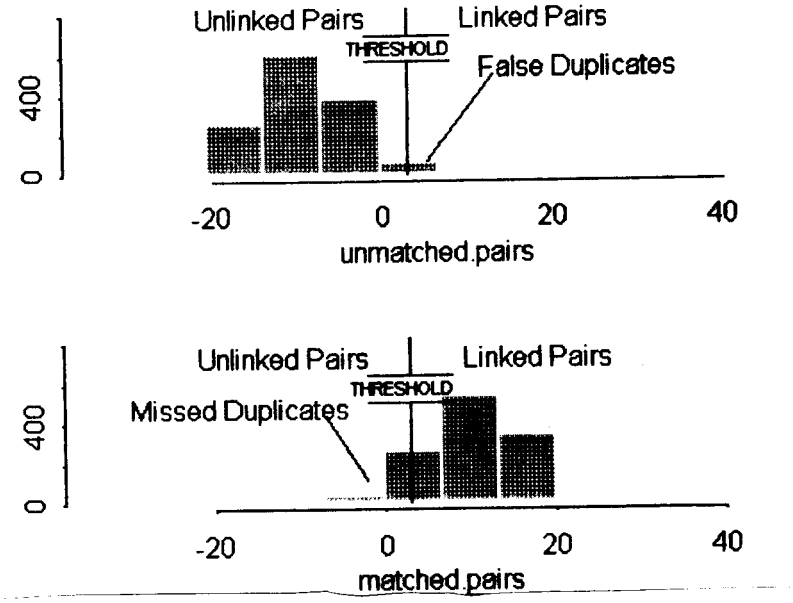
Minimizing False Duplicates

It will be noted that the scores for the unmatched pairs are consistently lower than for the matched pairs, but that occasionally, the scores for the unmatched pairs will be higher than some of the scores for the matched pairs. Figure 1, below, provides graphs for both histograms. The vertical line for both distributions represents the threshold, or value above which a pair will be linked (i.e., considered as representing the same person). In the top graph, consisting of the unmatched pairs, there are a few weight sums for pairs, which fall above the threshold and thus will be "linked" (i.e., considered as representing the same person, even though the pair was judged by the worker to represent different people). These are the "false duplicates."

Minimizing Missed Duplicates

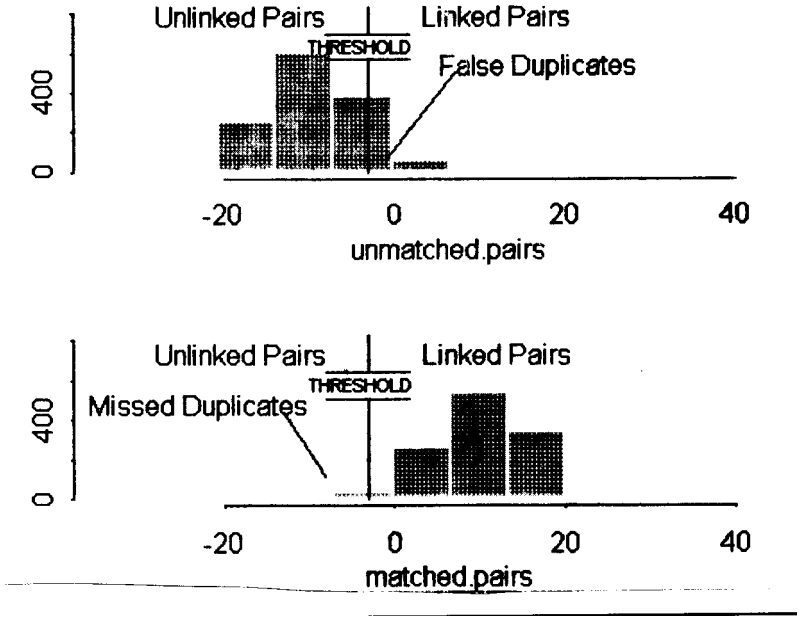
Now, consider the distribution of duplicates or Amatched pairs@ in the lower graph of Figure 1. Notice that with the threshold illustrated, there is a substantial proportion of duplicates that will be unlinked (i.e., considered as representing different people). Rather few of the unmatched pairs will be considered as matched (i.e., will be linked); but substantially more of the duplicates (matched pairs) will fail to be linked. These are the "missed duplicates." The higher proportion of these is due to minimizing the false duplicate error.

Figure 1. – Threshold which Minimizes False Duplicates



It may be that we consider the "missed duplicate" problem as more serious than the "false duplicate" issue. In this case, we can minimize the missed duplicates by moving the threshold to the left, as in Figure 2. Here, the false duplicate rate (the proportion of non-duplicates which are linked) is now larger than that for the missed duplicates.

Figure 2. – Threshold which Minimizes Missed Duplicates



We have now considered two kinds of errors:

- We can fail to identify a genuine match because our "linking" algorithm did not give the sum of the

weights above the threshold (missed duplicates).

- We can "link" a non-duplicate because our algorithm gave the sum of the weights above the threshold (false duplicates).

Table 2 gives both errors for ten alternative threshold values. Note that increasing the threshold value decreases false duplicates but increases the percentage of missed duplicates.

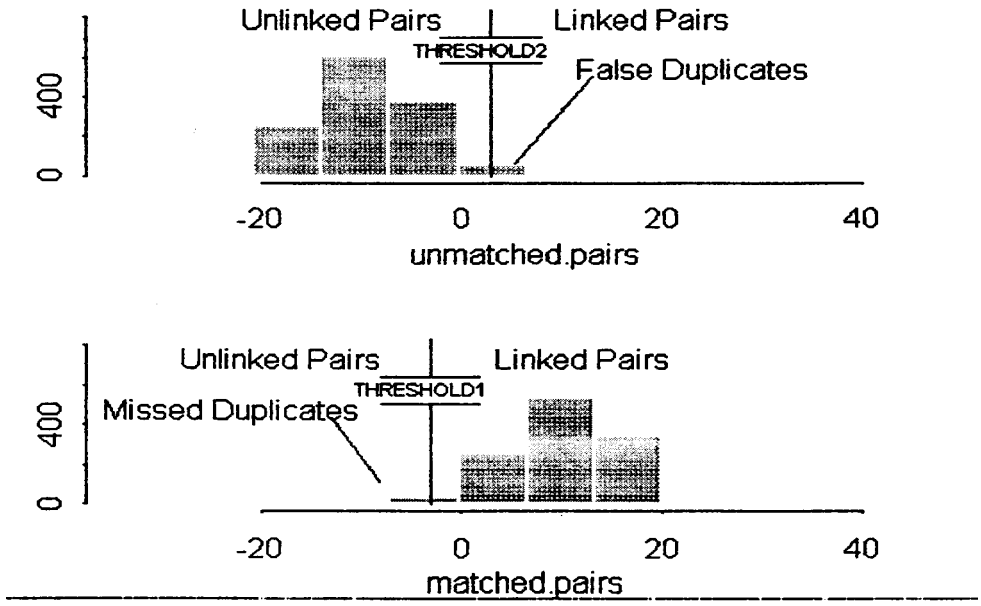
Table 2. -- Duplicates and Missed Duplicates for Alternative Thresholds

Threshold Value for Sum of Log=Odds	% False Duplicates in Nonmatched Sample	% Missed Duplicates in Matched Sample
-8.81	26.85	0.00
-7.03	26.52	0.08
-5.25	26.38	0.08
-3.48	7.66	2.21
-1.70	5.11	2.62
0.06	4.66	2.78
1.84	4.66	2.78
3.61	1.54	16.47
5.39	0.21	23.60
7.16 7.16	0.21	23.93

Now consider Figure 3. If, now, we identify *non-links* as those to the left of the lower threshold (THRESHOLD l) and *links* as those to the right of the upper threshold (THRESHOLD 2), we have a small error rate for both decisions -- but now, we have a new problem.

There is a "gray" area between THRESHOLD 1 and THRESHOLD 2 where there is no rule on how to make a decision. If the pairs in the gray area need to be inspected manually in order to make a decision, this becomes a task of prohibitive magnitude with large files. If one does not need to make a decision with every pair, the use of two thresholds may be the best alternative. NeSmith notes on the final page of her paper that the purpose of linking needs to be considered when setting the threshold. If missed duplicates are the most serious risk, then a lower threshold as in Figure 2 would be preferred. This would make sense for genealogical queries where the failure to find a genuine link could not be compensated for, while a false duplicate would ordinarily be easy to detect on examination. If a large file is to be Acleaned up,@ however, a false duplicate might be more serious, since merging two individuals would then lose information on one of them. A duplicate of an individual would also be a problem, but possibly less serious, and the higher threshold of Figure 1 might be better. If the file were small enough, then cleaning it up might be best with the two thresholds of Figure 3, with manual inspection of the pairs whose linkage scores fell in the gray area.

Figure 3. – Using Two Thresholds to Control Both Kinds of Errors



Summary

The procedure has several main phases:

- Select a file (or set of files) in which to identify duplicates.
- Pick fields for ordering the records to put likely duplicates close together, using a data base management system with "browsing" capacity (blocking).
- Manually identify between 1,000 and 1,500 duplicate pairs.
- Use the duplicate pairs and the preceding formulas to construct weights for all fields, except those used for blocking.
- Select one or two thresholds to use for "linking" pairs of records as estimated duplicates. The position of the threshold or thresholds depends on the desired type and size of the error rates (see Figures 1-3).
- Merge records which have been linked, allowing storage space for possible conflicts. If the entries for a specific field do not match, both entries should be stored, so that a genealogical researcher using the data bank can evaluate them both.
- Use these algorithms to identify duplicates when records are added to the file, and when queries are being made.

This type of project can be repeated with many different geographical areas: the problems and sets of

weights appropriate for use with patronymics will be much different than those associated with the U.S. and Canada. There are many refinements, which need to be investigated, including the use of value-specific techniques, partial agreements, lack of independence between field entries, and the use of other statistical procedures to enhance current techniques.

Note

David White is professor emeritus at Utah State University, Department of Mathematics, Logan, Utah 84322 and has been statistical consultant to the Record Linkage Team, Family History Department, Church of Jesus Christ of Latter-Day Saints.

References

Baldwin, J. A.; Acheson, E. D.; and Graham, W.J. (Eds.) (1987). *Textbook of Medical Record Linkage*, New York: Oxford University Press.

NeSmith, Nancy P. (1992). Record Linkage and Genealogical Files, *Genealogical Journal*, 20, 3-4**,** 113-119.

Newcombe, H. B. (1988). *Handbook of Record Linkage*, New York: Oxford University Press.

This paper is designed as a companion to "Record Linkage and Genealogical Files," by Nancy P. NeSmith (in this volume) and parallels as much as possible the description of record linkage given there with the formulas used to put the theory into practice. The material included here is not in any sense original but derives from the work of H. B. Newcombe (1988) and researchers, such as those included in Baldwin, Acheson, and Graham (1987), who have been working in this area, primarily from the decade of the 1960's and after.

If there are any questions relative to the material of these papers, please contact Ms. NeSmith at the address given in her paper, or Dr. White, with respect to the statistics.

Chapter
11

Matching and Record Linkage

William E. Winkler, Bureau of the Census

Matching has a long history of uses for statistical surveys and administrative data files. Business registers of names, addresses, and other information such as total sales are constructed by combining tax, employment, or other administrative databases (see Chapter 2). Surveys of retail establishments or farms often combine results from an area frame and a list frame. To produce a combined estimator, units must be identified from the area frame sample that are also found on the list frame (see Chapter 11). To estimate the size of a population via capture–recapture techniques, units common to two or more independent listings must be accurately determined (Sekar and Deming 1949; Scheuren 1983; Winkler 1989b). Samples must be drawn appropriately to estimate overlap (Deming and Gleser 1959).

Rather than develop a special survey to collect data for policy decisions, it is sometimes more appropriate to match data from administrative data sources. An economist, for instance, might wish to link a list of companies and the energy resources they consume with a comparable list of companies and the types, quantities, and dollar amounts of the goods they produce. There are potential advantages to using administrative data in analyses. Administrative data sources may contain greater amounts of data and that data may be more accurate due to improvements over time. In addition, virtually all cost of data collection is borne by the administrative programs, and respondent burden associated with a special survey is eliminated. Brackstone (1987) discusses these and other advantages of administrative sources as a substitute for surveys. Methods of adjusting analyses for matching error in merged databases are also available (Neter et al. 1965, Scheuren and Winkler 1993).

[1]The author appreciates many useful comments by Brenda G. Cox, the section editor, and an anonymous reviewer. The opinions expressed are those of the author and not necessarily those of the U.S. Bureau of the Census.

Business Survey Methods, Edited by Cox, Binder, Chinnappa, Christianson, Colledge, Kott.
ISBN 0-471-59852-6

This chapter addresses exact matching in contrast to statistical matching (Federal Committee on Statistical Methodology 1980). An *exact match* is a linkage of data for the same unit (e.g., business) from different files; linkages for units that are not the same occur only because of error. Exact matching uses identifiers such as name, address, or tax unit number. *Statistical matching*, on the other hand, attempts to link files that have few units in common. Linkages are based on similar characteristics rather than unique identifying information, and strong assumptions about joint relationships are made. Linked records need not correspond to the same unit.

Increasingly, computers are used for exact matching to reduce or eliminate manual review and to make results more easily reproducible. Computer matching has the advantages of allowing central supervision of processing, better quality control, speed, consistency, and reproducibility of results. When two records have sufficient information for making decisions about whether the records represent the same unit, humans can exhibit considerable ingenuity by accounting for unusual typographical errors, abbreviations, and missing data. For all but the most difficult situations, however, modern computerized record linkage can achieve results at least as good as a highly trained clerk. When two records have missing or contradictory name or address information, then the records can only be correctly matched if additional information is obtained. For those cases when additional information cannot be adjoined to files automatically, humans are often superior to computer matching algorithms because they can deal with a variety of inconsistent situations.

In the past, most record linkage has been done manually or via elementary but ad hoc computerized rules. This chapter focuses on computer matching techniques that are based on formal mathematical models subject to testing via statistical and other accepted methods. A description is provided of how aspects of name, address, and other file information affect development of automated procedures. The algorithms I describe are based on optimal decision rules that Fellegi and Sunter (1969) developed for methods first introduced by Newcombe et al. (1959). Multidisciplinary in scope, these automated record linkage approaches involve (1) string comparator metrics, search strategies, and name and address parsing/standardization from computer science; (2) discriminatory decision rules, error rate estimation, and iterative fitting procedures from statistics; and (3) linear programming methods from operations research. This chapter contains many examples because its purpose is to provide background for practitioners. While proper theory plays an important role in modern record linkage, my intent is to summarize theoretical ideas rather than rigorously develop them. The seminal paper by Fellegi and Sunter (1969) is still the best reference on theory and related computational methods.

20.1 TERMINOLOGY AND DEFINITION OF ERRORS

Much theoretical work and associated software development for matching and record linkage have been done by different groups working in relative isola-

tion, resulting in varied terminology across groups. In this chapter I use terminology consistent with Newcombe (Newcombe et al. 1959; Newcombe 1988) and Fellegi and Sunter (1969).

In the product $\mathbf{A} \times \mathbf{B}$ of files A and B, a *match* is an $a_i\ b_j$ pair that represents the same business entity and a *nonmatch* is a pair that represents two different entities. Within a single list, a *duplicate* is a record that represents the same business entity as another record in the same list. Rather than consider all pairs in $\mathbf{A} \times \mathbf{B}$, attention is sometimes restricted to those pairs that agree on certain identifiers or *blocking criteria*. Blocking criteria are also called *pockets* or *sort keys*. For instance, instead of making detailed comparisons of all 90 billion pairs from two lists of 300,000 records representing all businesses in a particular state, it may be reasonable to limit comparisons to the set of 30 million pairs that agree on U.S. Postal ZIP code. Errors of omission can result from use of such blocking criteria; *missed matches* are those false nonmatches that do not agree on a set of blocking criteria.

A *record linkage decision rule* is a rule that designates a pair either as a link, a possible link, or a nonlink. *Possible links* are those pairs for which the identifying data are insufficient to decide if the pair is a match. Typically, clerks review possible links and determine their match status. In a list of farms, name information alone is not sufficient for deciding whether "John K Smith, Jr, Rural Route 1" and "John Smith, Rural Route 1" represent the same operation. The second "John Smith" may be the same person as "John K Smith, Jr" or may be his father or grandfather. Mistakes can and do occur in matching. *False matches* are those nonmatches that are erroneously designated as links by a decision rule. *False nonmatches* are either (1) matches designated as nonlinks by the decision rule as it is applied to a set of pairs or (2) missed matches that are not in the set of pairs to which the decision rule is applied. Generally, *link/nonlink* refers to designations under decision rules and *match/nonmatch* refers to true status.

Matching variables are common identifiers (such as name, address, annual receipts, or tax code number) that are used to identify matches. Where possible, a business name such as "John K Smith Company" is parsed or separated into components such as first name "John," initial "K," surname "Smith," and business key word "Company." The parse allows better comparison of names and hence improves matching accuracy. Similarly, an address such as "1423 East Main Road" might be parsed into location number "1423," direction "East," street name "Main," and street type "Road." Matching variables do not necessarily uniquely identify matches. For instance, in constructing a frame of a city's retail establishments, name information such as "Hamburger Heaven" may not allow proper linkage if "Hamburger Heaven" has several locations. The addition of address information can sometimes help, but not if many businesses have different addresses on different lists. In such a situation there is insufficient information to separate new units from existing units that have different mailing addresses associated with them. The *matching weight* or *score* is a number assigned to a pair that simplifies assignment of link and nonlink status via decision rules. A procedure, or matching variable,

has more *distinguishing power* if it is better able to delineate matches and nonmatches than another.

20.2 IMPROVED COMPUTER-ASSISTED MATCHING METHODS

Historically, record linkage has been assigned to clerks who reviewed the lists, obtained additional information when matching information was missing or contradictory, and made linkage decisions following established rules. Typically these lists were sorted alphabetically by name or address characteristics to simplify the review process. If a name contained an unusual typographical variation, the clerks might not find its matches. For large files, matches could be separated by several pages of printouts, so that some matches might be missed. Even after extensive training, the clerks' matching decisions were not always consistent. All work required extensive review. Each major update required training the clerical staff again.

On the other hand, development of computer matching software can require person-years of time from proficient computer scientists. Existing software may not work optimally on files having characteristics significantly different from those for which they were developed. The advantages of automated methods far outweigh these disadvantages. In situations for which good identifiers are available, computer algorithms are fast, accurate, and yield reproducible results. Search strategies can be far faster and more effective than those applied by clerks. As an example, the best computer algorithms allow searches using spelling variations of key identifiers. Computer algorithms can also account for the relative distinguishing power of combinations of matching fields as input files vary. In particular, the algorithms can deal with the relative frequency that combinations of identifiers occur.

As an adjunct to computer operations, clerical review is still needed to deal with pairs having significant amounts of missing information, typographical errors, or contradictory information. Even then, using the computer to bring pairs together and having computer-assisted methods of review at terminals is more efficient than manual review of printouts.

By contrasting the creation of mailing lists for the U.S. Census of Agriculture in 1987 and 1992, the following example dramatically illustrates how enhanced computer matching techniques can reduce costs and improve quality. Absolute numbers are comparable because 1987 proportions were multiplied by the 1992 base of six million. To produce the address list, duplicates were identified in six million records taken from 12 different sources. Before 1982, listings were reviewed manually and an unknown proportion of duplicates remained in files.

In 1987, the development of effective name parsing and adequate address parsing software allowed creation of an ad hoc computer algorithm for automatically designating links and creating subsets for efficient clerical review. Within pairs of records agreeing on ZIP code, the ad hoc computer algorithm

used surname-based information, the first character of the first name, and numeric address information to designate 6.6 percent (396,000) of the records as duplicates and 28.9 percent as possible duplicates to be clerically reviewed. About 14,000 person-hours (as many as 75 clerks for 3 months) were used in this clerical review, and an additional 450,000 duplicates (7.5 percent) were identified. Many duplicates were not located, compromising subsequent estimates based on the list.

In 1992, Fellegi–Sunter algorithms were developed that used effective computer algorithms for dealing with typographical errors. The computer software designated 12.8 percent of the file as duplicates and another 19.7 percent as needing clerical review. About 6500 person-hours were used and an additional 486,000 duplicates (8.1%) were identified. Even without further clerical review, the 1992 computer procedures identified almost as many duplicates as the 1987 combination of computer and clerical procedures. The cost of software development was $110,000 in 1992. The rates of duplicates identified by computer plus clerical procedures were 14.1 percent in 1987 and 20.9 percent in 1992. The 1992 computer procedures lasted 22 days; in contrast, the 1987 computer plus clerical procedure needed 3 months.

20.3 STANDARDIZATION AND PARSING

Appropriate parsing of name and address components is crucial for computerized record linkage. Without it, many true matches would erroneously be designated as nonlinks because identifying information could not be adequately compared. For specific types of business lists, the drastic effect of parsing failure has been quantified (Winkler 1985b, 1986). DeGuire (1988) presents concepts needed for parsing and standardizing addresses; name parsing requires similar concepts.

20.3.1 Standardization of Names and Addresses

The basic ideas of *standardization* are to (1) replace the many spelling variations of commonly occurring words with standard spellings such as fixed abbreviations or spellings and (2) use key words found during standardization as hints for parsing subroutines. In standardizing names, words of little distinguishing power such as "Corporation" or "Limited" are replaced with consistent abbreviations such as "CORP" and "LTD," respectively. First name spelling variations such as "Rob" and "Bobbie" might be replaced with a consistent, assumed, original spelling such as "Robert" or an identifying root word such as "Robt" because "Bobbie" could refer to a woman with "Roberta" as her legal first name. The purpose of name standardization is to allow name-parsing software to work better, by presenting names consistently and by separating out name components that have little value in matching. When business-associated words such as "Company" or "Incorporated" are en-

countered, flags are set that force entrance into different name-parsing routines than would be used otherwise.

Standardization of addresses operates like standardization of names. Words such as "Road" or "Rural Route" are typically replaced by appropriate abbreviations. For instance, when a variant of "Rural Route" is encountered, a flag is set that forces parsing into routines different from routines associated with house-number/street-name addresses. When reference lists containing city, state or province, and postal codes are available from the national postal service or another source, then city names in address lists can be placed in a standard form that is consistent with the reference list.

20.3.2 Parsing of Names and Addresses

Parsing divides a free-form name field into a common set of components that can be compared. Parsing algorithms often use hints based on words that have been standardized. For instance, words such as "CORP" or "CO" might cause parsing algorithms to enter different subroutines than words such as "MRS" or "DR." In the examples of Table 20.1, "Smith" is the name component with the most identifying information. PRE refers to a prefix, POST1 and POST2 refer to postfixes, and BUS1 and BUS2 refer to commonly occurring words associated with businesses. While exact, character-by-character comparison of the standardized but unparsed names would yield no matches, use of the subcomponent last name "Smith" might help designate some pairs as links. Parsing algorithms are available that deal with either last-name-first types of names such as "John Smith" or last-name-last types such as "Smith, John." None are available that can accurately parse both types of names in a single file.

Humans can easily compare many types of addresses because they can associate corresponding subcomponents in free-form addresses. To be most effective, matching software requires address subcomponents to be in identified locations. As the examples in Table 20.2 show, parsing software divides a free-form address field into a set of corresponding components in specific locations on the data record.

20.3.3 Examples of Names

The main difficulty with business names is that even when they are properly parsed, the identifying information may be indeterminate. In each example of Table 20.3, the pairs refer to the same business entity in a survey frame. Alternatively, in Table 20.4, each pair refers to different business entities that have similar names. Because the name information in Tables 20.3 and 20.4 may be insufficient for accurately determining match status, address information or other identifying characteristics may have to be obtained via clerical review. If the additional address information is indeterminate, then at least one establishment in each pair may have to be contacted.

Table 20.1 Examples of Name Parsing

	Parsed							
Standardized	PRE	FIRST	MIDDLE	LAST	POST1	POST2	BUS1	BUS2
DR John J Smith MD	DR	John	J	Smith	MD			
Smith DRY FRM				Smith			DRY	FRM
Smith & Son ENTP				Smith		Son	ENTP	

Table 20.2 Examples of Address Parsing

	Parsed									
Standardized	Pre2	Hsnm	Stnm	RR	Box	Post1	Post2	Unit1	Unit2	Bldg
16 W Main ST APT 16	W	16	Main			ST		16		
RR 2 BX 215				2	215					
Fuller BLDG SUITE 405									405	Fuller
14588 HWY 16 W		14588	HWY				W			

Table 20.3 Names Referring to the Same Business Entities

Name	Reason
John J Smith ABC Fuel Oil	One list has owner name while the other list has business entity name.
John J Smith, Inc. J J Smith Enterprises	Either name may be used by the business.
Four Star Fuel, Exxon Distrib. Four Star Fuel	Independent fuel oil dealer is associated with major oil company.
Peter Knox Dairy Farm Peter J Knox	One list has establishment name while the other has owner name.

Table 20.4 Names Referring to Different Businesses

Name	Reason
John J Smith Smith Fuel	Similar initials or names but different companies
ABC Fuel ABC Plumbing	Same as previous
North Star Fuel, Exxon Distrib. Exxon	Independent affiliate and company with which affiliated

20.4 MATCHING DECISION RULES

For many projects, automated matching decision rules are developed using ad hoc, intuitive approaches. For instance, the decision rule might be as follows:

- If the pair agrees on a specific three characteristics or agrees on four or more within a set of five characteristics, designate the pair as a link.
- If the pair agrees on a specific two characteristics, designate the pair as a possible link.
- Otherwise, designate the pair as a nonlink.

Ad hoc rules are easily developed and may yield good results. The disadvantage is that ad hoc rules may not be applicable to pairs that are different from those used in defining the rule. Users seldom evaluate ad hoc rules with respect to false match and false nonmatch rates.

In the 1950s, Newcombe et al. (1959) introduced concepts of record linkage that were formalized in the mathematical model of Fellegi and Sunter (1969). Computer scientists independently rediscovered the model (Cooper and Maron 1978, Van Rijsbergen et al. 1981, Yu et al. 1982) and showed that the model's decision rules work best among a variety of rules based on competing mathe-

matical models. Fellegi and Sunter's ideas are a landmark in record linkage theory because they introduce many ways of computing key parameters needed for the matching process. Their paper provides (1) methods of estimating outcome probabilities that do not rely on intuition or past experience, (2) estimates of error rates that do not require manual intervention, and (3) automatic threshold choice based on estimated error rates. In my view the best way to build record linkage strategies is to start with formal mathematical techniques based on the Fellegi–Sunter model and then make ad hoc adjustments only as necessary. The adjustments may be likened to the manner in which early regression procedures were informally modified to deal with outliers and collinearity.

20.4.1 Crucial Likelihood Ratio

The record linkage process attempts to classify pairs in a product space $\mathbf{A} \times \mathbf{B}$ from two files A and B into M, the set of true matches, and U, the set of true nonmatches. Fellegi and Sunter (1969) considered ratios of probabilities of the form

$$R = \frac{P(\gamma \in \Gamma \mid M)}{P(\gamma \in \Gamma \mid U)}, \tag{20.1}$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement on the largest name component, street name, and street number. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific values of name components such as "Smith," "Zabrinsky," "AAA," and "Capitol" occur.

20.4.2 Theoretical Decision Rule

The decision rule is equivalent to the one originally given by Fellegi and Sunter [1969, equation (19)]. In the following, r represents an arbitrary pair, $\gamma \in \Gamma$ is the agreement pattern associated with r, and R is the ratio corresponding to r that is given by equation (20.1). The decision rule d provides three designated statuses for pairs and is given by:

$$d(r) = \begin{cases} \text{link} & \text{if } R > UPPER \\ \text{possible link} & \text{if } LOWER \leq R \leq UPPER \\ \text{nonlink} & \text{if } R < LOWER. \end{cases} \tag{20.2}$$

The cutoff thresholds *UPPER* and *LOWER* are determined by *a priori* error bounds on false matches and false nonmatches. Rule 20.2 agrees with intuition. If $\gamma \in \Gamma$ consists primarily of agreements, then it is intuitive that $\gamma \in \Gamma$ would be more likely to occur among matches than nonmatches and ratio (20.1)

would be large. On the other hand, if $\gamma \in \Gamma$ consists primarily of disagreements, then ratio (20.1) would be small.

Fellegi and Sunter (1969) showed that rule (20.2) is optimal in that for any pair of fixed upper bounds on the rates of false matches and false nonmatches, the clerical review region is minimized over all decision rules on the same comparison space Γ. The theory holds on any subset such as pairs agreeing on a postal code, street name, or part of a name field. The ratio R or any monotonically increasing transformation of it (such as given by a logarithm) is referred to as a *matching weight* or *total agreement weight*. In actual applications, the optimality of rule (20.2) is heavily dependent on the accuracy of the estimated probabilities in equation (20.1). The probabilities in equation (20.1) are called *matching parameters*.

20.4.3 Basic Parameter Estimation Under the Independence Assumption

Fellegi and Sunter (1969) were the first to observe that certain parameters needed for rule (20.2) could be obtained directly from observed data if certain simplifying assumptions were made. For each $\gamma \in \Gamma$, they considered

$$P(\gamma) = P(\gamma \mid M)P(M) + P(\gamma \mid U)P(U) \tag{20.3}$$

and noted that the proportion of pairs with $\gamma \in \Gamma$ could be computed directly from available data. If $\gamma \in \Gamma$ consists of a simple agree/disagree pattern associated with three variables satisfying the conditional independence assumption that there exist vector constants (marginal probabilities) $m \equiv (m_1, m_2, \cdots, m_K)$ and $u \equiv (u_1, u_2, \cdots, u_K)$ such that, for all $\gamma \in \Gamma$,

$$P(\gamma \mid M) = \prod_{i=1}^{K} m_i^{\gamma^i}(1 - m_i)^{(1-\gamma^i)} \quad \text{and} \quad P(\gamma \mid U) = \prod_{i=1}^{K} u_i^{\gamma^i}(1 - u_i)^{(1-\gamma^i)}, \tag{20.4}$$

then Fellegi and Sunter provide the seven solutions for the seven distinct equations associated with equation (20.3).

If $\gamma \in \Gamma$ represents more than three variables, then it is possible to apply general equation-solving techniques such as the method of moments (e.g., Hogg and Craig 1978, pp. 205–206). Because the method of moments has shown numerical instability in some record linkage applications (Jaro 1989) and with general mixture distributions (Titterington et al. 1988, p. 71), maximum-likelihood-based methods such as the Expectation-Maximization (EM) algorithm (Dempster et al. 1977, Wu 1983, Meng and Rubin 1993) may be preferred.

The EM algorithm has been used in a variety of record linkage situations. In each, it converged rapidly to unique limiting solutions over different starting

points (Thibaudeau 1989; Winkler 1989a, 1992). The major difficulty with the parameter-estimation techniques (EM or an alternative such as method of moments) is that they may yield solutions that partition the set of pairs into two sets that differ substantially from the desired sets of true matches and true nonmatches. In contrast to other methods, the EM algorithm converges slowly and is stable numerically (Meng and Rubin 1993).

20.4.4 Adjustment for Relative Frequency

Newcombe et al. (1959) introduced methods for using the specific values or relative frequencies of occurrence of fields such as surname. The intuitive idea is that if surnames such as "Vijayan" occur less often than surnames such as "Smith," then "Vijayan" has more distinguishing power. A variant of Newcombe's ideas was later mathematically formalized by Fellegi and Sunter (1969; see also Winkler 1988, 1989c for extensions). Copas and Hilton (1990) introduced a new theoretical approach that, in special cases, has aspects of the Newcombe's approach; it has not yet applied in a record linkage system. While the value-specific approach can be used for any matching field, strong assumptions must be made about independence between agreement on specific value states of one field versus agreement on other fields.

The concepts of Fellegi and Sunter (1969, pp. 1192–1194) describe the problem well. To simplify the ideas, files A and B are assumed to contain no duplicates. The true frequencies of specific values of a string such as first name in files A and B, respectively, are given by

$$f_1, f_2, \cdots, f_m; \sum_{j=1}^{m} f_j = N_A$$

and

$$g_1, g_2, \cdots, g_m; \sum_{j=1}^{m} g_j = N_B.$$

If the mth string, say "Smith," occurs f_m times in File A and g_m times in File B, then pairs agree on "Smith" $f_m g_m$ times in $\mathbf{A} \times \mathbf{B}$. The corresponding true frequencies in M are given by

$$h_1, h_2, \cdots, h_m; \sum_{j=1}^{m} h_j = N_M.$$

Note that $h_j \leq \min(f_j, g_j)$, where $j = 1, 2, \cdots, m$. For some implementations, h_j is assumed to equal the minimum, and $P(\text{agree } j\text{th value of string} \mid M) = h_j/N_M$ and $P(\text{agree } j\text{th value of string} \mid U) = (f_j g_j - h_j)/(N_A \cdot N_B - N_M)$. In practice, observed values rather than true values must be used. The variants of how the h_j frequencies are computed involve differences in how typographical errors are modeled, what simplifying assumptions are made, and how fre-

quency weights are scaled to simple agree/disagree probabilities (Newcombe 1988; Fellegi and Sunter 1969; Winkler 1988, 1989c). As originally shown by Fellegi and Sunter (1969), the scaling can be thought of as a means of adjusting for typographical error. The scaling is

$$P(\text{agree on string} \mid M) = \sum_{j=1}^{m} P(\text{agree on } j\text{th value of string} \mid M),$$

where the probability on the left is estimated via the EM algorithm or another method. With minor restrictions, the ideas of Winkler (1989c) include those of Fellegi and Sunter (1969), Newcombe (1988, pp. 88–89), and Rogot et al. (1986) as special cases.

In some situations, the frequency tables are created "on-the-fly" using the files actually being matched (Winkler 1989c); in others, the frequency tables are created *a priori* using large reference files. The advantage of on-the-fly tables is that they can use different relative frequencies in different geographic regions; for instance, Hispanic surnames in Los Angeles, Houston, or Miami and French surnames in Montreal. The disadvantage of on-the-fly tables is that they must be based on files that cover a large percentage of the target population. If the data files contain samples from a population, then the frequency weights should reflect the appropriate population frequencies. For instance, if two small lists of companies in a city are used and "George Jones, Inc" occurs once on each list, then a pair should not be designated as a link using name information only. Corroborating information such as address should also be used because the name "George Jones, Inc" may not uniquely identify the establishment.

20.4.5 Jaro String Comparator Metrics for Typographical Error

Jaro (1989) introduced methods for dealing with typographical error such as "Smith" versus "Smoth." Jaro's procedure consists of two steps. First, a string comparator returns a value based on counting insertions, deletions, transpositions, and string length. Second, the value is used to adjust a total agreement weight downward toward the total disagreement weight. Jaro's string comparator was extended by making agreement in the first few characters of the string more important than agreement on the last few (Winkler 1990b). As Table 20.5 illustrates, the original Jaro comparator and the Winkler-enhanced comparator yield a more refined scale for describing the effects of typographical error than do standard computer science methods such as the Damerau–Levenstein metric (Winkler 1985a, 1990b).

Jaro's original weight-adjustment strategy was based on a single adjustment function developed via ad hoc methods. Using calibration files having true matching status, Jaro's strategy has been extended by applying crude statistical curve fitting techniques to define several adjustment functions. Different curves were developed for first names, last names, street names, and house numbers.

Table 20.5 Comparison of String Comparators Rescaled Between 0 and 1

Strings		Winkler	Jaro	Damerau–Levenstein
billy	billy	1.000	1.000	1.000
billy	bill	0.967	0.933	0.800
billy	blily	0.947	0.933	0.600
massie	massey	0.944	0.889	0.600
yvette	yevett	0.911	0.889	0.600
billy	bolly	0.893	0.867	0.600
dwayne	duane	0.858	0.822	0.400
dixon	dickson	0.853	0.791	0.200
billy	susan	0.000	0.000	0.000

When used in actual matching contexts, the new set of curves and enhanced string comparator improve matching efficacy when compared to the original Jaro methods (Winkler 1990b). With general business lists, the same set of curves could be used or new curves could be developed. In a large experiment using files for which true matching status was known, Belin (1993) examined effects of different parameter-estimation methods, uses of value-specific weights, applications of different blocking criteria, and adjustments using different string comparators. Belin demonstrated that the original Jaro string comparator and the Winkler extensions were the two best ways of improving matching efficacy in files for which identifying fields had significant percentages of minor typographical errors.

20.4.6 General Parameter Estimation

Two difficulties arise in applying the EM procedures of Section 20.4.3. The first is that the independence assumption is often false (Smith and Newcombe 1975, Winkler 1989b). The second is that, due to model misspecification, EM or other fitting procedures may not naturally partition the set of pairs into the desired sets of matches M and nonmatches U.

To account for dependencies between the agreements of different matching fields, an extension of an EM-type algorithm due to Haberman (1975, see also Winkler 1989a) can be applied. Because many more parameters are associated with general interaction models than with independence models, only a fraction of all interactions may be fit. For instance, if there are 10 matching variables, the degrees of freedom are only sufficient to fit all three-way interactions (e.g., Bishop et al. 1975, Haberman 1979); with fewer matching variables, it may be necessary to fit various subsets of the three-way interactions.

To address the natural partitioning problem, $\mathbf{A} \times \mathbf{B}$ is partitioned into three sets of pairs C_1, C_2, and C_3 using an equation analogous to (20.3). The EM procedures are then divided into three-class or two-class procedures. When appropriate, two of the three classes are combined into a set that represents

either M or U. The remaining class represents the complement. When both name and address information is used for matching, the two-class EM tends to divide a set of pairs into those agreeing on address information and those disagreeing. If address information associated with many pairs is indeterminate (e.g., Rural Route 1 or Highway 65 West), the three-class EM can yield a proper partition because it tends to divide the set of pairs into (1) matches at the same address, (2) nonmatches at the same address, and (3) nonmatches at different addresses.

The general EM algorithm is far slower than the independent EM algorithm because the M step is no longer in closed form. Convergence is speeded up by using variants of the Expectation-Conditional Maximization (ECM) and Multicycle ECM (MCECM) Algorithm (Meng and Rubin 1993, Winkler 1989a). The difficulty with general EM procedures is that different starting points often yield different limiting solutions. However, if the starting point is relatively close to the solution given by the independent EM algorithm, then the limiting solution is generally unique (Winkler 1992). The independent EM algorithm often provides starting points that are suitable for the general EM algorithm.

Figures 20.1–20.8 illustrate that the automatic EM-based parameter-estimation procedures can yield dramatic improvements. Because there were no available business files for which true matching status was known, files of individuals having name, address, and demographic characteristics such as age, race, and sex were used. Each figure contains a plot of the estimated cumulative distribution curve via equation (20.2) versus the truth that is given by the 45-degree line. Figures 20.1–20.4 for matches and Figures 20.5–20.8 for nonmatches successively display fits according to (1) iterative refinement (e.g., Newcombe 1988, pp. 65–66), (2) three-class, independent EM, (3) three-class, selected interaction EM, and (4) three-class, three-way interaction EM with

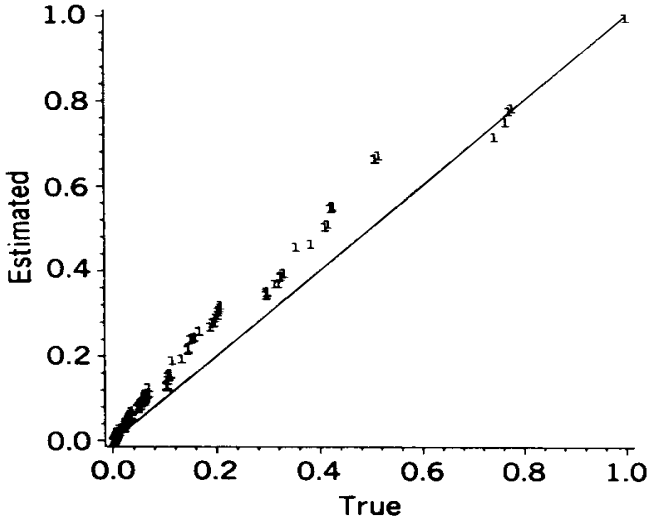


Figure 20.1 Estimates vs. truth, cumulative distribution of matches—two-class, iterative.

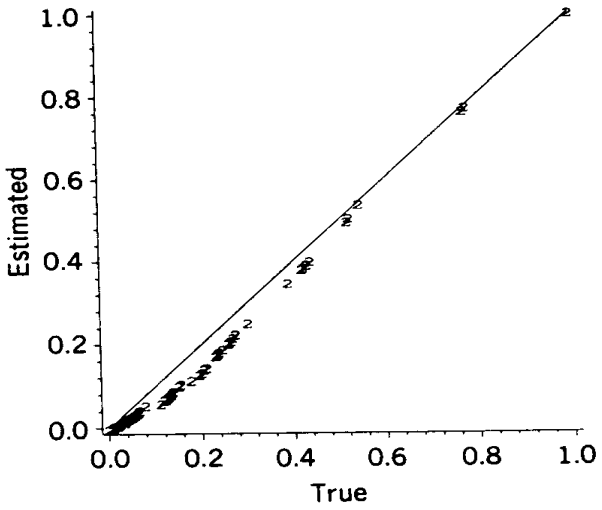


Figure 20.2 Estimates vs. truth, cumulative distribution of matches—three-class, independent EM.

convex constraints. *Iterative refinement* involves the successive manual review of sets of pairs and the reestimation of probabilities given a match under the independence assumption. Iterative refinement is chosen as a reference point (Figures 20.1 and 20.4) because it yields reasonably good matching decision rules (e.g., Newcombe 1988; Winkler 1990b). The algorithm for fitting selected interactions is due to Armstrong (1992). The EM algorithm with convex constraints that predispose a solution to the proper region of the parameter

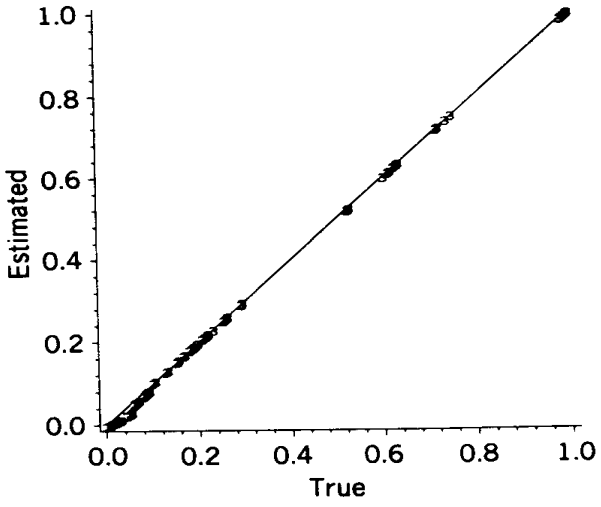


Figure 20.3 Estimates vs. truth, cumulative distribution of matches—three-class, selected interaction EM.

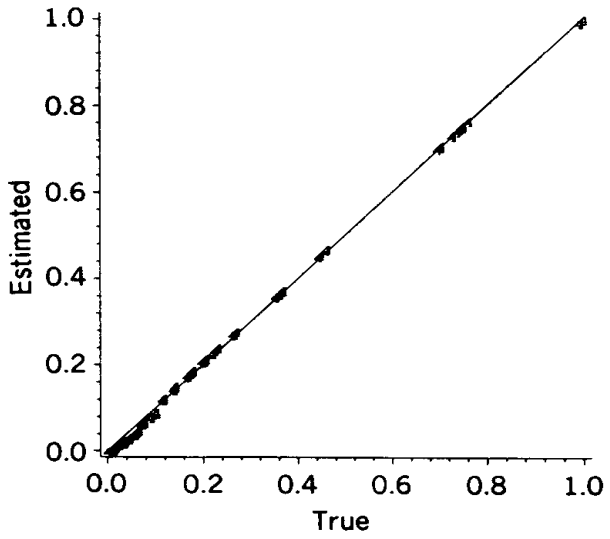


Figure 20.4 Estimates vs. truth, cumulative distribution of matches—three-class, three-way interaction EM, convex.

space is due to Winkler (1989a; also 1992, 1993b). All three-way interactions are used in the last model.

The basic reason that iterative refinement and three-class independent EM perform poorly is that independence does not hold. Three-class independent EM yields results that are closer to the truth because it divides the set of pairs that agree on address into those agreeing on name and demographic information and those that disagree. Thus, nonmatches such as husband–wife and

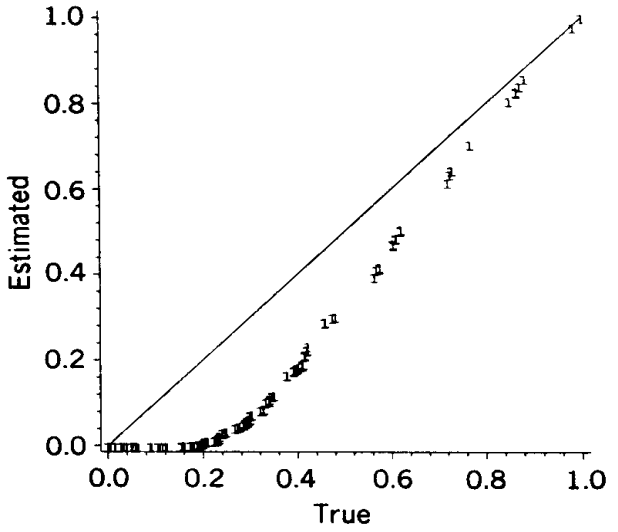


Figure 20.5 Estimates vs. truth, cumulative distribution of nonmatches—two-class, iterative.

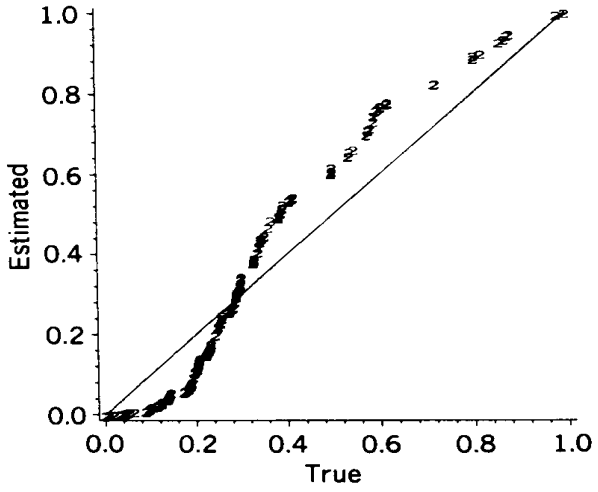


Figure 20.6 Estimates vs. truth, cumulative distribution of nonmatches—three-class, independent EM.

brother–sister pairs are separated from matches such as husband–husband and wife–wife. As shown by Thibaudeau (1993) with these data, departures from independence are moderate among matches whereas departures from independence among nonmatches (such as the husband–wife and brother–sister pairs at the same address) are quite dramatic.

The selected interaction EM does well (Figures 20.3 and 20.7) because true

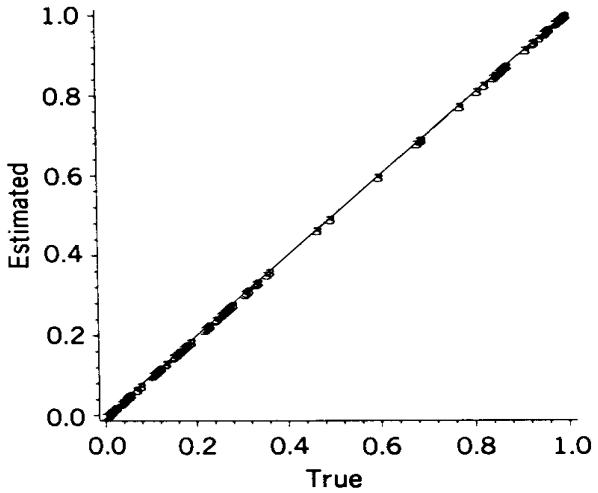


Figure 20.7 Estimates vs. truth, cumulative distribution of nonmatches—three-class, selected interaction EM.

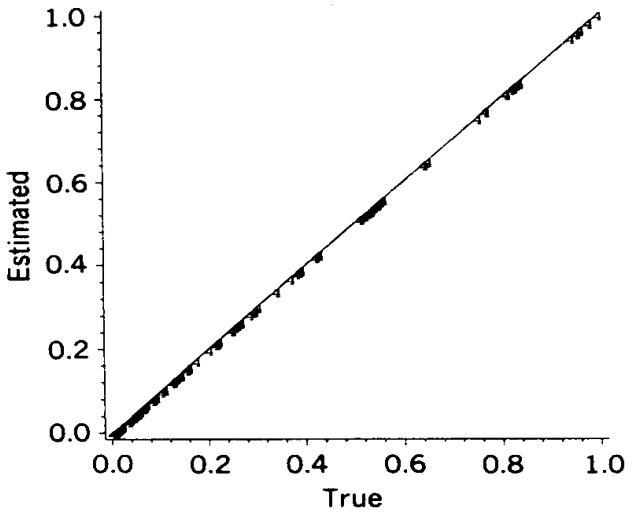


Figure 20.8 Estimates vs. truth, cumulative distribution of nonmatches—three-class, three-way interaction EM, convex.

matching status is used to determine the interactions that must be included. It is unreasonable to expect that true matching status will be available for many matching situations or that the exact set of interactions that were developed for one application will be suitable for use in another. Furthermore, loglinear modeling in latent-class situations is more difficult than for basic loglinear situations where such modeling is known to be difficult (e.g., Bishop et al. 1975). To alleviate the situation, it may be suitable to take a model having all three-way interactions and use convex constraints that bound some probabilities. The bounds would be based on similar matching situations. The all three-way interaction model without convex constraints does not provide accurate fits (Winkler 1992). If the convex constraints are chosen properly, then the three-way interaction EM with convex constraints provides fits (Figures 20.4 and 20.8) that are nearly as good as those obtained with the selected interaction EM (Winkler 1993b).

20.5 EVALUATING THE QUALITY OF LISTS

The quality of lists is primarily determined by how useful the available variables are for matching. For large files, the first concern is how effective common identifiers (blocking criteria) are at reducing the set of pairs to a manageable size. The effectiveness of blocking criteria is also determined by the estimated number of missed matches. Applying a greater number of matching variables generally improves matching efficacy. Name information generally provides more distinguishing power than receipts, sales, or address informa-

tion. Parameter estimates must be as good as possible. Improving parameter estimates can reduce clerical review regions by as much as 90 percent.

20.5.1 Quality of Blocking Criteria

While use of blocking criteria facilitates the matching process by reducing the number of pairs to be considered, it can increase the number of false nonmatches because some pairs disagree on the blocking criteria. The following describes an investigation of how well different sets of blocking criteria yield sets of pairs containing all matches (Winkler 1984, 1985b). The sets of pairs were constructed from 11 U.S. Energy Information Administration (EIA) lists and 47 state and industry lists containing 176,000 records. Within the set of pairs from the original set of files, name and address information allowed 110,000 matches to be identified. From the remaining 66,000 records, there were 3050 matches having similar names and addresses and 8510 matches having either a different name or a different address. The remaining 11,560 matches (18 percent of the 66,000 records) were identified via intensive manual review and were used in analyzing various blocking criteria.

In the subsequent analysis, only the 3050 matches having similar names and addresses were considered. In the blocking criteria displayed in Table 20.6, NAME represents an unparsed name field. Only the first few characters from different fields were used. These criteria were the best subset of several hundred criteria that were considered for blocking a list of sellers of petroleum products (Winkler 1984). Table 20.7 illustrates that for certain sets of lists it is quite

Table 20.6 Blocking Criteria

1. 3 digits ZIP code, 4 characters NAME
2. 5 digits ZIP code, 6 characters STREET
3. 10 digits TELEPHONE
4. 3 digits ZIP code, 4 characters of largest substring in NAME
5. 10 characters NAME

Table 20.7 Incremental Decrease in False Nonmatches—Each Set Consists of Pairs in the Union of Sets Agreeing on Blocking Criteria

Group of Criteria	Rate of False Nonmatches	Matches/ Incremental Increase	Nonmatches/ Incremental Increase
1	45.5	1460/NA	727/NA
1–2	15.1	2495/1035	1109/289
1–3	3.7	2908/413	1233/124
1–4	1.3	2991/83	1494/261
1–5	0.7	3007/16	5857/4363

difficult to produce groups of blocking criteria that give a set of pairs that include all matches. With the union of pairs based on the best two sets of criteria, 15.1 percent of the matches were dropped from further consideration; with three, 3.7 percent. The last (fifth) criterion was not useful because it enlarged the set of pairs with only 16 additional matches while adding 4363 nonmatches.

20.5.2 Estimation of False Nonmatches Not Agreeing on Multiple Blocking Criteria

If estimates of the numbers of missed matches are needed, then lists can be sampled directly. Even with very large sample sizes, the estimated standard deviation of the error rate estimate often exceeds the estimate (Deming and Gleser 1959). If samples are not used, then following the suggestion of Scheuren (1983), capture-recapture techniques as in Sekar and Deming (1949; see also Bishop et al. 1975, Chapter 6) can be applied to the set of pairs captured by the first four sets of blocking criteria of Section 20.5.1 (Winkler 1987). The best-fitting loglinear model yields the 95 percent confidence interval (27,160). The interval, which represents between 1 and 5 percent of true matches, contains the 50 matches that were known to be missed by the blocking criteria and found via intense clerical review.

20.5.3 Number of Matching Variables

As the number of matching variables increases, the ability to distinguish matches usually increases. For instance, with name information alone, it may only be feasible to create subsets of pairs that are held for clerical review. With name and address information, a substantial number of the matches can be correctly distinguished. With name, address, and financial information (such as receipts or income), it may be possible to distinguish most matches automatically.

Exceptions occur if some matching variables have extreme typographical variations and/or are correlated with other matching variables. For instance, consider the following. Two name fields are available for each record of the pairs. The first is a general business name that typically agrees among matches. The second name field in one record corresponds to the owner of a particular business license (e.g., in some states, all fuel storage facilities must be licensed) and in the other record the name field corresponds to the accounting entity that keeps financial records. While the owner of a particular business license will sometimes correspond to the financial person (owner of a gasoline service station), the two names will often disagree among true matches. When both name fields are used in software that assumes that agreements are uncorrelated, contradictory information can cause loss of distinguishing power. Expedient solutions are to drop the contradictory information in the second name field or to alleviate the problem via custom software modifications.

Table 20.8 Examples of Agricultural Names

John A Smith
John A and Mary B Smith
John A Smith and Robert Jones
Smith Dairy Farm

20.5.4 Relative Distinguishing Power of Matching Variables

Without a unique identifier such as a verified employer identification number (EIN), the name field typically has more distinguishing power than other fields such as address. The ability of name information to distinguish pairs can vary dramatically from one set of pairs to another. For instance, in one situation properly parsed name information, when combined with other information, may produce good automatic decision rules; in other situations it may not.

As an example of the first situation, consider the 1992 U.S. Census of Agriculture in which name parsing software was optimized to try to find surnames (or suitable surrogates) and first names. Because the overwhelming majority of farming operations have names of the form given in Table 20.8, the resultant parsed names will likely all have "Smith" as a surname that will yield good distinguishing power when combined with address information. The exception can occur when two names containing "Smith" have the same address. A similar situation occurs with the 1992 match of the Standard Statistical Establishment List (SSEL) of U.S. businesses with a list of small nonemployers from an Internal Revenue Service (IRS) 1040C file of records for which EIN was unavailable.

General business lists can signify the second situation of the poor decision rule because of the ways in which the name field can be represented. For instance, the same business entity may appear in the following forms given in Table 20.9. Even if name parsing software can properly represent name components, it may be difficult to use the components to distinguish matches. If the name information and clerical-review status were retained, then clerical review could be reduced during future updates. Each business could be represented by a unique record that has pointers to significant name variations of matches and nonmatches along with match status. If a potential update record

Table 20.9 Examples of Business Names That Are Difficult to Compare

John A Smith and Son Manufacturing Company, Incorporated
John Smith Co
John Smith Manufacturing
J A S Inc.
John Smith and Son

is initially designated as a possible link because of a name variation, then the associated name variations could be searched to decide whether a record with a name similar to the potential update record had previously been clerically reviewed. If it had, then the prior follow-up results could be used to determine whether the new record is a match.

20.5.5 Good Matching Variables But Unsuitable Parameter Estimates

Even when name and other matching variables can be properly parsed and have agreeing components, automatic parameter estimation software may not yield good parameter estimates because the lists have little overlap or because model assumptions in the parameter-estimation software are incorrect. In either situation, matching parameters are usually estimated via an iterative procedure involving manual review. Generally, matching personnel start with an initial set of parameters. The personnel review a moderately large sample of matching results and estimate new parameters via ad hoc means. The review-reestimation process is repeated until matching personnel are satisfied that parameters and matching results will not improve much.

The most straightforward means of parameter reestimation is the iterative refinement procedure of Statistics Canada (e.g., Newcombe 1988, pp. 65–66; Statistics Canada 1983; Jaro 1992). After each review and clerical resolution of match results, marginal probabilities given a match are reestimated and matching (under the independence assumption) is repeated. Marginal probabilities given a nonmatch are held as constant because they are approximated by probabilities of random agreement over the entire set of pairs. If the proportion of nonmatches within the set of pairs is very high, then the random-agreement approximation is valid because decision rules using the random agreement probabilities are virtually the same as decision rules using true marginal probabilities given a nonmatch.

For the 1992 U.S. Census of Agriculture, initial estimates obtained via the independent EM algorithm were replaced by refined estimates that accounted for lack of independence. The refined estimates were determined by reviewing a large sample of pairs, creating adjusted probability estimates, and repeating the process. For instance, if two records simultaneously agreed on surname and first name, their matching weight was adjusted upward from the independent weight.

20.6 ESTIMATION OF ERROR RATES AND ADJUSTMENT FOR MATCHING ERROR

Fellegi and Sunter (1969) introduced methods for automatically estimating error rates when the conditional independence assumption (20.4) is valid. Their methods do not involve sampling and can be extended to more general situations. This section provides different methods for estimating error rates within

a set of pairs than those given in Section 20.4.6. Estimation of false nonmatches due to pairs missed because of disagreement on blocking criteria is covered in Section 20.5. This section also describes new work that investigates how statistical analyses can be adjusted for matching error.

20.6.1 Sampling and Clerical Review

Estimates of the number of false matches and nonmatches can be obtained by reviewing a sample of pairs designated as links and nonlinks. Sample size can be minimized by concentrating the sample in weight ranges in which error is likely to take place. Using a weighting strategy that yields good distinguishing power with rule (20.2), most error among computer-designated links and nonlinks occurs among weights that are close to the thresholds *UPPER* and *LOWER*. Within the set of possible links that are clerically designated as links and nonlinks, simple random samples can be used. While the amount of manual review needed for confirming or correcting the link–nonlink designations can require substantial resources, reasonable estimates within the fixed set of pairs can be obtained. An alternative to sampling is to develop effective statistical models that allow automatic estimation of error rates. At present, such methods are the subject of much research and should show improvements in the future.

20.6.2 Rubin–Belin Estimation

Rubin and Belin (1991) developed a method of estimating matching error rates when the curves (ratio R versus frequency) for matches and nonmatches are somewhat separated and the failure of the independence assumption is not too severe. Their method is applicable to weighting curves R obtained via a one-to-one matching rule (Jaro 1989) and to which a number of ad hoc adjustments are made (Winkler 1990b). The one-to-one matching rule can dramatically improve matching performance because it can eliminate nonmatches such as husband–wife or brother–sister pairs that agree on address information. Without one-to-one matching, such pairs receive sufficiently high weights to be designated as possible links.

To model the shape of the curves of matches and nonmatches, Rubin and Belin require true matching status for a representative set of pairs. For a variety of basic settings, the procedure yields reasonably accurate estimates of error rates and is not highly dependent on *a priori* curve shape parameters (Rubin and Belin 1991; Scheuren and Winkler 1993; Winkler 1992). The SEM algorithm of Meng and Rubin (1991) is used to get 95 percent confidence intervals for the estimates.

While the Rubin–Belin procedures were developed using files of individuals (for which true match status was known), I expect that the procedures are also applicable for files of businesses. When one-to-one matching is used, the Rubin and Belin method can give better error rate estimates than a modified version

of the Winkler method given in Section 20.4.6 (e.g., Winkler 1992). If one-to-one matching is not used, then the Winkler method can yield accurate parameter estimates whereas the Rubin–Belin method cannot be applied because the curves associated with matches and nonmatches are not sufficiently separated.

20.6.3 Scheuren–Winkler Adjustment of Statistical Analyses

Linking information that resides in separate files can be useful for analysis and policy decisions. For instance, an economist might wish to evaluate energy policy by matching a file with fuel and commodity information for businesses against a file with the values and types of goods produced by the businesses. If the wrong businesses are matched, then analyses based on the linked files can yield erroneous conclusions. Scheuren and Winkler (1993) introduced a method of adjusting statistical analyses for matching error. If the probability distributions for matches and nonmatches are accurately estimated, then the adjustment method is valid in simple cases where one variable is taken from each file. Accurate estimates can sometimes be obtained via the method of Rubin and Belin (1991). Empirical applications have been performed for ordinary linear regression models (Winkler and Scheuren 1991) and for simple loglinear models (Winkler 1991). Extensions to situations of more than one variable from each file are under investigation.

20.7 COMPUTING RESOURCES AND AUTOMATION

Many large record linkage projects require new software or substantial modification of existing software. The chief difficulty with these projects is developing the highly skilled programmers required for the task. Few programmers have the aptitude or are allowed the years needed to acquire proficiency in advanced algorithm development and the multi-language, multi-machine approaches needed to modify and enhance existing software. For example, a government agency may use software that another agency spent several years developing in PL/I because PL/I is the only language their programmers know. Possibly more appropriate software written in C may not be used because the same programmers do not know how to compile and run C programs. The same PL/I programmers may not have the skills that allow them to make major modifications in PL/I software that they did not write or to port new algorithms in other languages to PL/I.

A secondary concern is lack of appropriate, general-purpose software. In many situations for which name, address, and other comparable information are available, existing matching software will work well if names and addresses can be parsed correctly. Directly comparable information might consist

of receipts for comparable time periods. Nondirectly comparable information might consist of receipts in one source and sales in another. To use such data, custom software modifications have to be added to software. The advantage of some existing software is that, without modification, they often parse a substantial percentage of the records in files.

20.7.1 Need for General Name-Parsing Software and What Is Available

At present, the only general-purpose business-name-parsing software that has been used by an assortment of agencies is the NSKGEN software from Statistics Canada. The software is written in a combination of PL/I and IBM Assembly language. NSKGEN software is primarily intended to create search keys that bring appropriate pairs of records together. Because it does a good job of parsing and standardizing names, it has been used for record linkage (Winkler 1986, 1987). I recently wrote general business-name-parsing software that was used in a match of the U.S. SSEL list of business establishments with the U.S. IRS 1040C list that contains many small establishments (Winkler 1993a). The software achieves better than a 99 percent parsing rate with an error rate of less than 0.2 percent with these lists. It has not yet been tested on a variety of general lists. The code is ANSI-standard C and, upon recompilation, runs on a number of computers. While name parsing software is written and used by commercial firms, the associated source code is generally considered proprietary.

20.7.2 Need for General Address-Parsing Software and What Is Available

Statistics Canada has the ASKGEN package (again written in PL/I and IBM Assembly language) which does a good job of parsing addresses (Winkler 1986, 1987). ASKGEN has recently been superseded by Postal Address Analysis System (PAAS) software. PAAS has not yet been used at a variety of agencies but, with limitations, has been used in creating an address register for the 1991 Canadian Census. The limitations were that most of the source address lists required special preprocessing to put individual addresses in a form more suitable for input to PAAS software (Swain et al. 1992). In addition to working on English-type addresses, the ASKGEN and PAAS software works on French-type addresses such as "16 Rue de la Place."

At the U.S. Bureau of the Census, address-parsing software has been written in ANSI-standard C and, upon recompilation, currently runs on an assortment of computers. The software has been incorporated in all major Census Bureau geocoding systems, has been used for the 1992 U.S. Census of Agriculture, and was used in several projects involving the 1992 U.S. SSEL. As with name-parsing software, source code for commercial address-parsing software is generally considered proprietary.

20.7.3 Matching Software

At present, I am unaware of any general software packages that have been specifically developed for matching lists of businesses. While the ASKGEN and NSKGEN standardization packages were used with the Canadian Business Register in 1984, associated matching was based on search keys generated through compression and standardization of corporate names. One-to-many matches were reviewed by clerks who selected the best match with the help of interactive computer software. At the U.S. Bureau of the Census, I have been involved with the development of software for large projects in which the Fellegi–Sunter model was initially used and a number of ad hoc modifications were made to deal with name-parsing failure, address-parsing failure, sparse and missing data, and data situations unique to the files being matched. In every case, the ad hoc modifications improved matching performance substantially over performance that would have been available from the software alone. The recent projects were the 1992 U.S. Census of Agriculture, the 1993 match of the SSEL file of U.S. businesses with the IRS 1040C list of nonemployers, and the 1993 matching of successive years' SSEL files and the unduplication of individual years' files. The latter two projects used files from 1992. A set of software for agricultural lists and several packages for files of individuals are described below.

The U.S. Department of Agriculture (1980) has a system for matching lists of agricultural businesses, which was written in FORTRAN for IBM mainframes in 1979 and has never been updated. Name-parsing software is available as part of the system. The software applies Fellegi–Sunter matching to the subsets of pairs corresponding to individuals. The remaining records that are identified as corresponding to partnerships and corporations are matched clerically when an exact character-by-character match fails. If the pairs of businesses generally have names that allow them to be represented in forms similar to the ways that files of individuals have their names represented, then matching software (or modifications of it) designed for files of individuals can be used.

While the ASKGEN and NSKGEN packages from Statistics Canada have been given out to individuals for use on IBM mainframes, associated documentation does not cover installation or details of the algorithms. To a lesser extent, the lack of detailed documentation is also true for the USDA system. The software packages require systems analysts and matching experts for installation and use.

General matching software has only been used on files of individuals due to the difficulties of name and address standardization and consistency in business files. Available systems are Statistics Canada's GRLS system (Hill 1991, Nuyens 1993), the system for the U.S. Census (Winkler 1990a), Jaro's commercial system (Jaro 1992), and University of California's CAMLIS system. None of the systems provides name- or address-parsing software. Only the Winkler system is free and, upon recompilation, runs on a large collection of

computers. Source code is available with the GRLS system and the Winkler system. The GRLS system has the best documentation.

20.8 CONCLUDING REMARKS

This chapter provides background on how the Fellegi–Sunter model of record linkage is used in developing automated matching software for business lists. The presentation shows how a variety of existing techniques have been created to alleviate specific problems due to name- and/or address-parsing failure or inappropriateness of assumptions used in simplifying computation associated with the Fellegi–Sunter model. Much research is needed to improve record linkage of business lists. The challenges facing agencies and individuals are great because substantial time and resources are needed for (1) creating and enhancing general name and address parsing/software; (2) performing, circulating, and publishing methodological studies; and (3) generalizing and adding features to existing matching software that improve its effectiveness when applied to business lists.

REFERENCES

Armstrong, J. A. (1992), "Error Rate Estimation for Record Linkage: Some Recent Developments," in *Proceedings of the Workshop on Statistical Issues in Public Policy Analysis*, Carleton University.

Belin, T. R. (1993), "Evaluation of Sources of Variation in Record Linkage Through a Factorial Experiment," *Survey Methodology*, **19**, pp. 13–29.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.

Brackstone, G. J. (1987), "Issues in the Use of Administrative Records for Administrative Purposes," *Survey Methodology*, **13**, pp. 29–43.

Cooper, W. S., and M. E. Maron (1978), "Foundations of Probabilistic and Utility-Theoretic Indexing," *Journal of the Association for Computing Machinery*, **25**, pp. 67–80.

Copas, J. R., and F. J. Hilton (1990), "Record Linkage: Statistical Models for Matching Computer Records," *Journal of the Royal Statistical Society*, Series A, **153**, pp. 287–320.

DeGuire, Y. (1988), "Postal Address Analysis," *Survey Methodology*, **14**, pp. 317–325.

Deming, W. E., and G. J. Gleser (1959), "On the Problem of Matching Lists by Samples," *Journal of the American Statistical Association*, **54**, pp. 403–415.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, **39**, pp. 1–38.

Federal Committee on Statistical Methodology (1980), *Report on Exact and Statistical*

Matching Techniques, Statistical Policy Working Paper 5, Washington, DC: U.S. Office of Management and Budget.

Fellegi, I. P., and A. B. Sunter (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, **64,** pp. 1183–1210.

Haberman, S. J. (1975), "Iterative Scaling for Log-Linear Model for Frequency Tables Derived by Indirect Observation," *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 45–50.

Haberman, S. (1979), *Analysis of Qualitative Data*, New York: Academic Press.

Hill, T. (1991), "GRLS-V2, Release of 22 May 1991," unpublished report, Ottawa: Statistics Canada.

Hogg, R. V., and A. T. Craig (1978), *Introduction to Mathematical Statistics*, 4th ed., New York: Wiley.

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, **89,** pp, 414–420.

Jaro, M. A. (1992), "AUTOMATCH Record Linkage System," unpublished, Silver Spring, MD

Meng, X., and D. B. Rubin (1991), "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, **86,** pp. 899–909.

Meng, X., and D. B. Rubin (1993), "Maximum Likelihood via the ECM Algorithm: A General Framework," *Biometrika*, **80,** pp. 267–278.

Neter, J., E. S. Maynes, and R. Ramanathan (1965), "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, **60,** pp. 1005–1027.

Newcombe, H. B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press.

Newcombe, H. B., J. M. Kennedy, S. J. Axford, and A. P. James (1959), "Automatic Linkage of Vital Records," *Science*, **130,** pp. 954–959.

Nuyens, C. (1993), "Generalized Record Linkage at Statistics Canada," *Proceedings of the International Conference on Establishment Surveys*, Alexandria, VA: American Statistical Association, pp. 926–930.

Rogot, E., P. Sorlie, and N. Johnson (1986), "Probabilistic Methods of Matching Census Samples to the National Death Index," *Journal of Chronic Disease*, **39,** pp. 719–734.

Rubin, D. B., and T. R. Belin (1991), "Recent Developments in Calibrating Error Rates for Computer Matching," *Proceedings of the Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, pp. 657–668.

Scheuren, F. (1983), "Design and Estimation for Large Federal Surveys Using Administrative Records," *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 377–381.

Scheuren, F., and W. E. Winkler (1993), "Regression Analysis of Data Files That Are Computer Matched," *Survey Methodology*, **19,** pp. 39–58.

Sekar, C. C., and W. E. Deming (1949), "On a Method of Estimating Birth and Death

Rates and the Extent of Registration," *Journal of the American Statistical Association*, **44**, pp. 101–115.

Smith, M. E., and H. B. Newcombe (1975), "Methods of Computer Linkage of Hospital Admission-Separation Records into Cumulative Health Histories," *Methods of Information in Medicine*, **14**, pp. 118–125.

Statistics Canada (1983), "Generalized Iterative Record Linkage System," unpublished report, Ottawa: Systems Development Division.

Swain, L., J. D. Drew, B. LaFrance, and K. Lance (1992), "The Creation of a Residential Address Register for Coverage Improvement in the 1991 Canadian Census," *Survey Methodology*, **18**, pp. 127–141.

Thibaudeau, Y. (1989), "Fitting Log-Linear Models When Some Dichotomous Variables Are Unobservable," *Proceedings of the Section on Statistical Computing, American Statistical Association*, pp. 283–288.

Thibaudeau, Y. (1993), "The Discrimination Power of Dependency Structures in Record Linkage," *Survey Methodology*, **19**, pp. 31–38.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1988), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

U.S. Department of Agriculture (1980), "Record Linkage System Documentation," unpublished report, Washington, DC: National Agricultural Statistics Service.

Van Rijsbergen, C. J., D. J. Harper, and M. F. Porter (1981), "The Selection of Good Search Terms," *Information Processing and Management*, **17**, pp. 77–91.

Winkler, W. E. (1984), "Exact Matching Using Elementary Techniques," technical report, Washington DC: U.S. Energy Information Administration.

Winkler, W. E. (1985a), "Preprocessing of Lists and String Comparison," in W. Alvey and B. Kilss (eds.), *Record Linkage Techniques—1985*, U.S. Internal Revenue Service, Publication 1299 (2-86), pp. 181–187.

Winkler, W. E. (1985b), "Exact Matching Lists of Businesses: Blocking, Subfield Identification, Information Theory," in W. Alvey and B. Kilss (eds.), *Record Linkage Techniques—1985*, U.S. Internal Revenue Service, Publication 1299 (2-86), pp. 227–241.

Winkler, W. E. (1986), "Record Linkage of Business Lists," technical report, Washington, DC: U.S. Energy Information Administration.

Winkler, W. E. (1987), "An Application of the Fellegi-Sunter Model of Record Linkage to Business Lists," technical report, Washington, DC: U.S. Energy Information Administration.

Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 667–671.

Winkler, W. E. (1989a), "Near Automatic Weight Computation in the Fellegi–Sunter Model of Record Linkage," *Proceedings of the Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, pp. 145–155.

Winkler, W. E. (1989b), "Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage," *Survey Methodology*, **15**, pp. 101–117.

Winkler, W. E. (1989c), "Frequency-Based Matching in the Fellegi–Sunter Model of

Record Linkage,'' *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 778–783.

Winkler, W. E. (1990a), ''Documentation of Record-Linkage Software,'' unpublished report, Washington, DC: U.S. Bureau of the Census.

Winkler, W. E. (1990b), ''String Comparator Metrics and Enhanced Decision Rules in the Fellegi–Sunter Model of Record Linkage,'' *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 354–359.

Winkler, W. E. (1991), ''Error Model for Analysis of Computer Linked Files,'' *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 472–477.

Winkler, W. E. (1992), ''Comparative Analysis of Record Linkage Decision Rules,'' *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 829–834.

Winkler, W. E. (1993a), ''Business Name Parsing and Standardization Software,'' unpublished report, Washington, DC: U.S. Bureau of the Census.

Winkler, W. E. (1993b), ''Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage,'' *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 274–279.

Winkler, W. E., and F. Scheuren (1991), ''How Matching Error Affects Regression Analysis: Exploratory and Confirmatory Results,'' technical report, Washington, DC: U.S. Bureau of the Census.

Wu, C. F. J. (1983), ''On the Convergence Properties of the EM Algorithm,'' *Annals of Statistics*, **11**, pp. 95–103.

Yu, C. T., K. Lam, and G. Salton (1982), ''Term Weighting in Information Retrieval Using the Term Precision Model,'' *Journal of the Association for Computing Machinery*, **29**, pp. 152–170.

Chapter 11

Linking Health Records: Human Rights Concerns*

Fritz Scheuren, Ernst and Young, LLP

1. Purpose

The purpose of this paper is to provide an introduction or "starter set" for reflecting on human rights issues that arise when bringing together or linking the health records of individuals. In particular, the paper will discuss the potential role of record linkages in the proposed new United States health information system; specifically, how linkage applications may affect both the rights of individuals to privacy and their rights of access to health care services.

Four potential types of record linkages will be covered (see Figure 1 below). The primary concern will be with linkages of health records, such as the computerized enrollment and encounter records proposed to be created under the Health Security Act or other health care reform legislation [1]. As the columns of Figure 1 indicate, linkages for both statistical and administrative purposes will be considered. As the rows of Figure 1 imply, there will be a discussion of record linkage within the health system, e.g., records of individuals may be linked to records of providers or insurers. The paper will also consider linkages of health care records with records from other systems, such as vital records or social security, income tax, and welfare program records.

In all, the paper is organized into eight sections: the present introduction and statement of purpose (Section 1); a background section on what is meant by record linkage -- both in general and with respect to health record systems (Section 2); then there are four short sections, each devoted to a cell in Figure 1 (Sections 3 to 6); and, finally, a brief overall summary with recommendations (Section 7). The main questions to be addressed throughout are the extent to which linkages should be permitted, for what purposes, and under what conditions. An Afterword has been included (as Section 8) to afford room for a more personal comment.

Figure 1. -- Potential Types of Health Record Linkages
(Cell entries reference paper section where topic covered)

Linkages	**Purposes**	
	Administrative	**Statistical**
Within health record system	Section 3	Section 4
With other Record systems	Section 6	Section 5

*Reprinted with permission. See Note at end of paper.

2. Background

This section is a review of automated record linkage techniques, the nature of record linkage errors, and some overall system concerns in a world where multiple opportunities exist to carry out record linkages.

2.1 Types of Record Linkages

It seems fairly safe to speculate that once human beings began to keep records there were efforts to link them together. Until well into this century, though, such work was done manually and often only with great difficulty and expense; however, there now exist four broad types of automated record linkage (see Figure 2) -- each of which will be described below by means of an example.

Figure 2. -- Examples of Linkage Types and System Structures

Type of Record Linkage	Record System Structure	
	Intended for Linkage	**Incidental to Linkage**
Deterministic	Social Security and Medicare systems	National Death Index (NDI)
Probabilistic	1990 Census Post Enumeration Survey	NDI Links to the Current Population Survey

In the United States, the first National experience with automated record linkage systems was the assignment, beginning in 1935, of social security numbers (SSN's) to most wage workers. Initially this system was based on a single punch card for each worker; these cards were updated using the SSN as an account identifier and a cumulative total kept of taxable wages received under covered employment. Record linkages at the Social Security Administration were computerized in the 1950's and SSN's are issued now to virtually all Americans.

From its inception, the intended use of the social security number was to carry out record linkage. Efforts, not always successful, were made so that SSN's, when assigned, would be unique and each person would have just one [2]. Further, the wage reporting system was designed so that updates by SSN would be conducted in a manner relatively free of error. Put another way, the social security system was designed or **intended** all along for automated record linkage and a straightforward, so-called **deterministic** linkage rule of **exact matching** on SSN's was to be the basic approach.

Birth and death registration in the U.S. offers a useful contrast to social security. These vital registers, which became complete only in the 1930's, were not intended for automated linkage operations [3]. Identifying items, like names, are on these records, of course, and could be used as matching keys but would not always be unique alone -- common surnames like Smith or Johnson or Williams being notable cases where linkage problems might be particularly severe. Automated linkages to U.S. death records did not begin nationally until the inception in the 1970's of the National Death Index or NDI. The NDI in its original operations relied on **multiple exact matches** as a way to locate potential linkages; [4] hence, as shown in Figure 2, the NDI may serve as an example of a **deterministic** automated linkage approach that was **added on** to a system not initially designed for such a use.

Deterministic match rules are easy to automate but do not adequately reflect the uncertainty that may exist for some potential links. They can also require costly manual intervention when errors occur in the matching keys. More complicated methods were needed that weighed the linkage information, allowing for errors and incompleteness, and minimizing the clerical intervention required to select the best link from all those possible. Such techniques are called **probabilistic.** The main theoretical underpinnings for probabilistic matching methods were firmly established by the late nineteen sixties with the papers of Tepping [5] and, especially, Fellegi and Sunter [6]. Sound practice dates back even earlier, at least to the nineteen fifties and the work of Newcombe and his collaborators [7].

The Fellegi-Sunter approach is basically a direct extension of the classical theory of hypothesis testing to the problem of record linkage. A mathematical model is developed for recognizing records in two files which represent identical units (said to be matched). As part of the process there is a comparison between all possible pairs of records (one from each file) and a decision made as to whether or not the members of the comparison-pair represent the same units, or whether there is insufficient evidence to justify either of these decisions. The three outcomes from this process can be referred to as a "link," " nonlink," or "potential link."

In point of fact, Fellegi and Sunter contributed the underlying theory to the methods already being used by Newcombe and showed how to develop and optimally employ probability weights to the results of the comparisons made. They also dealt with the implications of restricting the comparison pairs to be looked at, that is of "blocking" the files, something that generally has had to be done when linking files that are at all large.

Many of the major public health research advances made in recent decades have benefitted at least in part from probabilistic linkage techniques. Included are such well known epidemiological findings as the effects of smoking, risks from radiation exposure, asbestos and many other carcinogens arising in the workplace, through diet or other exposures -- increasingly in populations with genetic predispositions [8]. These benefits have to be considered when exploring record linkage impacts on privacy and other rights. We will return to this point at the end of this paper where trade-offs are explicitly considered.

Most of these automated linkages, like Newcombe's studies of radiation exposure at Chalk River (and elsewhere), were not envisioned when the records were originally created. Some probabilistic linkage systems were intended, however -- notably for "post enumeration" surveys (PES's), carried out to evaluate U.S. decennial census coverage. For example, the PES for 1990 was particularly well designed for carrying out probabilistic linkages [9]. Another good example of a continuing probabilistic linkage that has been a real success for statistical purposes is the bringing together of the NDI and Current Population Survey [10]. This linkage, though, was not planned into the design of either of the data sets being employed.

2.2 Nature of Linkage Errors and Identifying Information

All linkage operations are subject to two main types of errors: matching records together that belong to different entities (false matches) and failing to put records together that belong to the same entity (false non-matches). These errors can have different human rights implications, depending on what the linkages are used for (see Figure 3).

Figure 3. -- Linkage Error Implications on Human Rights

Types of Linkage Error	Linkages Used for --	

False Matches	Data about that individual	Information about a class of individuals
False Nonmatches	Potentially very serious	May be less serious

If the linkage is to assemble data about an individual so an administrative or diagnostic determination can be made about ***that*** individual, then the consequences of any error could be grave indeed. Potentially, a different (lower) standard of accuracy could be tolerated, provided a suitable adjustment is made when analyzing the results of linkage operations whose purpose is to obtain information about a group [11]. More will be said about these issues in later sections, particularly how this distinction affords an opportunity to both preserve individual privacy rights -- through group matches, say -- but still attain societal information needs.

If an efficient (low cost, essentially error free) health care linkage system is a goal, then consideration needs to be given to the establishment of a health identification "number." In ideal circumstances, personal identifying information on a medical record should satisfy the following requirements [12].

- The identifying information should be **permanent**; that is, it should exist at the birth of a person to whom it relates or be allocated to him/her at birth, and it should remain unchanged throughout life.
- The identifying information should be **universal**; that is, similar information should exist for every member of the population.
- The identifying information should be **reasonable**; that is, the person to whom it relates and others, should have no objection to its disclosure for medical purposes.
- The identifying information should be **economical**; that is, it should not consist of more alphabetic, digits and other characters than necessary.
- The identifying information should be **simple**; that is, it should be capable of being handled easily by a clerk and computers.
- The identifying information should be **available**.
- The identifying information should be **known**; that is, either the person to whom it relates or an informant acting on his/her behalf should be able to provide it on demand.
- The identifying information should be **accurate**; that is, it should not contain errors that could result in its discrepancy on two records relating to the same person.
- The identifying information should be **unique**; that is, each member of the population should be identified differently.

The social security number, incidentally, fails several of these tests. Only now is it beginning to be issued at birth; also it is far from being accurately reported. In practice, too, because of incentives created by the SSN's use in the tax system, the number is not always unique. Some people use ***more than one*** SSN, even in

the same year, and more often over longer periods of time. Multiple uses of the ***same*** SSN by different people have been common , as well.

Concerns about the risks to health records from unauthorized disclosures are greater with an identifier like the SSN which is widely available on many large private data bases, like credit files, and of course many non-health related Federal, state and other government files [13]. In the Office of Technology Assessment's 1993 report [14] on privacy the following recommendation is made with regard to the SSN.

> ***The use of the social security number as a unique patient identifier has far-reaching ramifications for individual health care information privacy that should be carefully considered before it is used for that purpose.***

Elsewhere [15] the stronger recommendation has been made not to use the SSN as a health identifier. Its use could lead to matching errors and might greatly increase the potential for unregulated linkages between health and nonhealth data sets.

2.3 Some Proposed Health Record Linkage Systems

The proposed Health Security Act [16] calls for the establishment of a National Health Board to oversee the creation of an electronic data network. The types of information collected would include: enrollment and disenrollment in health plans; clinical encounters and other items and services from health care providers; administrative and financial transactions and activities of participating states, regional alliances, corporate alliances, health plans, health care providers, employers, and individuals; number and demographic characteristics of eligible individuals residing in each alliance area; payment of benefits; utilization management; quality management; grievances, and fraud or misrepresentation in claims or benefits [17].

The Health Security Act specifies, among other things, the use of uniform paper forms containing standard data elements, definitions, and instructions for completion; requirements for use of uniform health data sets with common definitions to standardize the collection and transmission of data in electronic form; uniform presentation requirements for data in electronic form; and electronic data interchange requirements for the exchange of data among automated health information systems.

A prototype health care record linkage system may be worth considering as well since it spells out an initial schematic of a person-level health or patient record. Data could come from an array of health care settings, linked together using a "linkage processor." This processor would determine the linkage and also assign the unique patient identifier in the actual patient record. Record types would differ by the type of provider from which they are derived. The functions of the record linkage software program are outlined in Figure 4. It is anticipated that the patient identifying information would be housed in a person's primary care unit. The linkage processor stores the patient identifying data and generates the unique identifier. It processes records from other providers and links the record as shown. Some initial data categories and identifying information are outlined in Figure 5 [18].

Figure 4. -- Patient Record Prototype

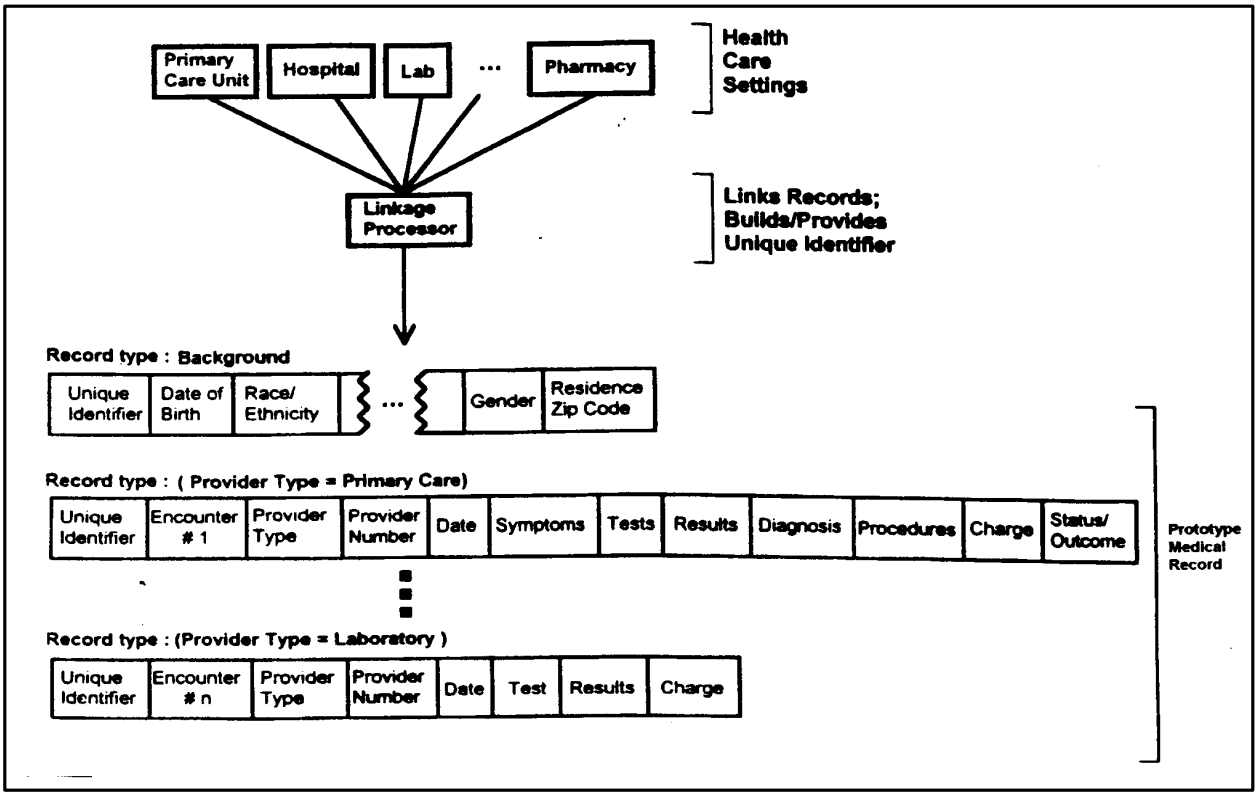
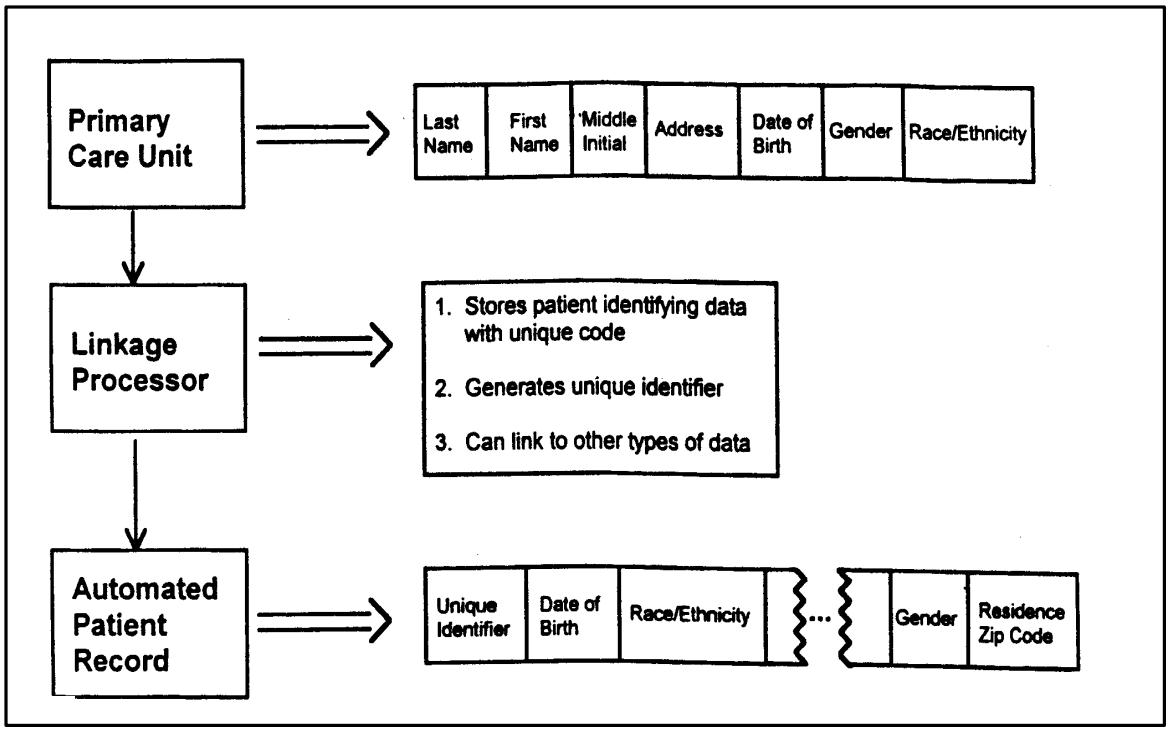


Figure 5. -- Record Linkage Architecture



2.4 Additional System Concern

In all data capture systems, of course, it is important to explicitly build-in the means to address privacy rights, the degree to which confidentiality promises are required (and kept), and the means used to make individual data physically secure. While such concerns are general, record linkage systems have some unique aspects that may bear discussion -- particularly the systems described in Section 2.3 above. Figure 6 summarizes these, emphasizing the additional complexity introduced by the linkage environment and the degree to which linkage systems are or should be "auditable." By "auditable" is meant that, at a minimum, each access to identifiable data is controlled and a log kept of the individuals who obtained the data and of all transactions that occurred (in other words, an **audit trail** is kept so that outside monitoring is possible).

Figure 6. -- Some Overall Record Linkage System Concerns

Linkage Issues	Complexity	Auditability
Privacy	Extremely high, may be beyond our current understanding, without training and experience	May be very difficult to establish, maintain, or use in monitoring access
Confidentiality		
Security		

Tore Dalenius has provided a good review of privacy, confidentiality, and security goals in statistical settings [19]. His work may afford a point of departure for the discussion here. In common speech, the words: privacy, confidentiality and security partially overlap in usage and often have meanings that depend greatly on context. Each can also have an emotional content which makes precise definitions difficult, even contentious. For example, Dalenius quotes Westin (1967) as saying about privacy:

> *Few values so fundamental to society as privacy have been left so undefined in social theory or have been the subject of such vague and confused writing by social scientists.*

A good start on giving meaning to the word "privacy," or "information privacy" (our context here), might be the definition first articulated by Justice Brandeis as the "right to be let alone...the most comprehensive of rights and the right most valued by civilized man" [20]. Attempts to update this definition have been many and will undoubtedly continue. All afford the individual or data subject some, sometimes sole, rights over what matters they want to keep private and what matters they are willing -- or want -- to reveal.

Record linkage settings pose a particular challenge to an individual's ability to exercise his or her privacy rights. The sheer complexity of the setting makes it hard to clarify for the subject what the potential benefit or harm may be to permitting access. Consider the linkage of just two files, say, of "n" and "m" variables respectively. Cognitively for the individual involved the linkage decision may seem like one of no particular moment. The combined file will consist of data already given out earlier -- a single file of "n + m" -- rather than two separate files. But a deeper look -- at relationships, for example, between variables -- shows that a combinatorial explosion of facts about an individual has taken place -- from, say, $2^n + 2^m$ to 2^{n+m}. (Incidentally, to illustrate what this means, assume just that n=m=11; then the combined file has over 1000 times more information about the relationships between variables than the two files separately.)

Ready examples come to mind where individuals present themselves in one way (to get Medicaid or Medicare, say) but in another setting (perhaps a job interview) give a different, even a contradictory set of "facts."

When records from these two encounters are linked, obviously the implications may be many, since these differences would be revealed [21].

Obtaining data at different points in time and for different primary purposes is a difficulty that is peculiar to linkage settings. The privacy decision an individual may wish to make could, therefore, change over time and might depend on the particular data items as well as the purposes for which a release from their privacy rights is being sought. Singer et al., [22] for example, advocate that –

> *Patients sign an informed consent or a notification statement at regular intervals, not simply the first time a patient visits the provider's office.*

They then go on to recommend that the consent or notification statement spell out to whom the information about the patient may be disseminated, for what purposes, and what the patient's rights are with respect to this information. Such an approach, especially as it relates to secondary uses of data and the time period for which the informed consent is valid seems clearly required in a linkage setting where patient specific information may accumulate over time and from many sources (not just one provider). It may be necessary for a regulatory body to develop regulations standardizing the contents of informed consent and disclosure policies. These regulations could definitely state what constitutes an "informed consent" and legitimate non-consented disclosure. Even then, only experience will tell whether true informed consent will be possible for most individuals.

Indeed, without wishing to jump to conclusions, it may be reasonable to conjecture that, for some kinds of data linkage at least and certain individuals, our technological abilities to electronically merge data sets may have outstripped our sense of what a data subject would have to "consent to in an **informed** way" for the systems to be built on an entirely voluntary basis. If this is so, then simply creating the health linkage system envisioned might in and of itself take away the privacy rights of some people.

The problem of complexity in record linkage systems may warrant the attention being given to complexity in general systems[23]. Linear thinking alone may, in any case, be insufficient to address what will happen not only to the individual's ability to manage his or her own data but to the system's integrity overall. What confidentiality promises can be made and kept in such a world? How can one even speculate realistically about the risks to data corruption or unauthorized disclosure? Recent experiences elsewhere [24] do not encourage belief that reasonable ways exist of being clear about even what the threats are.

Among the crucial "fail safes" is to provide an audit trail for every query against a record and any retention of a data subset. Such systems already exist for some real time settings, although not necessarily in a way that would allow a simple scale-up. A crucial step is the maintenance of these systems so they operate properly [25]. While beyond the scope of this paper, it might be noted that the expense of this maintenance step and a mechanism to "monitor the monitoring" needs a lot of attention, too. Arguments in favor of doing record linkages for efficiency reasons have not fully weighed these costs. In Brannigan and Beier [26] still other sound system architecture issues and recommendations are made that would be needed to implement essential confidentiality and security procedures, especially if large scale record linkages are to be employed.

3. Administrative Data Linkages Within the Health System

By an administrative data linkage we mean a linkage of data about an entity done with the intention of taking some direct action regarding that entity. In a health setting the most obvious example would be to assemble (link) data about a patient from multiple sources in order to improve the diagnosis or treatment of that individual. We will start with this case (subsection 3.1) and then go on to discuss administrative health linkages more generally (subsection 3.2).

3.1 Linkages for Direct Patient Care

Figure 7 lays out some of the dimensions in administrative record linkages aimed at improving the health of a patient. The Figure has five rows and two columns. Each row covers a broad area dealing with, respectively, overall issues, technical (and administrative) aspects, legal matters, the perception of the public and of experts. The primary focus of the Figure is to directly address under what conditions linkages should be permitted (Column 1) and areas for future study (Column 2). Since the goal of this paper is to just be a "starter set," only illustrative suggestions have been made in the cells, both here and elsewhere.

Among the general conditions for linkage a signed notification statement seems needed [27]. In this context, a "notification statement" might tell the patient who will have access, for what purposes and with what oversight. Hoffman in a recent paper makes the observation that "too many people may already have insufficiently monitored access to hospital patient records. He seconds Mark Siegler's thesis that "medical confidentiality, as it has been traditionally understood by patients and doctors, no longer exists." Siegler, after a patient

Figure 7. -- Administrative Data Linkages Conducted for the Health of Patients

Broad Areas	**Possible Response**	
	Under what conditions (Column 1)	**For future study (Column 2)**
Overall Recommend-ations	Just notification needed; if for use of patient and patient caregivers only	Concerns about coercive aspects of government "monopoly" in health care
Technical Aspects	Encryption to prevent unauthorized access and reduce risks of reidentification	Concerns about how to monitor operation
Legal Questions	For federal records, subject to the Privacy Act; use seemingly fully permitted now.	Electronic data linkages across governmental jurisdictions deserve more study; also roles of intermediaries (e.g., Health Information Trustee -- HR 4077)
Public Views	Direct evidence lacking but indirect evidence suggests that health uses to aid patients would be seen very positively.	Concerns about public view of risks associated with system need to be better understood.
Expert Opinion	An obvious use, seemingly favored by all.	Need to continue research on uniform reporting issues so as to obtain promised benefits of electronic linkages without an undue burden.

expressed concern about the confidentiality of his hospital record, scanned his medical chart and enumerated "at least 25 and possibly as many as 100 health professionals and administrative personnel...[who] had access to the patient's record," all with legitimate reasons to examine the chart [28].

Secure physical access is essential and any linkage stipulated as done for diagnosis and treatment of a patient should be available only for the use of the patient and his or her caregivers. Concerns exist about the patient data requested for such encounters and whether the demands and burdens on the patient are reasonable. The collection of uniform patient data has clear advantages; the specific data required, though, will need external review, possibly by a regulatory body -- similar to that discussed earlier on consent standards.After all there are privacy rights given up by patients to their caregivers and these should be limited to an essential minimum.

Patient and primary caregiver controlled access might involve encryption techniques or other measures designed to prevent or at least reduce the risks of unauthorized (unmonitored) use. Linkages might be time limited to reduce exposure further. As noted, Brannigen and Beier [29] have made numerous other important suggestions. System administrative issues are extensive and concerns about monitoring operations deserve continued study.

Fair information practices must be adhered to -- as required, say, in the Privacy Act and reinforced by pending legislation [30]. Continuing study of state and local restrictions [31] should be pursued to find good working models and to anticipate areas where weaknesses may arise in the National System, if litigation occurs. The *Privacy Journal* has regularly compiled state and Federal privacy laws and is a useful resource here [32].

Direct evidence of public reaction is lacking on linkages used solely to aid the patient. Such use is presumed to be very positively received. There is a large segment, though, of the public [33] that are concerned about any electronic record linkage system of the scope envisioned, mainly because of their general mistrust of government and other large institutions. These individuals or some of them, at least, might not think the benefits to be derived warrant the risks they perceive for abuse inherent in such a large-scale record linkage effort.

Virtually all "experts" take the position that notification of the use envisioned here is enough. One exception is Goldman [34] which states:

> *Personally-identifiable health records must be in the control of the individual. Personal information should only be disclosed with the knowing, meaningful consent of the individual.*

The distinction between consent and notification may not be as important here as elsewhere. With notification there is always a "quid pro quo" -- give this data about yourself if you want to participate. In this setting patients are often asked to give what amounts to "coerced" consent; therefore, the distinction may be in name only. Logically, however, it seems inconsistent to withhold information about yourself that could be used to aid you. Unquestionably, though, a refusal to comply could mean denial of access to health care services.

3.2 Other Health Administrative Linkages

Many other health linkages are possible besides those directly involved with patient care. These could range --

- From linking treatments received by a patient to the costs of those treatments;
- To associating outcome measures (death or survival, say) to the types of medical procedures employed; and
- Even to linkages whose intent was to detect fraud or malpractice.

Data about a hospital or other health facility might be sought by looking at all the records of the patients that can be linked to that hospital. The number of possibilities, in fact, is very large -- too large to cover in any depth here. Some observations may be helpful, nonetheless, to fix a few of the ideas about what the privacy rights dimensions are:

- First, in administrative linkages such as these, the patient may become just a data point in an endeavor focussed elsewhere [35]. The dehumanizing aspect of this change of focus is inherently unsettling. Provisions like those in Figure 7 seem insufficient when the person looking at the data is not the pri-

mary caregiver but an administrator concerned about financial results, the efficiency of a medical technique, etc. -- i.e., someone without any personal relationship to the patient.

- Second, to handle the changed circumstances, among other things a "need to know" principle [36] might be applied to limit the routine availability of detailed health and demographic data. To illustrate: If data about, say, a hospital's performance is needed only hospital-level patient aggregates might be provided, rather than complete individually identifiable patient detail.

Clearly much greater safeguards seem needed once there is no longer a personal bond between the patient and the individual using the data about that patient. Arguably, establishing a convincing system that would warrant the patient and public trust required here may be exceedingly difficult.

An important issue that may deserve comment is the "final" disposition of a patient's health (and related financial) records when the patient dies. Even for federal record systems, the Privacy Act no longer offers any protection, for example. We are learning more and more about the genetic causes of some illnesses. Matching records from deceased patients could put their descendants (or other relatives) at risk for possible differential treatment. If the view is taken, as quoted above in Goldman , that the patient "owns" his or her records then, by inference, upon death the estate of the patient owns that patient's records and their disposition is a matter to be settled by the heirs. In any event, inter or intra-generational record linkage needs careful consideration and might be done, as a rule, only with the consent of **all** individuals so linked.

4. Research Data Linkages Within the Health System

It can be argued that some research uses of data linkages within the Health System are administrative and so are already covered by the discussion in Section 3, especially subsection 3.2. There can be a fine line between applied research (intending to serve a permissible administrative purpose) and basic research (involving possibly an unanticipated analysis of variables originally obtained for another purpose).

Rather than try to draw the line, however, we will confine our attention to "basic research" since this involves some potentially new issues. In particular, our discussion will focus on researchers who are in some sense outside the Health Care System -- i.e., individuals that do *not already have access* to the patient data. Such a decision has consequences, of course. For example, important issues, like what research doctors do when using data about their own patients, go undiscussed. On the other hand, there is already an extensive body of practice on this topic and record linkage issues do not seem primary.

In any event, for the basic research setting we have confined attention too, figure 8 attempts to set out a summary of the main issues. As in Figure 7 earlier, included are some overall recommendations, legal and procedural questions are addressed, as well as perceptions concerns (both by the public and among the experts). These are further elaborated below.

Notification of patients about basic research uses may be sufficient in some settings while a specific consent may be needed in others. All basic research should be authorized by a review board mechanism of some sort with an annual public report, perhaps, to an outside citizens body. Requirements for securing consent pose difficult logistical and statistical problems that need extensive study. Anonymous group matching offers a potentially promising middle ground that could allow individual consent decisions to be honored, yet may not greatly sacrifice approved scientific ends [37]. However, as Figure 8 states, an extensive development and evaluation period is needed before this approach will prove its value.

Figure 8. -- Basic Research Data Linkages within the Health System

Broad Areas	Possible Response	
	Under what conditions (Column 1)	**For future study (Column 2)**
Overall Recommend-ations	Notification and even maybe consent required for individual linkages, plus research review board authorization	Statistical properties of group linkages and their use need extensive study when consent not given.
Technical Aspects	Elimination of all obvious (and not so obvious) identifiers. Access to data also limited by reidentification risks and "need to know"	Research on use of synthetic data. Continuous study of (ever) changing reidentification risks.
Legal Questions	Laws often unreasonably require no risk of redisclosure.	Research on "proof of harm" issue. Legislative and litigation research on contract based research access.
Public Views	Significant negative sentiment tied to distrust of government and lack of a specific clear purpose.	Study reactions to longterm (lifelong) record linkage
Expert Opinion	For the most part strongly favor broad basic research uses requiring only notification.	Nonmedical uses of health system records need more study.

The elimination of all identifying items about a patient would seem to be a necessary prerequisite for broad access to the health system data base by outside researchers. The risks of potential reidentification [38] are an ongoing concern, especially as nonhealth electronic systems grow in size and potentially have common variables which overlap those in health data bases. Research access through contractual arrangements as proposed by Herriot [39] has already begun in some settings (where it might be evaluated) and deserves study in others (where is has yet to be applied). The development of wholly synthetic data sets [40] also warrants work

and may be potentially promising because of the public assurances that can be given which might satisfy even those who greatly distrust government.

As noted earlier, there are a significant minority of individuals who oppose linkages and this group grows larger when there is no clear and compelling purpose for such linkage, except an ill-defined one -- like "basic research." [41] Lifelong patient linkage projects which are particularly attractive basic research tools may be subject to potentially severe public reaction if done without continuing consent (as occurred in Sweden [42]).

In general, even the strongest human rights advocates make an exception for research uses of individual data, stating [43] that "Information that is not personally-identifiable may be provided for research and statistical purposes." Given the growing power of probabilistic matching, though, we may not be far from the day when the only way to remove personally-identifiable information about some individuals is to remove all direct data concerning such individuals from a research file. Additionally, there may be some concerns about the appropriateness of nonmedical uses of health care records as, say, for the decennial census, [44] a point more appropriately covered in the next section.

5. Research Linkages between Health and Other Record Systems

Our discussion of basic research issues within the health system (Section 4) forms a bridge to a discussion of research data linkages between health and other record systems. Many parallels exist , as may be seen by comparing Figure 8 with Figure 9 below. There are, however, some new elements too:

- First, deterministic matching algorithms should be possible within the health system, assuming some form of health identifier is settled on. Generally, though, unless the SSN is used as the health identifier, only probabilistic matching methods will be available between health and nonhealth record systems; hence greater uncertainty about linkage quality will exist.

- Second, these nonhealth systems were clearly intended for nonhealth purposes; thus, their use in health record linkage research, through the simple expedient of health legislation, say, seems problematic. In fact, a strong case might be made for "consent only access" to at least some of them. Also any retroactivity in this expanded use should not be taken lightly either.

- Third, there seems to be a wide range of record linkage options, spanning matches to vital records at one end of the spectrum [45] (a traditional epidemiological tool) with tax records at the other [46] (something seldom done). The views of experts and the public appear to move predictably along this continuum from some acceptance to almost none [47].

- Fourth, even anonymous group matching methods need more study in this setting and not just their statistical efficiency as noted in Figure 8 but their public acceptability. Black males seem particularly opposed to, at least, some linkages. Concerns like those in Fisher et al. [48] merit examination here too.

As already noted, at least some experts are concerned about proposals using health records to improve the accuracy of the decennial census population count [49]. In fact, except in cases where explicit consent is obtained, it may make sense to confine all matches of health records to nonhealth records solely to those research purposes related to health. The control of any linkages between health and nonhealth records, say with Census Bureau data, needs careful study too [50]. Most Federal statistical agencies, for example, currently **lack** auditable record linkage systems [51]. and would have to greatly increase internal controls to meet what should be stringent electronic access (and audit) standards [52].

Figure 9. -- Research Data Linkages between Health and Other Record Systems

Broad Areas	**Possible Response**	
	Under what conditions (Column 1)	**For future study (Column 2)**
Overall	Generally consent should be required plus research review board authorization	Same as Figure 8.
Technical Aspects	Same as Figure 8.	Same as Figure 8.
Legal Questions	Conforming legislation needed to Tax Code, Social Security Act, etc.	Research on "proof of harm" issue. Legislative and litigation research on contract-based research access.

Public Views	Significant minority would not consent to individual linkages	Research on reactions to group linkages for statistical purposes. Study parallel to HIV testing.
Expert Opinion	For the most part strongly favor health research uses only requiring notification.	Nonmedical uses of any linkages need more study.

6. Nonresearch Linkages between Health and Other Record Systems

As may be apparent by now, in this paper there has been a progression from linkage opportunities that might be viewed by most individuals as beneficial, even to be encouraged, to linkages that are more problematic. This section discusses linkages that, in the view of many, may be dangerous and should generally be discouraged.

Figure 10 sets out a summary of possible issues in nonresearch linkages between health and nonhealth systems. Some overall observations on this figure might be worth making too -- highlighting what is new or controversial.

With the exception of a court order in a criminal case, all nonresearch linkages for nonhealth reasons should be prohibited. Even health administrative linkages (say, to use IRS address information to locate a person for health reasons) should be carefully limited (as is the case now). Areas for future study might include research on notification issues and consent-based exceptions. After all, new health needs keyed to helping individuals may arise over time and hence notification statements might need to be changed or at least their understanding reviewed periodically.

Figure 10. -- Nonresearch Administrative Data Linkages between Health and Nonhealth Record Systems

Broad Areas	**Possible Response**	
	Under what conditions (Column 1)	**For future study (Column 2)**
Overall Recommendations	For nonhealth reasons only with a court order. For health reasons only to directly aid patients.	Continuing research on (changing?) understanding of all consent or notification statements.
Technical Aspects	Minimizing redisclosure risks, especially to open or decentralized systems like vital records.	Continuing research on record keeping practices in nonhealth record systems, government and private.
Legal Questions	Ban any use of a new health identifier in nonhealth record systems.	Study conforming legislative needs.
Public and Expert	In generally close agreement, with a major-	Continuous routine monitoring.

Opinion	ity favoring restrictions on nonhealth uses.	

Existing systems, especially vital records, have many variables in common with health care record systems. Vital records are also quite open and hence they pose a significant risk of redisclosure, especially in public use (or other widely available) research files. If an independent health identifier is *not* used, then perhaps the SSN, for example, should be removed, or access to it restricted on birth and death records.

A legal ban, of course as generally advocated, should be imposed on the use of any new health identifier created, *except in health systems*. Research on other obvious and not so obvious identifiers, e. g., geographic details, should be ongoing to be sure that (legislated?) health record practices keep up with technology and the changing nature of unauthorized disclosure risks.

Public and expert opinion appear to both strongly oppose nonhealth administrative use of health record systems [53]. Additional public opinion research, though, seems needed on this point and others. For example, what are the public's views on the risks to any *new* health system from the *existing* centralized federal record systems (at IRS and SSA, for instance)? What about their views on the real danger of probabilistic matches to private data bases or to open or decentralized government systems, like vital records?

7. Summary Recommendations

Throughout this paper recommendations have been made that address aspects of privacy concerns in any large scale record linkage activity involving the proposed new health system or between that system and others. Figure 11 below provides a brief summary of these.

Figure 11. -- Selected Permissible Record Data Linkages by Purpose and Under What Conditions

Type of Data Linkage	Permissible and Under What Conditions
Administrative Data Linkages for the health of the patient	Just notification needed; if for use of patient and patient caregivers only
Other Administrative Data Linkages of Patient Records within the health system.	Greater safeguards seem needed once there is no longer a personal bond between patient and service provider (caregiver)
Basic Research Data Linkages within the Health System	Notification and even maybe consent required for individual linkages; research review board authorization.
Research Data Linkages between Health and Other Record systems	Generally consent should be required plus research review board authorization.
Nonresearch Administrative Data Linkages between	For nonhealth reasons, only with a court order. For

Health and Nonhealth Record Systems	health reasons, only to directly aid patients.

The overall treatment of linkage opportunities in this paper has gone from situations that simply called for a signed notification statement, preferably at regular intervals (Section 3), to suggested (Section 4) or required (Section 5) informed consent -- for linkage research in the health system or linked record research more generally . Finally (in Section 6), there was a brief discussion of how to **prevent** matching for nonhealth administrative purposes, except in rare instances. In all of these discussions, recommendations have been given along with the views of others; also areas for future study have been highlighted.

Frankly, this paper advocates a "go slow," careful approach to any attempt at data linkages undertaken as part of health care reform. It is unlikely that all the potential vulnerabilities of the new linkage system will be learned by anything other than experience -- hopefully not too hard won. Prototyping linkage experiments [54] are key. Patient consent and notification experiments will also be needed, as well as continuous study of public and patient opinion. An evolutionary rather than revolutionary strategy seems to represent the kind of humility and listening needed to avoid major blunders, especially in any advertent or inadvertent "takings" of privacy rights.

Much of the motivation around health reform speaks to efficiencies that can be gained with standardization of reporting and electronic data networking. These arguments seem to have merit; however, even if true, such changes will require a great many people to learn to do things in new ways and potentially paper records may need to continue to be employed for a long time (even if all new encounters are captured electronically).

Because the job is so big, it is important to begin **now but incrementally**. If structured properly, an orderly transition could be conducted, leaving ample time for human rights impacts to be respected.

8. An Afterword

An afterword may be worth making concerning the recommendations about "rights" in this paper; in particular, the rights to privacy and consent need to be set alongside the rights to universality and nondiscriminatory treatment [55].

Record linkage can aid a society in achieving advances in the well being its citizens. This point may have been lost in the detailed discussion of privacy and consent concerns. For example, the epidemiological literature is full of health studies that use record linkage techniques to advance knowledge [56].

The benefit side of record linkage can be oversold, however. A recent *Science* article may be worth quoting in this regard [57].

> *Over the past 50 years, epidemiologists have succeeded in identifying the more conspicuous determinants of noninfectious diseases -- smoking, for instance, which can increase the risk of developing lung cancer by as much as 3000%. Now they are left to search for subtler links between diseases and environment causes or lifestyles. And that leads to the Catch-22 of modern epidemiology. On the one hand, these subtle risks--say, the 30% increase in the risk of breast cancer from alcohol consumption that some studies suggest -- may affect such a large segment of the population that they have potentially huge impacts on public health. On the other, many epidemiologists concede that their studies are so plagued with biases, uncertainties, and methodological weaknesses that they may be inherently incapable of accurately discerning such weak associations. As Michael Thun, the director of analytic epidemiology for the American Cancer Society, puts it, "With epidemiology you can tell a little thing from a big thing. What's very hard to do is to tell a little thing from nothing*

> *at all." Agrees Ken Rothman, editor of the journal Epidemiology: "We're pushing the edge of what can be done with epidemiology." With epidemiology stretched to its limits or beyond, says Dimitios Trichopoulos, head of the epidemiology department at the Harvard School of Public Health, studies will inevitably generate false positive and false negative results "with disturbing frequency."*

Where does all of this leave things? The claim that the present paper is just a "starter set" is believed mainly to be true; but, in some places, even that may exceed current knowledge. What, in fact, many of the recommendations call for is simply more empirical work and hard thinking. Particularly crucial are two of these:

- Establishing ongoing programs of experimentation (e.g., on consent and notification statements), plus public opinion research on privacy issues, both in general and with a particular focus on record linkage [58].

- Instituting statistical work on group matching or other techniques that would lessen the tradeoff between the competing values of furthering scientific research **and** safeguarding personal privacy [59].

In the end, of course, the recommendations made here are simply the author's weighing of the evidence from the perspective of nearly 25 years of experience working on record link

Footnotes

[1] Health Security Act (1993). Washington, DC: U.S. Government Printing Office. See also, for example, Donaldson, M. S. and Lohr, K. N., (eds.) (1994*). Health Data in the Information Age: Use Disclosure and Privacy*, Committee on Regional Health Data Networks, Institute of Medicine: National Academy Press.

[2] Herriot, R. and Scheuren, F. (1975). The Role of the Social Security Number in Matching Administrative and Survey Records, *Studies from Interagency Linkages*, U. S. Social Security Administration.

[3] Despite early advocates, like Dunn, H. L. (1946). Record Linkage, *American Journal of Public Health*, 36, 1412-1416.

[4] Patterson, J. E. and Bilgrad, R. (1985). The National Death Index Experience: 1981-1985, *Record Linkage Techniques -- 1985*, Proceedings of the Workshop on Exact Matching Methodologies, Arlington Va.; Washington, DC: U. S. Department of Treasury.

[5] Tepping, B. (1968). A Model for Optimum Linkage of Records, *Journal of the American Statistical Association*, 63, 1321-1332.

[6] Fellegi, I. P. and Sunter, A. (1969). A Theory of Record Linkage, *Journal of the American Statistical Association*, 64, 1183 - 1210.

[7] Newcombe, H. B. (1967). Record Linking: The Design of Efficient Systems for Linking Records into Individual and Family Histories, *American Journal of Human Genetics*, 19, 335-359. Newcombe, H. B.; Kennedy, J. M.; Axford, S. J.; and James, A. P. (1959), Automatic Linkage of Vital Records, *Science*, 130, 3381, 954-959. Newcombe, H. B. and Kennedy, J. M. (1962), Record Linking: Making Maximum Use of the Discrminating Power of Identifying Information, *Communications of the Association for Computing Machinery*, 5, 563-566.

[8] See, for example, Beebe, G. W. (1985). Why are Epidemiologists Interested in Matching Algorithms? *Record Linkage Techniques*, Proceedings of the Workshop on Exact Matching Methodologies, Arlington, Va.; Washington, DC: U.S. Department of Treasury. See also, [56] and [57].

[9] See, for example, Winkler, W. and Thibaudeau, Y. (1991). An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U. S. Census, *Statistical Division Report Series, CENSUS/SRD/RR - 91/09*. See also, Belin, T. and Rubin, D. (1995). A Method of Calibrating False-Match Rates in Record Linkages, *Journal of the American Statistical Association*, 90, 694 - 707.

[10] Rogot, E.; Sorlie, P. D.; Johnson, N. J.; Glover, C. S.; and Treasure, D. W. (1988). *A Mortality Study of One Million Persons: First Data Book*, NIH Publication No. 88-2896, Bethesda, MD: Public Health Service, National Institute of Health.

[11] Oh, H. L. and Scheuren, F. (1975). Fiddling Around with Matches and Nonmatches, *Proceedings of Social Statistics Section, American Statistics Association*. Also, Scheuren, F. and Winkler, W. E. (1997), Regression Analysis of Data Files that Are Computer Matched -- Parts I and II, in this volume: *Record Linkage Techniques -- 1997*, Washington, DC: National Academy Press. (Part I appeared previously in *Survey Methodology*, (1993), 19 (1) 39-58, Statistics Canada; Part II was delivered at the XII Methodology Symposium, Ottawa Canada, November 1, 1995, under the title Linking Data to Create Information and will be included in a forthcoming issue of *Survey Methodology*.)

[12] Fair, M. (1995). An Overview of Record Linkage in Canada, presented at the American Statistical Asso-

ciation Annual Meetings in Orlando, FL, August 1995.

[13] Davis, K. (1995). Guarding Your Financial Privacy, *Kiplinger's Personal Finance Magazine*, 49.

[14] Office of Technology Assessment (1993). *Protecting Privacy in Computerized Medical Information*, Washington, DC: U.S. Government Printing Office.

[15] Scheuren, F. (1993). Correspondence with Dr. Elmer Gabrieli on a health identifcation number, in *Guide for Unique Healthcare Identifier Model,* ASTM, Philadelphia, May, 1993 draft. Ironically, public opinion poll data suggest that the American people favor the adaptation of the SSN, rather than the introduction of a new health identifier. See [33] for details.

[16] *Health Security Act* (1993). Washington, DC: U.S. Government Printing Office.

[17] Donaldson, M. S. and Lohr, K. N., (eds.)(1994). *Health Data in the Information Age, Use, Disclosure, and Privacy*, Committee on Regional Health Data Networks, Institute of Medicine: National Academy Press.

[18] Schwartz, H.; Kunitz, S.; and Kozloff, R. (1995). Building Data Research Resources From Existing Data Sets: A Model for Integrating Patient Data to Form a Core Data Set, presented at the American Statistical Association Annual Meetings in Orlando, FL, August 1995.

[19] Dalenius, T. (1988). *Controlling Invasion of Privacy in Surveys*, Continuing Education Series, Statistics Sweden.

[20] Olmstead v. United States. 277 U.S. 438. 478 (1928) (Justice Brandeis dissenting).

[21] Some implications are obvious. For example, "information in medical records can conceivably affect you for the rest of your life if revealed to an employer or insurance company," (*The Washington Post* Health Section, February 8, 1994). The obvious cases are not the only ones to be worried about, though. The combinatorial possibilities are so great that they may not only impair full consent to linkage by patients but also access decisions by data stewards.

[22] Singer, E.; Shapiro, R.; and Jacobs, L (1995). Privacy of Health Care Data: What Does the Public Know? How Much Do They Care? Paper submitted with support from the American Association for the Advancement of Science, Science and Human Rights Program.

[23] Horgan, J. (1995). From Complexity to Perplexity, *Scientific American*, June 1995, 104-109. See also, Waldrop, M. M. (1992), *Complexity*. New York: Simon and Schuster.

[24] Superhack, *Scientific American,* July Issue, 1994, 17. This is a story of a group of about 600 computer "hacks,'" collaborating over the internet, who broke a computer security encryption algorithm. About 17 years earlier, it was predicted that this feat would take 40 quardrillion years. Once the effort started, it took 8 months! For more on this, see also, *Science,* May, 1994, 776-777.

[25] In contrast, consider *The Washington Post*, July 18, 1994, where there is a story about how, despite an existing monitoring system, inadequate controls were used for access to sensitive information.

[26] Brannigan and Beier (1995), Medical Data Protection and Privacy in the United States: Theory and Reality, paper submitted with support from American Association of the Advancement of Science, Science and Human Rights Program.

[27] Singer, E., Shapiro, R., and Jacobs, L (1995), *op. Cit.*

[28] Hoffman, B. (1990). Patient Confidentiality and Access to Medical Records: A Physician's Perspective, *Health Law in Canada,*10:210-12. Siegler, M. (1982), Confidentiality in Medicine -- A Decrepit Concept, *New England Journal of Medicine,* 307:1518-21, as summarized by Cummings, N. (1993), Patient Confidentiality, *Second Opinion*, 112-116.

[29] Brannigan and Beier (1995), *op. cit.*

[30] Introduced by Condit as HR 4077 in the103rd Congress; also reintroduced (again by Condit) in the 104th Congress as HR 435.

[31] As recommended by OTA (1993), *op. cit.*

[32] For example, Smith, R.E. (1992). Compilation of State and Federal Privacy Laws, *Privacy Journal*.

[33] Inferred from Harris-Equifax (1993), *Health Care Information Privacy: A Survey of the Public and Leaders*, New York: Louis Harris and Associates. See also, Blair, J. (1995), Ancillary Uses of Government Administrative Data on Individuals: Public Perceptions and Attitudes, Unpublished Working Paper, Committee on National Statistics, National Academy of Sciences. As Blair points out (and this author confirmed by calling Harris and Associates), the Harris-Equifax survey has important limitations on its interpretability; nonetheless, its main conclusions are in essential agreement with other research on privacy concerns. Blair summarizes these as well. Roughly, almost no matter how you ask the question, there are always about one sixth to one fifth of the population who oppose electronic record linkages on privacy grounds. Conversely, again almost no matter how you ask the question, about the same fraction will favor beneficial sounding linkages on efficiency grounds. The two thirds or so in the middle will differ in their opinions depending on the specifics. See also, [55].

[34] Goldman, J. (1994). Regarding H.R. 3137: Data needs and related issues for implementing health care reform, Statement before the House Post Office and Civil Service Subcommittee on Census, Statistics and Postal Personnel, Washington, DC. For an excellent expression of an alternative view, see Newcombe (1995), When Privacy Threatens Public Health, *Canadian Journal of Public Health*, 86, 188-192.

[35] Kluge, E. H. (1993). Advanced Patient Records: Some Ethical and Legal Considerations Touching Medical Information Space, *Methods of Information in Medicine*, 95-103.

[36] Brannigan and Beier (1995), *op. cit.*

[37] Spruill, N. and Gastwirth, J. (1982). On the Estimation of the Correlation Coefficient from Grouped Data, *Journal of the American Statistical Association*, 77, 614-620. Gastwirth, J., and Johnson, W.O. (1994), Screening With Cost-Effective Quality Control: Potential applications to HIV and Drug Testing, *Journal of the American Statistical Association*, 89, 972-981. Contrast Gastwirth, J. (1986), Ethical Issues in Access to and Linkage of Data Collected by Government Agencies, *Proceedings of the American Statistical Association, Social Statistics Section*, 6-13.

[38] See, for example, Jabine, T.B. and Scheuren, F. (1985). Goals for Statistical Uses of Administrative Records: The Next Ten Years, *Journal of Business and Economic Statistics*.

[39] Wright, D. and Ahmed, S. (1990). Implementing NCES's New Confidentiality Protections, American Statistical Association, *1990 Proceedings on the Section on Survey Research Methods*, Alexandria, Va.: American Statistical Association.

[40] Rubin, D. B. (1993). Comments on Confidentiality, A Proposal for Satisfying all Confidentiality Constraints through the Use of Multiple-Imputed Synthetic Microdata, *Journal of Official Staistics*.

[41] Harris-Equifax (1993) and Blair (1995), *op. cit.*, See also, [33]. Clearly, though, we do not know enough to be sure.

[42] Dalenius, T. (1988). *op. cit.*

[43] Goldman, J. (1994). *op. cit.*

[44] Singer, E.; Shapiro, R.; and Jacobs, L (1995). *op. cit.*

[45] Fair, M. (1995). *op. cit.*

[46] But see, for example, Scheuren, F. (1994). Historical Perspectives on the Estate Multiplier Technique, *Statistics of Income, Estate Tax Wealth Compendiun,* U. S. Internal Revenue Service.

[47] Scheuren, F. (1985). Methodological Issues in Linkage of Multiple Data Bases, *Record Linkage Techniques -- 1985*, Washington, DC: Department of the Treasury, Internal Revenue Service,155-178. Scheuren, F. (1995), Review of Private Lives and Public Policy, *Journal of the American Statistical Association*, March 1995 Issue.

[48] Fisher, J. et al. (1995). Gaining Respondent Participation: Issues of Trust, Honesty and Reliability, Paper submitted with support from American Association of the Advancement of Science, Science and Human Rights Program.

[49] Singer, E.; Shapiro, R.; and Jacobs, L (1995). *op. cit.*

[50] One joint control option that may be of interest arose in the project described in Rogot, E. et al. (1988). *op. cit.*

[51] Scheuren, F. (1995). *op. cit.*

[52] Brannigan and Beier (1995). *op cit.*

[53] This might be inferred from the 1993 Harris-Equifax Questions on access to patient health data by insurance companies and employers. Harris-Equifax, *op. cit.* Also, from Blair (1995) and the other research started by Scheuren (1985). See [33] and [47].

[54] Schwartz, H. et al. (1995). *op. cit.*

[55] As elaborated in Chapman, Audrey R. (1997). Introduction: Developing Health Information Systems Consistent with Human Rights Criteria, *Health Care and Information Ethics: Protecting Fundamental Human Rights*, Kansas City, MO: Sheed and Ward, 3-30.

[56] Cited earlier were Beebe [8] and Fair [12], among others. See also, endnotes [7], [10], and [34]. Also of note in this context is the paper by Sugarman, Jonathan, et al. (1997). Improving Health Data among American Indians and Alaska Natives: An Approach from the Pacific Northwest, *Health Care and Information Ethics: Protecting Fundamental Human Rights,* Kansas City, MO: Sheed and Ward, 88-113.

[57] Taubes, G. (1995). Epidemiology Faces its Limits, *Science*, July 14, 1995, 164-169.

[58] As advocated in Scheuren, F. (1985). Methodological Issues in Linkage of Multiple Data Bases, *Record Linkage Techniques -- 1985*, Internal Revenue Service and as pursued by him over the past 10 years through the sponsorship of numerous public opinion polls, asking various questions about linkage. Most of these are discussed in Blair, J. (1994). *Ancillary Uses of Government Administrative Data*, College Park, MD: University of Maryland Survey Research Center. Work at the Bureau of Labor Statistics, with focus groups and other cognitive research techniques, has also been sponsored. At this point, the summary given already in endnote [33] represents the limited state of knowledge.

[59] Certainly, the seminal work of Spruill, Nancy and Gastwirth, Joseph (1982). On the Estimation of the Correlation Coefficient from Grouped Data, *Journal of the American Statistical Association*, 77, 614-620.

Additional References

Acheson, E.D. (1967). *Medical Record Linkage*, Oxford, U.K.: Oxford University Press.

Copas, J.B. and Hilton, F.J. (1990). Record Linkage: Statistical Models for Matching Computer Records, *Journal of the Royal Statistical Society*, Ser. A, 153 (Part 3), 287-320.

Dunn, H.L. (1946). Record Linkage, *American Journal of Public Health*, 36, 1412-1416.

Jaro, M. A. (1989). Advances in Record-Linking Methodology as Applied to Matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association*, 84, 414-420.

Kilss, B. and Alvey, W. (eds.) (1985). *Record Linkage Techniques -- 1985, Proceedings of the Workshop on Exact Matching Methodologies,* Arlington, Virginia, May 9-10, 1985), Washington, DC: Department of the Treasury, Internal Revenue Service.

Newcombe, H.B. (1967). Record Linking: The Design of Efficient Systems for Linking Records into Individual and Family Histories, *American Journal of Human Genetics*, 19. 335-359.

Newcombe, H.B., and Kennedy, J.M. (1962). Record Linking: Making Maximum Use of the Discriminating Power of Identifying Information, *Communications of the Association for Computing Machinery*, 5, 563-566.

Rogot, E.; Sorlie, P.D.; Johnson, N.J.; Glover, C.S.; and Treasure, D.W. (1988). *A Mortality Study of One Million Persons: First Data Book*, NIH Publication No. 88-2896, Bethesda, MD: Public Health Service, National Institute of Health.

Roos, L.L.; Wajda, A.; and Nicol, J.P. (1986). The Art and Science of Record Linkage: Methods that Work with Few Identifiers, *Computers in Biology and Medicine*, 16, 45-57.

Scheuren, F. (1985). Methodological Issues in Linkage of Multiple Data Bases, *Record Linkage Techniques – 1985*, Washington, DC: Department of the Tresury, Internal Revenue Service, pp. 155-178.

Scheuren, F.; Alvey, W.; and Kilss, B. (1986). Record Linkage for Statistical Purposes in the United States, *Proceedings of the Workshop in Computerized Record Linkage in Health Research*, held in Ottawa, Ontario, May 21 - 23, 1986, G.R. Howe and R.A. Spasoff, (Eds.), Toronto: University of Toronto Press, pp. 198-210.

Donaldson, M.S. and Lohr, K.N., (Eds.)(1994). *Health Data in the Information Age: Use, Disclosure and Privacy*, Committee on Regional Health Data Networks, Institute of Medicine: National Academy Press.

Note: This paper was commissioned for the Health Care and Information Ethics project, sponsored by the American Association for the Advancement of Science (AAAS) when Fritz Scheuren was working for George Washington University. It appeared as a chapter in Audrey R. Chapman (Ed.), (1997). *Health Care and Information Ethics: Protecting Fundamental Human Rights*, Kansas City, MO: Sheed and Ward. The paper is reprinted here with permission.

Chapter 11

Record Linkage in an Information Age Society

Martha E. Fair, Statistics Canada

Abstract

As we move into the 21st century the acquisition, generation, distribution, and application of statistical knowledge in a timely fashion will become more important. Required are innovations in terms of the products, technologies, and the way in which we generate, disseminate, and use statistical data and information. It is anticipated that work units will shrink, funding will be limited, and there will be greater analytical uses of administrative, as well as survey and census data. There may need to be a fundamental rethinking and radical redesign of business processes and workplaces. Today's market, customer values and technologies are changing rapidly. Standards, cooperation and collaboration of various agencies, and software developments are very important. Access and control of sensitive information as well as the technical aspects of confidentiality are necessary. Data integration of a number of different sources, including census, survey, registry and administrative files in a variety of economic and social areas are sometimes required. The quality of the statistical information is also of concern.

One useful tool that has been developed for generating and using statistical data is computerized record linkage. Anticipated new developments and applications of this methodology for the 21st century are described. Emphasis is placed on the health area, particularly in these times of health reform.

Over the past 15 years, generalized systems have been developed at Statistics Canada. Briefly described is a new version of a generalized record linkage system (GRLS.V3) that is being put into place to carry out internal and two-file linkages. With an earlier mainframe system, large-scale death and cancer linkages for members of survey and other cohorts have been shown to be practicable using the Canadian Mortality Data Base, the Canadian Cancer Data Base and the Canadian Birth Data Base. This approach has greatly reduced respondent burden, lowered survey costs, and greatly refined the detection and measurements of differences in geographic, socio-economic and occupational groups. Some of the past successes are described, particularly where longitudinal follow-up and creation of new sampling frames are required. For example, the Nutrition Canada Survey, the Canada Health Survey and Fitness Canada Surveys have been linked with mortality data. Some examples of the use of follow-up of census data are discussed (e.g., a study of farmers using 1971 Census of Agriculture and Census of Population).

This paper was reprinted with permission from the *Proceedings of the Census Bureau's Conference and Technology Interchange*, March 17-21, 1996.

Introduction -- Statistical Data Needs for the 21st Century

Purpose

The purpose of this article is to discuss some of the issues surrounding **statistical uses of record linkage**, with a view to the expanded uses of **probabilistic record linkage** in the 21st century, particularly with respect to the generation and use of administrative and survey data. Record linkage is the bringing together of two or more pieces of information relating to the same entity (e.g., individual, family, business). In probabilistic record linkage, the comparison or matching algorithm yields for each record pair, a probability or "weight" which indicates the likelihood that record pairs relate to the same entity (Fair, 1995).

In the 21st century, it is anticipated that those carrying out and requiring record linkage of data should be prepared for change. Hardware and software needs for record linkage will range from global statistical systems for giant organizations on large super computers, to requests for linkages of small area data sets on small laptops. Integration of a variety of statistical survey and administrative data sources may be required. There is a move to reduce the complexity of data, to avoid unnecessarily duplicating data, and to have a single, unified view of an organization's information, with the data's physical location being almost transparent to the user. There is considerable re-engineering of data acquisition processes, including the editing, manipulating and grouping of files. This should improve the quality of the input files. Data models may be centered around the same individual, family or entity over time rather than a cross-sectional snapshot of an event. It is anticipated that databases will become more comprehensive and inclusive. There will be a need to develop and revise international data standards, such as for disease, geographic, industrial coding, and data exchange. Timeliness is important with many organizations moving to electronic data capture and optical imaging. Dissemination of products will be via a spectrum of medium, with emphasis on the usefulness to the customer. On-line access may be required for inquiry, downloading and reporting. New links between agencies and countries may be required, and hence confidentiality issues will be of prime importance. Here, it is useful that statistical and administrative record linkage applications be differentiated.

Today, we will examine some **general topics** first, namely:

- evolving in response to customer needs in changing times;
- some comments regarding the "information age;"
- characteristics/indicators of success for an effective statistical system; and
- moving from data to information.

We will then look at **record linkage** in more depth and examine:

- today's situation;
- examples of present uses of record linkage;
- preparing for the future journey -- the life cycle of events;
- making the right connection; and
- summary.

I will use examples of statistical applications of record linkage, with emphasis on those from Statistics Canada and the health research area in particular.

Evolving in Response to Customer Needs in Times of Change

Many of the common social, economic, occupational and environmental concerns of today are complex and multi-faceted. **Change** seems to have become the operative word. **Policies, institutions, communities and businesses** are changing at the global, national, provincial/state, regional and local levels. Institutions in North America and worldwide have undergone an unprecedented wave of consolidation. There is concern to identify and strive toward **global statistical systems** that can produce national statistical services that are comparable and readily accessible (Haberman, 1995). The capabilities of **technology,** especially communication and information technology are changing daily.

The **tools and options for dissemination** are expanding. There is a corresponding rising consumer expectation, particularly with respect to timeliness and quality of statistical data products. This has implications in terms of standard data concepts, definitions, coding, methodology used for record linkage, and development of national and provincial/state data bases. Communication and collaboration of various countries, particularly with respect to software and methodological developments, have benefited through a series of seven workshops regarding record linkage held in Canada (e.g., Carpenter and Fair, 1989) and others held in the United States (Kilss and Alvey, 1985).

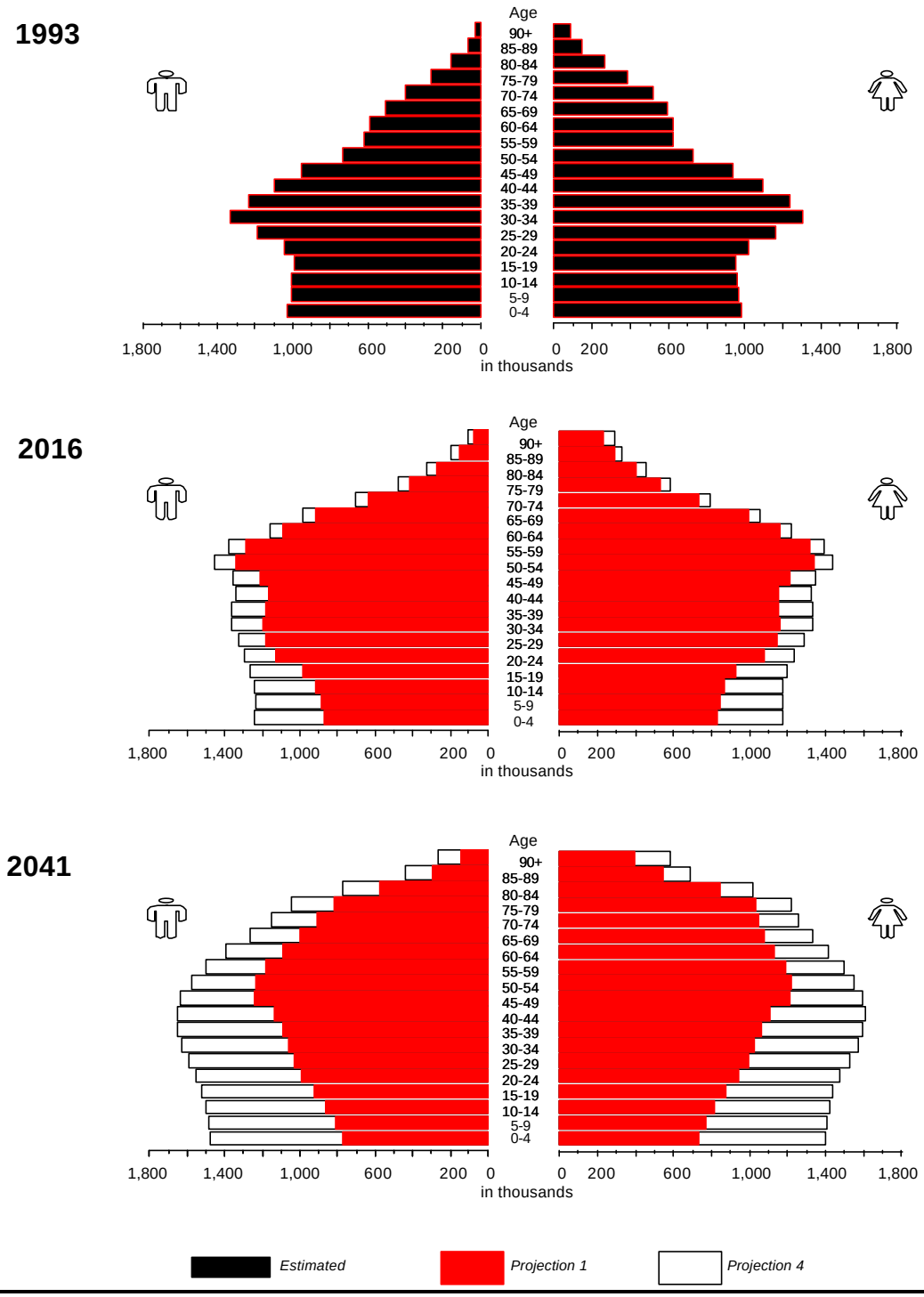
Analysis of data sources from different countries is helpful and comparative international statistics are required. Joint analysis of data from different countries is common (e.g., a joint analysis of 11 underground miners studies to examine radon and lung cancer risks). There is a need for **international collaborative works**, such as the United Nations Scientific Committee on the Effects of Atomic Radiation, which aims to provide to the scientific and world community, its latest evaluations of the sources of ionizing radiation and the effects of exposures (United Nations, 1993). Here the major aim is to assess the consequences to human health of a wide range of doses of ionizing radiation and to estimate the dose of people all over the world from natural and man-made radiation sources. Linkage of a variety of data sources are required.

The **social and economic structure** is changing. There are **new concepts of family, childhood and parenthood.** This has important implications for the follow-up of households and individuals for longitudinal surveys and for administrative files. *The Health of Canada's Children - A CICH Profile* discusses some of the recent trends in Canada (Canadian Institute of Child Health, 1994). Some of the examples given are as follows. **Families are changing** -- the structure of the families are different from what they used to be. In 1967, 65% of all Canadian families consisted of a male wage earner and a stay-at-home spouse. In 1990, this traditional family structure accounted for only 15% of families.

Our **society is becoming more diverse.** Families are rooted in more diverse cultural, religious, linguistic and ethnic backgrounds than in the past. In 1991, 13% of the Canadian population spoke a language other than French or English at home. Where surnames and forenames are used in probabilistic matching, we have found that special tables of weights have had to be developed by region, and sometimes over time. For example, there is quite a different distribution of name frequencies in Quebec, which is predominantly French, in British Columbia, where the number of Asian names have increased in recent years, and in Canada overall. Naming conventions are changing, with women often retaining their maiden name, as is particularly common in Quebec.

The **role of women** has changed. In the early 1960's less than a third of Canadian women worked outside the home. By 1996, about 80% of women are expected to be in the work force. **Lifelong learning has become a necessity.** Increasingly, the workplace requires a higher level of skills and a different set of skills than in the past. More and more jobs require people who can work in teams, who have high literacy, numeracy and computing skills, who can then critically and creatively solve problems -- and most of all, continue to learn new skills.

Figure 1. -- Population by Age Group and Sex, Canada, 1993, 2016 and 2041



Source: George, M.V.; Norris, M.J.; Nault, F.; Loh, S.; and Dai, S.Y. (1994), *Population Projections for Canada, Provinces and Territories 1993-2016*, Statistics Canada Catalogue No. 91-520, page 74.

Between 1971 and 1991, the **age profile** of the population changed from a traditional pyramid shape to a wide column, with fewer younger people and dramatically more **older** people. By 2041 the column will be top-heavy (see Figure 1-- Source: George et al., 1994). Similarly in the United States, by 2025 more than

30% of the population will be over 55 year old. Persons aged 80 and over will outnumber any younger 5-year age group (UNDIESA, 1991).

The **economy is restructuring.** There is a heightened sense of economic anxiety. Driven by technological innovation, global competition and new trade arrangements, the economy is undergoing a fundamental restructuring. **Governments are restructuring**. At all levels, they are tightening up their spending to make programs more cost effective and more relevant to changing needs. This has been most apparent in the health sector in term of **health reform**. (Blomquist and Brown, 1994). At the same time, there are **major reforms of social programs**, not only to improve efficiency, but also to remake these programs.

The Information Age

The Tofflers have described our times as being that of the third wave (Toffler and Toffler, 1995). The **First wave** was agricultural and it lasted thousands of years until the 18th century. Then the Industrial Revolution created a novel concept of massification -- mass production, mass markets, mass consumption, mass media, mass political parties, mass religion and weapons of mass destruction. This **Second wave** lasted about three hundred years. The **Third wave** is that of an **information-age society**. Because of the computer chip we are moving from an age in which we produce things to an age in which we produce information. But paradoxically, the more that national boundaries are usurped by our universal hook-up to the global computer network, the more we segment (Grant, 1996). With the complexities of the new system we require more and more information exchange among the various units of companies, government agencies, hospitals, associations, institutions and individuals. Factories, cities, even nations are receding and being replaced by smaller units of consumption and by minority political and religious interests. In the Tofflers' words, the world "de-massifies."

We are in a time of **redefining the workplace -- and work itself**. Work units are shrinking. The home may be the workplace of the future for many more people. **Customized and semi-customized, highly diversified statistical products** will be required -- yet the cost of producing these diversified products must be minimal. There is a requirement of flexibility and choice by many clients.

There is a growing **time crunch**. Time itself is one of the most important economic resources. The ability to shorten time -- by communicating swiftly or by bringing products in a timely fashion -- may mean the difference between profit and loss (Toffler and Toffler, 1995). In the health area, there is a need for more flexible, fast-paced, information-rich systems which can act as surveillance systems and assist in identifying present and emerging health issues.

We may have to **rethink and re-image our relationships**. Amidst societal change, people more than ever need an anchor, a refuge, a place where they belong (Bank of Montreal, 1995). Traditionally, a sense of **community** has helped fill that need. This in the past, was often built around a common geographic location, a common workplace, a common history or tradition. Individuals now form commitments to a wide variety of communities based on shared experiences and values -- family, profession, neighbourhood, age, ethnic background, talent, language. Barna (1990) notes that in the process of redefining what counts in life, many of us have decided that **commitment** is not in our best interest. Traditional concepts such as loyalty and the importance of memberships in various groups have been thrown out in favour of personal interest and self-preservation. This may have important implications for the workforce and for negotiations.

Characteristics and Indicators of an Effective Statistical System

Dr. Fellegi, the Chief Statistician at Statistics Canada, gave a 1995 Morris Hansen Lecture at the Washington Statistical Society. He described an effective statistical system as being characterized by its ability to:

- illuminate issues, not just monitor them;
- evolve in response to needs;
- be aware of priority needs;
- set priorities;
- have a high level of public credibility, since few in society can verify national statistics; and
- be free from undue political interference (Fellegi, 1995).

Three main indicators of success of statistical systems noted in this paper were:

- How adaptable is the system in adjusting its product line to evolving needs?
- How effective is the system in exploiting existing data to meet client needs?
- How credible is the system in terms of the statistical quality of its outputs and its non-political objectivity? (Fellegi, 1995)

Moving from Data to Information

Two recent methodology symposium topics held at Statistics Canada are relevant. The XIIth International Symposium on Methodological Issues, held at Statistics Canada on November 1-3, 1995, was entitled "From Data to Information." At this symposium topics included the role of statistics in making social policy, data integration, analytical methods, access and control of data, quality of statistical information, technical aspects of confidentiality, making data accessible to the general public, data warehousing, and electronic information dissemination. An earlier symposium dealt with re-engineering for statistical agencies (Statistics Canada, 1994). Re-engineering is a rethinking and radical redesign of the way business is carried out by an agency or corporation. The desired end results are lower production costs, quicker dissemination, and higher customer satisfaction.

There is a desire to understand and improve the performance of the health system. As noted *in Health Data in the Information Age -- Use, Disclosure and Privacy* (Donaldson and Lohr, 1994) this in turn motivates proposals for the creation and maintenance of comprehensive, population-based health care data bases. Regional health care databases are being established around the United States and Canada. Guidelines are needed to realize the full potential of these files, as well as to reduce respondent burden.

Two critical dimensions of databases are their **comprehensiveness and inclusiveness**. Comprehensiveness describes the completeness of the records (i.e., the amount of information one has for each patient and for an individual over time). Inclusiveness refers to which populations in a geographic area are included in a database. The more inclusive a database, the more it approaches coverage of 100 percent of the population. The Census of Population, the vital statistics and morbidity files are important data sources for a variety of national health studies because of their comprehensiveness and inclusiveness.

Record Linkage

Today's Situation

Just as we have just looked into the future in a more general fashion, it is also good to reflect on some of the past development of record linkage methods. Some of today's data sources were created by individuals with a view to record linkage in the future (e.g., in Canada the vital statistics birth records were linked with Family Allowance files to determine the eligibility of applicants when this program was first implemented).

The initial definition of record linkage was in terms of the book of life (Dunn, 1946). The early development work had to do with investigating the feasibility of probabilistic linkage (Newcombe et al., 1959), the theory of record linkage (Fellegi and Sunter, 1969), the development of specific computer programs, followed by the development of generalized software (Hill, 1981) and national files, commercial software (Jaro, 1995) and other software (e.g., Chad, 1993). Communication and collaboration with agencies in various provinces in Canada, in the United States, the United Kingdom and Australia have aided record linkage developmental work (e.g., Kilss and Alvey, 1985; Gill et al., 1993; Jaro, 1995; Winkler and Scheuren, 1995).

One key technological development is the shift from a paper-based system of records to an **electronic process** for creating, transmitting and disseminating products. At Statistics Canada, the 1990s brought about a major revolution in advanced technology with the wide-scale introduction of Computer Assisted Interviewing (CAI) for household, agriculture and business surveys. Computer Assisted Personal Interviewing (CAPI) has been introduced with the Labour Force Survey supplements and longitudinal household surveys covering a wide range of topics including Survey of Income and Labour Dynamics, the National Population Health Survey and the National Longitudinal Survey of Children (Gosselin, 1995). Vital statistics (Starr and Starr, 1995), census and cancer registries are additional examples where re-engineering and change may be anticipated in the future. There has been a move from microfilming of source documents to **optical imaging**.

A **generalized system** initiative at Statistics Canada was started in response to the use of repetitive processes, particularly in survey taking. This includes sampling, data collection and capture, automated coding, edit and imputation, estimation, and record linkage (Doucet, 1995). This **suite of software products** has been developed with technologies that make them highly portable across major computing platforms.

The original version of generalized record linkage software (GRLS.V1) that was developed at Statistics Canada was for a mainframe environment. Currently under development is GRLS.V3 which runs in a client-server environment with ORACLE and a C compiler (Statistics Canada, 1996). GRLS will run on a PC or workstation which supports the UNIX operating system. This software allows for an internal linkage within a file (e.g., to create health histories in a cancer registry) or a two-file linkage (e.g., linkage of a survey file to mortality). This software is particularly useful where there is no unique, reliable, lifetime identifier on the files being linked.

GRLS has three important stages:

- In the **searching stage** screens are used to specify the files, indicate the records to be compared (e.g., within pockets with similar phonetic code of the surname), specify the rules for comparison (e.g., agree, disagree, partially agreement, or user-defined functions), and specify the weights to be assigned to the outcomes.

- In the **decision stage**, the weights can be adjusted and threshold weights selected to define whether pairs are linked, possibly linked, or unlinked.

- In the final **grouping stage**, the records are brought together appropriately. You can have conflicts resolved automatically (e.g., two records linking to one death record). This is called mapping, and one can select the appropriate type (e.g., 1-1, 1-many, many-1, many-many). You may also have the option to resolve conflicts manually via on-screen updates. The final output of GRLS is an ORACLE table containing the GROUP information.

It is very important to note that GRLS V3 **does not modify** the files it is linking. This means that the same file may participate concurrently in several two-file linkages. For example, one might want to link several (and unrelated) files against the same master file.

Record Linkage in the Toolbox of Software -- Some Examples of its Use

Statistics Canada uses a common set of software products in re-engineering its administrative and statistical programs. This set of products is collectively referred to as the toolbox. Each toolbox product has a current release, an identified support level and a designated support centre. Currently the generalized record linkage software is part of this toolbox.

Record linkage is an important tool for the creation of statistical data, particularly in relation to census taking. Some of the important uses are as follows:

- **Data Quality.**--Some European countries use population registers instead of a census (e.g., Denmark). It is also possible to use administrative data and record linkage to help impute missing or inconsistent data. Data sources can be examined to eliminate duplicate records for individuals and to identify missing records in databases (e.g., by the linkage of infant deaths and birth records or by the linkage of births and deaths with census records).

- **Bias.**--The advantage of population-based record linkage includes the avoidance of selection bias, which can occur in cohort and case-control studies. Recall bias is usually avoided because the data are collected before the outcome or in ignorance of the outcome.

- **Coverage.**--In Canada record linkage data is used to improve the census coverage (e.g., address register) as well as to estimate its coverage (e.g., reverse record check). With disease-specific registries, it is possible to use linkage to identify underreporting of cases (e.g., by linkage of cancer registries with death registrations, the linkage of hospital records with deaths for heart disease). This has important implications for diseases such as AIDS and cancer.

- **Tracing Tool.**--Record linkage and administrative records are often used to follow-up cohorts to determine the individuals' vital status. Tracing is often needed for follow-up of industrial cohorts and for longitudinal surveys to obtain the cause of death and/or cancer. Mobility patterns of persons are important for the allocation of health resources.

- **Benchmarking/Calibration.**--Combining results from several data collection sources may give improved estimates (e.g., use of income from tax, survey and census sources).

- **Sampling Frame.**--Record linkage may be involved in setting up a sampling frame for surveys (e.g., census of agriculture farm register is used for the sampling of intercensal farm surveys).

- **Supplementary Surveys.**--Several postcensal surveys have been carried out following the Canadian census. Examples include the aboriginal peoples, and the health and activity limitation surveys. Data from the survey can be linked with that available on the census.

- **Release of Public-Use Tapes.**--Linkage can be used to examine public-use tapes for potential problems in their release (e.g., data crackers).

- **Building New Data Sources (e.g., Registries).**--Some cancer registries combine a variety of data sources using record linkage to generate their registry. Some of the data sources include hospital admissions, pathology reports, records from clinics, and death registrations.

- **Creation of Patient-Oriented, Rather than Event-Oriented Statistics.**-- (e.g., for hospital admissions, for cancer registries, (Dale, 1989)).

The uses of linkage in analytical studies have often been varied, and are generally tied in with increased use of administrative records for statistical purposes and with the reduction of respondent burden. (A roundtable luncheon of the Social Statistics Section at the 1995 American Statistical Association, chaired by G. Hole, discussed some of the above and future uses of administrative records to complement/ supplement data from household surveys.)

A more complete list of some of the uses of record linkage have been described earlier (Fair, 1995; Newcombe, 1994). Some examples are as follow:

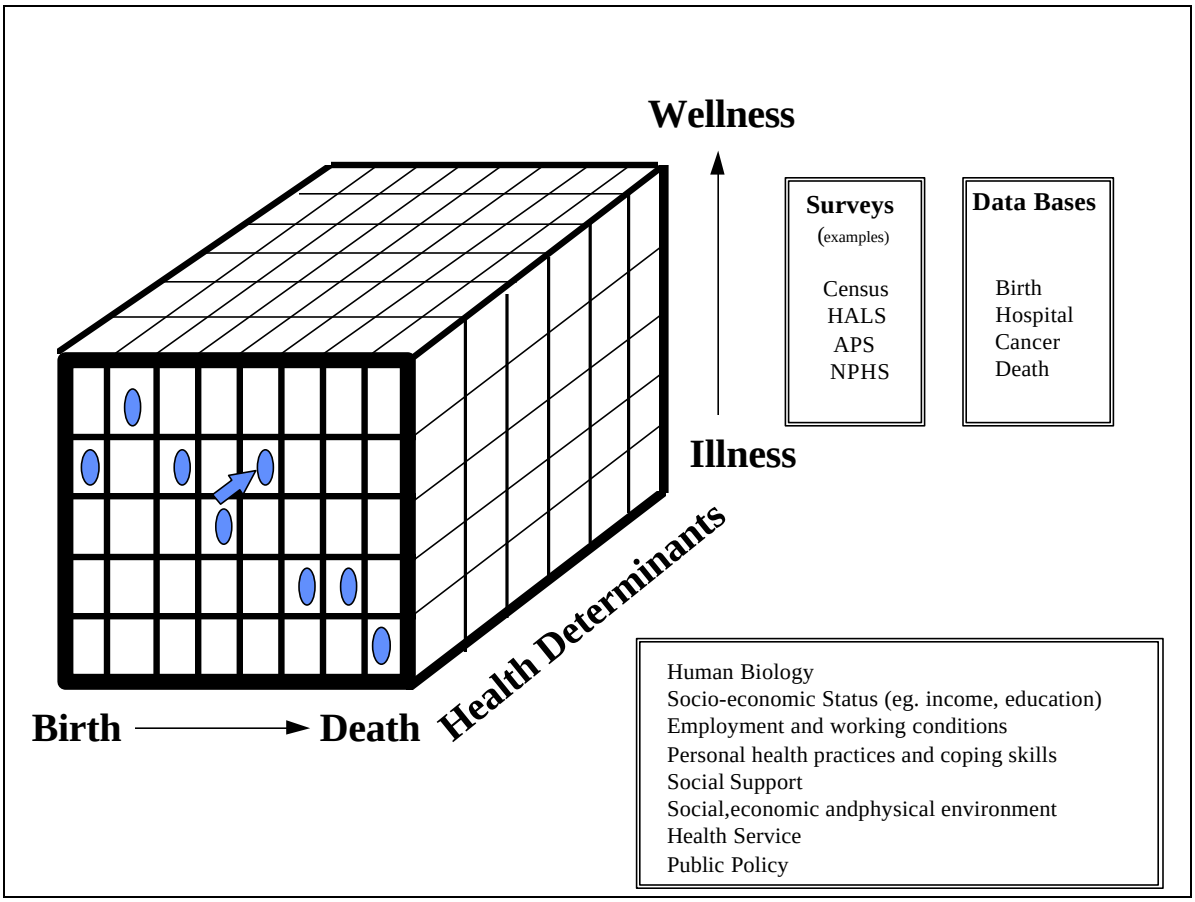
- Mortality, cancer and/or birth follow-up of
  - -- cohorts (e.g., miners, asbestos workers)
  - -- case/control studies
  - -- clinical trials (e.g., Canadian Breast Screening study);
- Building, maintaining and using registries (e.g., cancer and AIDS);
- Creation of patient-oriented histories;
- Follow-up of surveys (e.g., Nutrition Canada, Canada Health Survey, Fitness Canada);
- Occupational and environmental health studies;
- Examining factors which influence health care usage and costs; and
- Regional variations in the incidence of disease.

A longitudinal National Population Health Survey is currently in progress in Canada. In the original survey approximately 95% of the respondents agreed to have their survey data linked to their provincial health records. This linkage will strongly enhance the data set's potential usefulness because it will add respondents' interaction with the health care system.

Preparing for the Journey Ahead -- The Life Cycle of Events

In the health area, this usually involves the linkage of various data sources over time. Figure 2 is an example of how we can view the **life cycle** of events from birth to death, **health determinants**, and **outcomes in ranges of "illness" to "wellness."** Piecing together the various important components may involve gathering data from a number of different sources such as surveys to estimate the degree of "illness" or "wellness" of the population (e.g., Census, Health and Activity Limitation Survey (HALS), Aboriginal Peoples Survey (APS), National Population Health Survey (NPHS)), national databases of existing administrative records (e.g., Canadian Birth Data Base, the Canadian Cancer Data Base and the Canadian Mortality Data Base), and from a number of different perspectives. For example, within health determinants one may be interested in human biology, socio-economic status (e.g., income, education), employment and working conditions, personal health practices and coping skills, social support, social, economic and physical environment, health services, and public policy. As the population/individual progresses through the different stages of the life cycle, the degree of "wellness" can vary as indicated in the diagram. (See also Hertzman, Frank and Evans, 1994).

Figure 2.--The Life Cycle of Events



HALS -- Health and Activity Limitation Survey
APS -- Aboriginal Peoples Survey
NPHS -- National Population Health Survey

Some examples, involving the use of census data, are as follows:

- **Maternal Health and Infant Birth - Death Linkages**.--A study of regional differences in perinatal and infant mortality in the province of Ontario has been carried out. Infant and perinatal mortality in the 53 counties of Ontario were studied in two time periods -- 1970-72 and 1978-79. A considerable regional variation in the range of rates was found. Socio-economic factors were found to have an important influence on the maternal and infant determinants of mortality and in this way contributed to the variations in mortality over the province. Recently, there has been interest in establishing a Canadian Perinatal Health Surveillance system.

- **Occupational Studies.--**There are strong pressures from society to determine and reveal the health risks to which it is exposed, especially where the harm is cumulative or latent for an extended period of time. These pressures come from three main sources. **Organized labour** has a special interest in conditions in the workplace which might lead to delayed effects, such as cancers among its members. Both the **general public and environmental groups** have frequently expressed concern over the possible consequences of exposure of the population at large to chemical and other agents. These agents are being produced in increasing numbers and quantities, and distributed both as commercial products and as contaminant wastes in ways that may result in ingestion or inhalation. The third source of pressure originates with **professional groups** whose work involves them in the detection and measurement of health risks and in setting safety regulations. Cancer incidence and mortality data are a main source of information to assist in the determination of health risks.

  In light of urgent demands to protect workers' health, there is a need for a broad-based occupation-cancer database containing information on both cancer incidence and a wide range of occupations. A current feasibility study is examining the possibility of linking cancer, mortality and occupational, household and socio-economic data derived from the 1986 census data. The sample, consisting of seven geographic regions (4 urban and 3 rural), were selected based on census geography.

  As an occupational group, farmers have low overall mortality. However, a number of epidemiological studies suggest increased risk of certain cancers among farmers, including cancer of the stomach, lip, prostate, brain and skin, leukemia, Hodgkin's disease, multiple myeloma, and non-Hodgkin's disease.

  A mortality and cancer cohort study of about 326,000 Canadian male farm operators enumerated in the 1971 Census of Agriculture has been carried out in collaboration with Health Canada (Fair, 1993). Seven major files were linked to create the data required for the analysis file in this study, namely:

  -- the 1971 Census of Agriculture;
  -- the 1971 Census of Population;
  -- the 1971 Central Farm Register;
  -- the 1981 Central farm Register;
  -- the Canadian Mortality Data Base;
  -- the 1966-71-76-81-86 Census of Agriculture Longitudinal file; and
  -- the Canadian Cancer Data Base.

Analyses of these data have examined prostate (Morrison et al., 1993) and brain cancer (Morrison et al., 1992) in particular.

- **Socio-Economic Gradients**.--There has been an increasing awareness of the importance of supporting basic research designed to identify determinants of health in order to inform policy makers about how best to improve the population's heath and how best to accomplish this goal efficiently and cost-effectively. As a result, the Manitoba Centre for Health Policy and Evaluation has collaborated with Statistics Canada to determine the feasibility of linking provincial administrative health care utilization with census data for a sample of Manitobans (Mustard et al., 1995).

  Mortality and health services utilization have been described in relation to the socio-economic status measure, mortality and the use of health care services at seven different stages in the life course (ages 0-4, 5-14, 15-29, 30-49, 50-64, 65-74, 75+). The objective of the study was to identify those classes of morbidity which dominate utilization of health care services at each stage of life course and simultaneously, the classes of morbidity which show the greatest disparities in relation to socio-economic status. The research resource of this project was created at a fraction of the cost of a population survey. Some of the public policy responses indicated by these data were:

  - to consider directing an even greater share of health care services to lower socio-economic groups;
  - to more aggressively target preventive medical and health services, especially in early adulthood; and
  - to formulate explicit public policies addressing health inequalities. (Mustard et al., 1995, p. 67).

Making the Right Connections and Summary

We are in a time of rapid **changes** in terms of **markets**, **customer expectations** and technologies for record linkage **software** development, **hardware**, and **applications**. There often needs to be an optimal balance between cost, quality and timeliness. Many of the existing **data systems** are on the threshold of change. There is a shift from single data base applications to electronic data transfer and warehousing, data sharing within broad subject matter areas, and to enterprise wide systems and **data integration**. There are various hardware and software environments being used. A variety of approaches can be used to assess user's needs. These include professional advisory committees, client-oriented program evaluations, interactions with professional and other associations, market feedback, and analytic programs.

One needs to have the capacity to acquire, generate, distribute and apply knowledge strategically and operationally (Toffler and Toffler, 1995). To a large extent the quality of record linkage in the future is dependent on the **quality of the files** being linked -- quality in/quality out. There is a need to harmonize concepts and outputs. For example, it is anticipated that the Tenth International Revision of the Classification of Disease will be implemented. A restructured industry classification system known as the North American Industry Classification System is being developed. **Uniform lifetime** business and individual **numbers** are highly desirable for many of the new information systems. Further work is required in designing appropriate items for the data sets -- for example, more detail may be available at the local level than on a national basis.

There is a need to **integrate** a number of different sources of data. As governments and agencies regionalize services, there are additional requests for **small area data**. It is important to have the capacity to use multiple definitions of geographic population areas of interest (e.g., enumeration area, postal code areas, school districts, health units) depending on the nature of the investigation.

There is a need to develop **confidentiality procedures** and screening rules for the generation and release of public use data files. All studies involving record linkage at Statistics Canada must satisfy a prescribed review and approval process. For example, the purpose of the record linkage activity must be statistical or research in nature and must be consistent with the mandate of Statistics Canada as described in the Statistics Act. The record linkage activity must have demonstrable cost or respondent burden savings over

other alternatives, or be the only feasible option. It must also be shown to be in the public interest. A comprehensive list of recommendations for Federal statistics agencies in the United States is given in Duncan et al., 1993.

In conclusion, **analysis** of existing and future linked data sets is indispensable in illuminating the main social and economic issues we face not only today, but also into the future. We need to anticipate and look forward to issues of the 21st century where record linkage may serve as an important research tool.

References

Bank of Montreal (1995). *Bank of Montreal 178th Annual Report 1995,* Public Affairs Department of the Bank, Bank of Montreal Tower, 55 Bloor Street West, 4th Floor, Toronto, Ontario M4W 3N5.

Barna, G. (1990). *The Frog in the Kettle: What Christians Need to Know About Life in the Year 2000*, Regal Books, Ventura, California 93006.

Blomquist, A. and Brown D. M. (Eds.) (1994). *Limits to Care: Reforming Canada's Health Care System in an Age of Restraint.* Available from: Renouf Publishing Company Limited, 1294 Algoma Road, Ottawa, Ontario K1B 3W8.

Canadian Institute of Child Health (1994). *The Health of Canada's Children - A CICH Profile 2nd Profile.* Available from: Canadian Institute of Child Health, 885 Meadowlands Drive, Suite 512 Ottawa, Ontario K2C 3N2.

Carpenter, M., and Fair, M.E. (Eds.) (1989). *Canadian Epidemiology Research Conference -- Proceedings of the Record Linkage Session and Workshop.* Available from: Statistics Canada, Occupational and Environmental Health Research Section, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

Chad, R. (1993). *A Comparison of Three Different Computer Matches*. Special Census/Administrative Record Match Working Group in Conjunction with the Year 2000 Researcher Development Staff, U.S. Bureau of the Census, Washington, DC, September 1993, (Matchers -- Winkler, Slaven, Jaro).

Dale, D. (1989). Linkage As Part of a Production System. The Ontario Cancer Registry, in *Canadian Epidemiology Research Conference -- Proceedings of the Record Linkage Sessions and Workshop*, M. Carpenter and M.E. Fair, Eds. Available from: Statistics Canada, Occupational and Environmental Health Research Section, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

Donaldson, M.S., and Lohr, K.N. (Eds.) (1994). *Health Data in the Information Age -- Use, Disclosure, and Privacy*, Washington, D. C.: National Academy Press.

Doucet, E. (1995). Survey Re-Engineering: Is Our Information Technology Framework Up to It? *Proceedings of Statistics Canada Symposium 94 -- Re-Engineering for Statistical Agencies*, November 1994, Available from: Financial Operations Division, Statistics Canada, R.H. Coats Building, 6th Floor, Tunney's Pasture, Ottawa, Ontario K1A 0T6, P. 159-168.

Duncan, G.T.; Jabine, T.B.; and De Wolf, V.A. (Eds.) (1993). *Private Lives and Public Policies -- Confidentiality and Accessibility of Government Statistics*, Washington, D.C.: National Academy Press.

Dunn, H.L. (1946). Record Linkage, *American Journal of Public Health*, 36, 1412-1416.

Fair, M.E. (1995). An Overview of Record Linkage in Canada, *1995 Proceedings of the Social Statistics Section of the American Statistical Association*, American Statistical Association, 1429 Duke Street, Alexandria, Virginia 22314-3402, 25-33.

Fair, M.E. (1993). Recent Advances in Matching and Record Linkage from a Study of Canadian Farm Operators and Their Farming Practices, *1993 ICES Proceedings of the International Conference of Establishment Surveys*, American Statistical Association,1429 Duke Street, Alexandria, Virginia 22314-3402, 600-605.

Fellegi, I.P. (1995). Characteristics of an Effective Statistical System, Morris Hansen Lecture, presented at the Washington Statistical Society, October 25, 1995.

Fellegi, I.P. and Sunter, A.B. (1969). A Theory of Record Linkage, *Journal of the American Statistical Association*, 40, 1183-1210.

George, M.V.; Norris, M.J.; Nault, F.; Loh, S.; and Dai, S.Y. (1994). *Population, Projections for Canada, Provinces and Territories 1993-2016*, Statistics Canada, Demography Division, Catalogue No. 91-520. Available from: Marketing Division, Sales and Services, Statistics Canada, Ottawa, K1A 0T6.

Gill, L.; Goldacre, M.; Simmons, H.; Bettley, G.; and Griffith, M. (1993). Computerized Linking of Medical Records: Methodological Guidelines, *Journal of Epidemiology and Comm. Health*, 47:4, 316-319.

Gosselin, J. F. (1995). The Operational Framework at Statistics Canada, *Proceedings of Statistics Canada Symposium '94 -- Re-Engineering for Statistical Agencies*, November 1994. Available from: Financial Operations Division, R.H. Coats Building, 6th Floor, Tunney's Pasture, Ottawa, Ontario K1A 0T6, 170- 174.

Grant, Linda (of *The Guardian*) (1996). Riding the Wave, *The Ottawa Citizen*, January 20, 1996, B4.

Haberman, H. (1995). Towards a Global Statistical System, *Proceedings of Statistics Canada Symposium 94 -- Re-Engineering for Statistical Agencies*, November 1994. Available from: Financial Operations Division, R.H. Coats Building, 6th Floor, Tunney's Pasture, Ottawa, Ontario K1A 0T6, 53-60.

Hertzman, C.; Frank, J.; and Evans, R.G. (1994). Heterogeneitics in Health Status and the Determinants of Population Health, *Why Are Some People Healthy and Others Not? The Determinants of Health of Populations*, R.G. Evans; L. Barer; and T.M. Marmor, Eds. New York: Aldine De Gruyter, 74f.

Hill, T. (1981). Generalized Iterative Record Linkage System, Ottawa, Canada: Statistics Canada.

Jaro, M.A. (1995). Probabilistic Linkage of Large Public Health Data Files, *Statistics in Medicine*, 14, 491-498.

Kilss, B., and Alvey, W. (Eds.) (1985). *Record Linkage Techniques -- 1985. Proceedings of the Workshop on Exact Matching Methodologies*, Arlington, Virginia, May 9-10, 1985, Washington, DC: Department of Treasury, Internal Revenue Service.

Morrison, H.; Savitz, D.; Semenciw, R.; Hulka, B.; Mao, Y.; Morison, D.; and Wigle, D. (1993), Farming and Prostate Cancer Mortality, *American Journal of Epidemiology*, 137, 270-280.

Morrison, H.I.; Semenciw, R.M.; Morison, D.; Magwood, S.; and Mao, Y. (1992). Brain Cancer and Farming in Western Canada, *Neuroepidemiology*, 11, 267-276.

Mustard, C.; Derksen, S.; Berthelot, J.M.; Wolfson, M.; Roos, L.L.; and Carriere, K.S. (1995). *Socio-economic Gradients in Mortality and the Use of Health Care Services at Different Stages in the Life Course*, Manitoba Centre for Health Policy and Evaluation, Department of Community Health Sciences, Faculty of Medicine, University of Manitoba.

Newcombe, H.B. (1994). Cohorts and Privacy, *Cancer Causes and Control*, 5, 287-292.

Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Social Studies, Administration and Business*, Oxford, U.K.: Oxford University Press.

Newcombe, H. B.; Kennedy, J. M.; Axford, S. J.; and James, A. P. (1959). Automatic Linkage of Vital Records, *Science*, 130, 3381, 954-959.

Starr, P. and Starr, S. (1995). Reinventing Vital Statistics, The Impact of Changes in Information Technology, Welfare Policy and Health Care, *Public Health Reports*, 110, 535-544.

Statistics Canada (1996). *Generalized Record Linkage System Concepts*, Draft version dated 1996 February 14, Research and General Systems Development Division, Ottawa, K1A 0T6.

Statistics Canada (1994). *Symposium '94 Re-engineering for Statistical Agencies*, Catalogue No. 11-522E, Occasional -- November 1994. Available from: Marketing Division, Sales and Service, Statistics Canada, Ottawa, K1A 0T6.

Toffler A. and Toffler H. (1995). *Creating A New Generation -- The Politics of the Third Wave*, Atlanta: Turner Publishing, Inc.

United Nations (1993). *Sources and Effects of Ionizing Radiation -- United Nations Scientific Committee on the Effects of Atomic Radiation*, New York: United Nations Publication, United Nations.

United Nations Department of International Economic and Social Affairs (UNDIESA) (1991). *The Sex and Age Distribution of Population*, ST/ESA/ SER. A/122, New York.

Winkler, W. and Scheuren, F. (1995). Linking Data to Create Information, *Proceedings of Statistics Canada Symposium '95 -- From Data to Information -- Methods and Systems*, November 1995, Statistics Canada, Ottawa K1A 0T6 (in press).

Chapter 11

Computational Disclosure Control for Medical Microdata: The Datafly System

Latanya Sweeney, Massachusetts Institute of Technology

Abstract

We present a computer program named Datafly that uses computational disclosure techniques to maintain anonymity in medical data by automatically generalizing, substituting and removing information as appropriate without losing many of the details found within the data. Decisions are made at the field and record level at the time of database access, so the approach can be used on the fly in role-based security within an institution, and in batch mode for exporting data from an institution. Often organizations release and receive medical data with all explicit identifiers, such as name, address, phone number, and social security number, removed in the incorrect belief that patient confidentiality is maintained because the resulting data look anonymous; however, we show that in most of these cases, the remaining data can be used to re-identify individuals by linking or matching the data to other databases or by looking at unique characteristics found in the fields and records of the database itself. When these less apparent aspects are taken into account, each released record can be made to ambiguously map to many possible people, providing a level of anonymity which the user determines.

Introduction

Sharing and disseminating electronic medical records while maintaining a commitment to patient confidentiality is one of the biggest challenges facing medical informatics and society at large. To the public, patient confidentiality implies that only people directly involved in their care will have access to their medical records and that these people will be bound by strict ethical and legal standards that prohibit further disclosure (Woodward, 1996). The public is not likely to accept that their records are kept "confidential" if large numbers of people have access to their contents.

On the other hand, analysis of the detailed information contained within electronic medical records promises many advantages to society, including improvements in medical care, reduced institution costs, the development of predictive and diagnostic support systems, and the integration of applicable data from multiple sources into a unified display for clinicians; but these benefits require sharing the contents of medical records with secondary viewers, such as researchers, economists, statisticians, administrators, consultants, and computer scientists, to name a few. The public would probably agree these secondary parties should know some of the information buried in the record, but such disclosure should not risk identifying patients.

In 1996, the National Association of Health Data Organizations (NAHDO) reported that 37 states had legislative mandates to gather hospital-level data. Last year, 17 of these states reported they had started collecting ambulatory care (outpatient) data from hospitals, physician offices, clinics, and so on. Table 1 contains a list of the fields of information which NAHDO recommends these states accumulate. Many of these states have subsequently given copies of collected data to researchers and sold copies to industry. Since the information has no explicit identifiers, such as name, address, phone number or social security number, confidentiality is incorrectly believed to be maintained.

Table 1. -- Data Fields Recommended by NAHDO for State Collection of Ambulatory Data

Patient Number
Patient ZIP Code
Patient Racial Background
Patient Birth Date
Patient Gender
Visit Date
Principal Diagnosis Code (ICD9)
Procedure Codes (up to 14)
Physician ID#
Physician ZIP code
Total Charges

In fairness, there are many sources of administrative billing records with fields of information similar to those listed in Table 1. Hospital administrators often pass medical records along in part to independent consultants and outside agencies. There are the records maintained by the insurance companies. Pharmaceutical companies run longitudinal studies on identified patients and providers. Local drug stores maintain individualized prescription records. The list is quite extensive. Clearly, we see the possible benefits from sharing information found within the medical record and within records of secondary sources; but on the other hand, we appreciate the need for doctor-patient confidentiality. The goal of this work is to provide tools for extracting needed information from medical records while maintaining a commitment to patient confidentiality. These same techniques are equally applicable to financial, demographic and educational microdata releases, as well.

Background

We begin by first stating our definitions of de-identified and anonymous data. In de-identified data, all explicit identifiers, such as social security number, name, address and phone number, are removed, generalized or replaced with a made-up alternative. De-identifying data does not guarantee that the result is anonymous however. The term anonymous implies that the data cannot be manipulated or linked to identify any individual. Even when information shared with secondary parties is de-identified, we will show it is often far from anonymous.

There are three major difficulties in providing anonymous data. One of the problems is that anonymity is in the eye of the beholder. For example, consider Table 2. If the contents of this table are a subset of an extremely large and diverse database then the three records listed in Table 2 may appear anonymous. Suppose the ZIP code 33171 primarily consists of a retirement community; then there are very few people of such a young age living there. Likewise, 02657 is the ZIP code for Provincetown, Massachusetts, in which we found about 5 black women living there year-round. The ZIP code 20612 may have only one Asian family. In these cases, information outside the data identifies the individuals.

Table 2. -- De-identified Data that Are not Anonymous

ZIP Code	Birthdate	Gender	Ethnicity
33171	7/15/71	m	Caucasian
02657	2/18/73	f	Black
20612	3/12/75	m	Asian

Most towns and cities sell locally collected census data or voter registration lists that include the date of birth, name and address of each resident. This information can be linked to medical microdata that include a date of birth and ZIP code, even if the names, social security numbers and addresses of the patients are not present. Of course, local census data are usually not very accurate in college towns and areas that have a large transient community, but for much of the adult population in the United States, local census information can be used to re-identify de-identified microdata since other personal characteristics, such as gender, date of birth, and ZIP code, often combine uniquely to identify individuals.

The 1997 voting list for Cambridge, Massachusetts contains demographics on 54,805 voters. Of these, birth date alone can uniquely identify the name and address of 12% of the voters. We can identify 29% by just birth date and gender, 69% with only a birth date and a 5-digit ZIP code, and 97% (53,033 voters) when the full postal code and birth date are used. These values are listed in Table 3. Clearly, the risks of re-identifying data depend both on the content of the released data and on related information available to the recipient.

Table 3. -- Uniqueness of Demographic Fields in Cambridge Voter List

Birth date alone	12%
birth date and gender	29%
birth date and 5-digit ZIP	69%
birth date and full postal code	97%

A second problem with producing anonymous data concerns unique and unusual information appearing within the data themselves. Instances of uniquely occurring characteristics found within the original data can be used by reporters, private investigators and others to discredit the anonymity of the released data even when these instances are not unique in the general population. Also, unusual cases are often unusual in other sources of data as well making them easier to identify. Consider the database shown in Table 4. It is not surprising that the social security number is uniquely identifying, or given the size of the database, that the birth date is also unique. To a lesser degree the ZIP codes in Table 4 identify individuals since they are almost unique for each record. Importantly, what may not have been known without close examination of the particulars of this database is that the designation of Asian as a race is uniquely identifying. During an interview, we could imagine that the janitor, for example, might recall an Asian patient whose last name was Chan and who worked as a stockbroker for ABC Investment since the patient had given the janitor some good investing tips.

Table 4. -- Sample Database in which Asian is

A Uniquely Identifying Characteristic

SSN	Ethnicity	Birth	Sex	ZIP
819491049	Caucasian	10/23/64	m	02138
749201844	Caucasian	03/15/65	m	02139
819181496	Black	09/20/65	m	02141
859205893	Asian	10/23/65	m	02157
985820581	Black	08/24/64	m	02138

Any single uniquely occurring value or group of values can be used to identify an individual. Consider the medical records of a pediatric hospital in which only one patient is older than 45 years of age. Or, suppose a hospital's maternity records contained only one patient who gave birth to triplets. Knowledge of the uniqueness of this patient's record may appear in many places including insurance claims, personal financial records, local census information, and insurance enrollment forms. Remember that the unique characteristic may be based on diagnosis, treatment, birth year, visit date, or some other little detail or combination of details available to the memory of a patient or a doctor, or knowledge about the database from some other source.

Measuring the degree of anonymity in released data poses a third problem when producing anonymous data for practical use. The Social Security Administration (SSA) releases public-use files based on national samples with small sampling fractions (usually less than 1 in 1,000); the files contain no geographic codes, or at most regional or size of place designators (Alexander et al., 1978). The SSA recognizes that data containing individuals with unique combinations of characteristics can be linked or matched with other data sources. So, the SSA's general rule is that any subset of the data that can be defined in terms of combinations of characteristics must contain at least 5 individuals. This notion of a minimal bin size, which reflects the smallest number of individuals matching the characteristics, is quite useful in providing a degree of anonymity within data. The larger the bin size, the more anonymous the data. As the bin size increases, the number of people to whom a record may refer also increases, thereby masking the identity of the actual person.

In medical databases, the minimum bin size should be much larger than the SSA guidelines suggest. Consider these three reasons: most medical databases are geographically located and so one can presume, for example, the ZIP codes of a hospital's patients; the fields in a medical database provide a tremendous amount of detail and any field can be a candidate for linking to other databases in an attempt to re-identify patients; and, most releases of medical data are not randomly sampled with small sampling fractions, but instead include most if not all of the database.

Determining the optimal bin size to ensure anonymity is tricky. It certainly depends on the frequencies of characteristics found within the data as well as within other sources for re-identification. In addition, the motivation and effort required to re-identify released data in cases where virtually all possible candidates can be identified must be considered. For example, if we release data that maps each record to 10 possible people and the 10 people can be identified, then all 10 candidates may even be contacted or visited in an effort to locate the actual person. Likewise, if the mapping is 1 in 100, all 100 could be phoned since visits may then be impractical, and in a mapping of 1 in 1000, a direct mail campaign could be employed. The amount of effort the recipient is willing to spend depends on their motivation. Some medical files are quite valuable, and valuable data will merit more effort. In these cases, the minimum bin size must be further increased or the sampling fraction reduced to render these efforts useless.

Of course, the expression of anonymity most semantically consistent with our intention is simply the probability of identifying a person given the released data and other possible sources. This conditional probability depends on frequencies of characteristics (bin sizes) found within the data and the outside world. Unfortu-

nately, this probability is very difficult to compute without omniscience. In extremely large databases like that of SSA, the database itself can be used to compute frequencies of characteristics found in the general population since it contains almost all the general population; small, specialized databases, however, must estimate these values. In the next section, we will present a computer program that generalizes data based on bin sizes and estimates. Following that, we will report results using the program and discuss its limitations.

Methods

Earlier this year, Sweeney presented the Datafly System (1997) whose goal is to provide the most general information useful to the recipient. Datafly maintains anonymity in medical data by automatically aggregating, substituting and removing information as appropriate. Decisions are made at the field and record level at the time of database access, so the approach can be incorporated into role-based security within an institution as well as in exporting schemes for data leaving an institution. The end result is a subset of the original database that provides minimal linking and matching of data since each record matches as many people as the user had specified.

Diagram 1 provides a user-level overview of the Datafly System. The original database is shown on the left. A user requests specific fields and records, provides a profile of the person who is to receive the data, and requests a minimum level of anonymity. Datafly produces a resulting database whose information matches the anonymity level set by the user with respect to the recipient profile. Notice how the record containing the Asian entry was removed; social security numbers were automatically replaced with made-up alternatives; and birth dates were generalized to the year, and ZIP codes to the first three digits. In the next three paragraphs we examine the overall anonymity level and the profile of the recipient, both of which the user provides when requesting data.

Diagram 1. -- The Input to the Datafly System is the Original Database and Some User Specifications, and the Output is a Database Whose Fields and Records Correspond to the Anonymity Level Specified by the User, in this Example, 0.7.

User
-fields & records
-recipient profile
-anonymity 0.7

Original Medical Database

SSN	Race	Birth	Sex	ZIP
819491049	Caucasian	10/23/64	m	02138
749201844	Caucasian	03/15/65	m	02139
819181496	Black	09/20/65	m	02141
859205893	Asian	10/23/65	m	02157
985820581	Black	08/24/64	m	02138

Datafly

Resulting Database, anonymity 0.7

SSN	Race	Birth	Sex	ZIP
444444444	Caucasian	1964	m	02100
555555555	Caucasian	1965	m	02100
333333333	Black	1965	m	02100
222222222	Black	1964	m	02100

The overall anonymity level is a number between 0 and 1 that specifies the minimum bin size for every field. An anonymity level of 0 provides the original data, and a level of 1 forces Datafly to produce the most general data possible given the profile of the recipient. All other values of the overall anonymity level between 0 and 1 determine the minimum bin size *b* for each field. (The institution is responsible for mapping the ano-

nymity level to actual bin sizes though Sweeney provides some guidelines.) Information within each field is generalized as needed to attain the minimum bin size; outliers, which are extreme values not typical of the rest of the data, may be removed. When we examine the resulting data, every value in each field will occur at least *b* times with the exception of one-to-one replacement values, as is the case with social security numbers.

Table 5 shows the relationship between bin sizes and selected anonymity levels using the Cambridge voters database. As A increased, the minimum bin size increased, and in order to achieve the minimal bin size requirement, values within the birth date field, for example, were re-coded as shown. Outliers were excluded from the released data and their corresponding percentages of N are noted. An anonymity level of 0.7, for example, required at least 383 occurrences of every value in each field. To accomplish this in the birth date field, dates were re-coded to reflect only the birth year. Even after generalizing over a 12 month window, the values of 8% of the voters still did not meet the requirement so these voters were dropped from the released data.

Table 5. -- Anonymity Generalizations for Cambridge Voters Data with Corresponding Bin Sizes *

Anonymity	BinSize	BirthDate	Drop%
1.0			
.9	493	24	4%
.8	438	24	2%
.7	383	12	8%
.6	328	12	5%
.5	274	12	4%
.4	219	12	3%
.3	164	6	5%
.2	109	4	5%
.1	54	2	5%
0.0			

* The birth date generalizations (in months) required to satisfy the minimum bin size are shown and the percentages of the total database dropped due to outliers is displayed. The user sets the anonymity level as depicted above by the slide bar at the 0.7 selection. The mappings of anonymity levels to bin sizes is determined by the institution.

In addition to an overall anonymity level, the user also provides a profile of the person who receives the data by specifying for each field in the database whether the recipient could have or would use information external to the database that includes data within that field. That is, the user estimates on which fields the recipient might link outside knowledge. Thus each field has associated with it a profile value between 0 and 1, where 0 represents full trust of the recipient or no concern over the sensitivity of the information within the field, and 1 represents full distrust of the recipient or maximum concern over the sensitivity of the field's contents. The role of these profile values is to restore the effective bin size by forcing these fields to adhere to bin sizes larger than the overall anonymity level warranted. Semantically related sensitive fields, with the exception of one-to-one replacement fields, are treated as a single concatenated field which must meet the minimum bin size, thereby thwarting linking attempts that use combinations of fields.

Consider the profiles of a doctor caring for a patient, a clinical researcher studying risk factors for heart disease and a health economist assessing the admitting patterns of physicians. Clearly, these profiles are all different. Their selection and specificity of fields are different; their sources of outside information on which they could link are different; and, their uses for the data are different. From publicly available birth certificate,

driver license, and local census databases, the birth dates, ZIP codes and gender of individuals are commonly available along with their corresponding names and addresses; so these fields could easily be used for re-identification. Depending on the recipient, other fields may be even more useful, but we will limit our example to profiling these fields. If the recipient is the patient's caretaker within the institution, the patient has agreed to release this information to the care-taker, so the profile for these fields should be set to 0 to give the patient's caretaker full access to the original information. When researchers and administrators make requests that do not require the most specific form of the information as found originally within sensitive fields, the corresponding profile values for these fields should warrant a number as close to 1 as possible but not so much so that the resulting generalizations do not provide useful data to the recipient. But researchers or administrators bound by contractual and legal constraints that prohibit their linking of the data are trusted, so if they make a request that includes sensitive fields, the profile values would ensure that each sensitive field adheres only to the minimum bin size requirement. The goal is to provide the most general data that are acceptably specific to the recipient. Since the profile values are set independently for each field, particular fields that are important to the recipient can result in smaller bin sizes than other requested fields in an attempt to limit generalizing the data in those fields; a profile for data being released for public use, however, should be 1 for all sensitive fields to ensure maximum protection. The purpose of the profile is to quantify the specificity required in each field and to identify fields that are candidates for linking; and in so doing, the profile identifies the associated risk to patient confidentiality for each release of data.

Results

Numerous tests were conducted using the Datafly System to access a pediatric medical record system (Sweeney, 1997). Datafly processed all queries to the database over a spectrum of recipient profiles and anonymity levels to show that all fields in medical records can be meaningfully generalized as needed since any field can be a candidate for linking. Of course, which fields are most important to protect depends on the recipient. Diagnosis codes have generalizations using the International Classification of Disease (ICD-9) hierarchy. Geographic replacements for states or ZIP codes generalize to use regions and population size. Continuous variables, such as dollar amounts and clinical measurements, can be treated as categorical values; however, their replacements must be based on meaningful ranges in which to classify the values; of course this is only done in cases where generalizing these fields is necessary.

The Group Insurance Commission in Massachusetts (GIC) is responsible for purchasing insurance for state employees. They collected encounter-level de-identified data with more than 100 fields of information per encounter, including the fields in Table 1 for approximately 135,000 patients consisting of state employees and their families (Lasalandra, 1997). In a public hearing, GIC reported giving a copy of the data to a researcher, who in turn stated she did not need the full date of birth, just the birth year. The average bin size based only on birth date and gender for that population is 3, but had the researcher received only the year of birth in the birth date field, the average bin size based on birth year and gender would have increased to 1125 people. It is estimated that most of this data could be re-identified since collected fields also included residential ZIP codes and city, occupational department or agency, and provider information. Furnishing the most general information the recipient can use minimizes unnecessary risk to patient confidentiality.

Comparison to μ-ARGUS

In 1996, The European Union began funding an effort that involves statistical offices and universities from the Netherlands, Italy and the United Kingdom. The main objective of this project is to develop specialized software for disclosing public-use data such that the identity of any individual contained in the released data cannot be recognized. Statistics Netherlands has already produced, but has not yet released, a first version of a program named μ-Argus that seeks to accomplish this goal (Hundepool, et al., 1996). The μ-Argus program is considered by many as the official confidentiality software of the European community even though Statistics Netherlands admittedly considers this first version a rough draft. A presentation of the concepts on which μ-Argus is based can be found in Willenborg and De Waal (1996).

The program μ-Argus, like the Datafly System, makes decisions based on bin sizes, generalizes values within fields as needed, and removes extreme outlier information from the released data. The user provides an overall bin size and specifies which fields are sensitive by assigning a value between 0 and 3 to each field. The program then identifies rare and therefore unsafe combinations by testing 2- or 3-combinations across the fields noted by the user as being identifying. Unsafe combinations are eliminated by generalizing fields within the combination and by local cell suppression. Rather than removing entire records when one or more fields contain outlier information, as is done in the Datafly System, the μ-Argus System simply suppresses or blanks out the outlier values at the cell-level. The resulting data typically contain all the rows and columns of the original data though values may be missing in some cell locations.

In Table 6a there are many Caucasians and many females, but only one female Caucasian in the database. Tables 6b and 6c show the resulting databases when the Datafly System and the μ-Argus System were applied to this data. We will now step through how the μ -Argus program produced the results in Table 6c.

Table 6a. -- There is Only One Caucasian Female, Even Though There are Many Females and Caucasians

SSN	Ethnicity	Birth	Sex	ZIP	Problem
819181496	Black	09/20/65	m	02141	shortness of breath
195925972	Black	02/14/65	m	02141	chest pain
902750852	Black	10/23/65	f	02138	hypertension
985820581	Black	08/24/65	f	02138	hypertension
209559459	Black	11/07/64	f	02138	obesity
679392975	Black	12/01/64	f	02138	chest pain
819491049	Caucasian	10/23/64	m	02138	chest pain
749201844	Caucasian	03/15/65	f	02139	hypertension
985302952	Caucasian	08/13/64	m	02139	obesity
874593560	Caucasian	05/05/64	m	02139	shortness of breath
703872052	Caucasian	02/13/67	m	02138	chest pain
963963603	Caucasian	03/21/67	m	02138	chest pain

Table 6b. -- Results from Applying the Datafly System to the

Data in Table 6a *

SSN	Ethnicity	Birth	Sex	ZIP	Problem
902387250	Black	1965	m	02140	shortness of breath
197150725	Black	1965	m	02140	chest pain
486062381	Black	1965	f	02130	hypertension
235978021	Black	1965	f	02130	hypertension
214684616	Black	1964	f	02130	obesity
135434342	Black	1964	f	02130	chest pain
458762056	Caucasian	1964	m	02130	chest pain
860424429	Caucasian	1964	m	02130	obesity
259003630	Caucasian	1964	m	02130	shortness of breath
410968224	Caucasian	1967	m	02130	chest pain
664545451	Caucasian	1967	m	02130	chest pain

*The minimum bin size is 2. The given profile identifies only the demographic fields as being likely for linking. The data are being made available for semi-public use so the Caucasian female record was dropped as an outlier.

Table 6c. -- Results from Applying the Approach of the m-Argus System to the Data in Table 6a*

SSN	Ethnicity	Birth	Sex	ZIP	Problem
	Black	1965	m	02141	shortness of breath
	Black	1965	m	02141	chest pain
	Black	1965	f	02138	hypertension
	Black	1965	f	02138	hypertension
	Black	1964	f	02138	obesity
	Black	1964	f	02138	chest pain
	Caucasian	1964	m	02138	chest pain
			f	*02139*	hypertension
	Caucasian	1964	m	02139	obesity
	Caucasian	1964	m	02139	shortness of breath
	Caucasian	1967	m	02138	chest pain
	Caucasian	1967	m	02138	chest pain

*The minimum bin size is 2. SSN was marked as being most identifying, the birth, sex, and ZIP fields were marked as being more identifying, and the ethnicity field was simply marked as identifying. Combinations across these were examined; the resulting suppressions are shown. The uniqueness of the Caucasian female is suppressed; but, there still remains a unique record for the Caucasian male born in 1964 that lives in the 02138 ZIP code.

The first step is to check that each identifying field adheres to the minimum bin size. Then, pairwise

combinations are examined for each pair that contains the "most identifying" field (in this case, SSN) and those that contain the "more identifying" fields (in this case, birth date, sex and ZIP). Finally, 3-combinations are examined that include the "most" and "more" identifying fields. Obviously, there are many possible ways to rate these identifying fields, and unfortunately different identification ratings yield different results. The ratings presented in this example produced the most secure result using the μ-Argus program though admittedly one may argue that too many specifics remain in the data for it to be released for public use.

The value of each combination is basically a bin, and the bins with occurrences less than the minimum required bin size are considered unique and termed outliers. Clearly for all combinations that include the SSN, all such combinations are unique. One value of each outlier combination must be suppressed. For optimal results, the μ-Argus program suppresses values which occur in multiple outliers where precedence is given to the value occurring most often. The final result is shown in Table 6c. The responsibility of when to generalize and when to suppress lies with the user. For this reason, the μ-Argus program operates in an interactive mode so the user can see the effect of generalizing and can then select to undo the step.

We will briefly compare the results of these two systems, but for a more in-depth discussion, see Sweeney (1997). The μ-Argus program checks at most 2- or 3-combinations of identifying fields, but not all 2- or 3-combinations are necessarily tested. Even if they were, there may exist unique combinations across 4 or more fields that would not be detected. For example, Table 6c still contains a unique record for a Caucasian male born in 1964 that lives in the 02138 ZIP code, since there are 4 characteristics that combine to make this record unique, not 2. Treating a subset of identifying fields as a single field that must adhere to the minimum bin size, as done in the Datafly System, appears to provide more secure releases of microdata.

Discussion

The Datafly and μ-Argus systems illustrate that medical information can be generalized so that fields and combinations of fields adhere to a minimal bin size, and by so doing, confidentiality can be maintained. Using such schemes we can even provide anonymous data for public use. There are two drawbacks to these systems but these shortcomings may be counteracted by policy.

One concern with both μ-Argus and Datafly is the determination of the proper bin size and its corresponding measure of disclosure risk. There is no standard which can be applied to assure that the final results are adequate. What is customary is to measure risk against a specific compromising technique, such as linking to known databases, that we assume the recipient is using. Several researchers have proposed mathematical measures of the risk which compute the conditional probability of the linker's success (Duncan, et al., 1987).

A policy could be mandated that would require the producer of data released for public use to guarantee with a high degree of confidence that no individual within the data can be identified using demographic or semi-public information. Of course, guaranteeing anonymity in data requires a criterion against which to check resulting data and to locate sensitive values. If this is based only on the database itself, the minimum bin sizes and sampling fractions may be far from optimal and may not reflect the general population. Researchers have developed and tested several methods for estimating the percentage of unique values in the general population based on a smaller database (Skinner, et al., 1992). These methods are based on subsampling techniques and equivalence class structure. In the absence of these techniques, uniqueness in the population based on demographic fields can be determined using population registers that include patients from the database, such as local census data, voter registration lists, city directories, as well as information from motor vehicle agencies, tax assessors and real estate agencies. To produce an anonymous database, a producer could use population registers to identify sensitive demographic values within a database, and thereby obtain a measure of risk for the release of the data.

The second drawback with the µ-Argus and Datafly systems concerns the dichotomy between researcher needs and disclosure risk. If data are explicitly identifiable, the public would expect patient consent to be required. If data are released for public use, then the producer should guarantee, with a high degree of confidence, that the identity of any individual cannot be determined using standard and predictable methods and reasonably available data. But when sensitive de-identified, but not necessarily anonymous, data are to be released, the likelihood that an effort will be made to re-identify an individual increases based on the needs of the recipient, so any such recipient has a trust relationship with society and the producer of the data. The recipient should therefore be held accountable.

The Datafly and µ-Argus systems quantify this trust by profiling the fields requested by the recipient. But recall that profiling requires guesswork in identifying fields on which the recipient could link. Suppose a profile is incorrect; that is, the producer misjudges which fields are sensitive for linking. In this case, these systems might release data that are less anonymous than what was required by the recipient, and as a result, individuals may be more easily identified. This risk cannot be perfectly resolved by the producer of the data since the producer cannot always know what resources the recipient holds. The obvious demographic fields, physician identifiers, and billing information fields can be consistently and reliably protected. However, there are too many sources of semi-public and private information such as pharmacy records, longitudinal studies, financial records, survey responses, occupational lists, and membership lists, to account a priori for all linking possibilities.

Table 7. -- Contractual Requirements for Restricted-Use of Data Based on Federal Guidelines and the Datafly System

There must be a legitimate and important research or administrative purpose served by the release of the data. The recipient must identify and explain which fields in the database are needed for this purpose.

1. The recipient must be strictly and legally accountable to the producer for the security of the data and must demonstrate adequate security protection.
2. The data must be de-identified. It must contain no explicit individual identifiers nor should it contain data that would be easily associated with an individual.
3. Of the fields the recipient requests, the recipient must identify which of these fields, during the specified lifetime of the data, the recipient could link to other data the recipient will have access to, whether the recipient intends to link to such data or not. The recipient must identify those fields for which the recipient will link the data.
4. The provider should have the opportunity to review any publication of information from the data to insure that no potential disclosures are published.
5. At the conclusion of the project, and no later than some specified date, the recipient must destroy all copies of the data.
6. The recipient must not give, sell, loan, show, or disseminate the data to any other parties.

What is needed is a contractual arrangement between the recipient and the producer to make the trust explicit and share the risk. Table 7 contains some guidelines that make it clear which fields need to be protected against linking since the recipient is required to provide such a list. Using this additional knowledge and the techniques presented in the Datafly System, the producer can best protect the anonymity of patients in data even when the data are more detailed than data for public-use. Since the harm to individuals can be extreme and irreparable and can occur without the individual's knowledge, the penalties for abuses must be stringent. Significant sanctions or penalties for improper use or conduct should apply since remedy against abuse lies outside the Datafly System and resides in contracts, laws and policies.

Acknowledgments

The author acknowledges Beverly Woodward, Ph.D., for many discussions, and thanks Patrick Thompson, for editorial suggestions. The author also acknowledges the continued support of Henry Leitner and Harvard University DCE. This work has been supported by a Medical Informatics Training Grant (1 T15 LM07092) from the National Library of Medicine.

References

Alexander, L. and Jabine, T. (1978). Access to Social Security Microdata Files for Research and Statistical Purposes, *Social Security Bulletin,* (41), 8.

Duncan, G. and Lambert, D. (1987). The Risk of Disclosure for Microdata, *Proceedings of the Bureau of the Census Third Annual Research Conference*, Washington, D.C.: Bureau of the Census.

Hundepool, A. and Willenborg, L. (1996). μ - and τ-ARGUS: Software for Statistical Disclosure Control, *Third International Seminar on Statistical Confidentiality*, Bled.

Lasalandra, M. (1997). Panel Told Releases of Medical Records Hurt Privacy, *Boston Herald*, Boston, (35).

National Association of Health Data Organizations. (1996). A Guide to State-Level Ambulatory Care Data Collection Activities, Falls Church, VA.

Skinner, C. and Holmes, D. (1992). Modeling Population Uniqueness, *Proceedings of the International Seminar on Statistical Confidentiality*, International Statistical Institute, 175-199.

Sweeney, L. (1997). *Guaranteeing Anonymity When Sharing Medical Data, The Datafly System*, MIT Artificial Intelligence Laboratory Working Paper, Cambridge, 344.

Willenborg, L. and De Waal, T. (1996). *Statistical Disclosure Control in Practice*, New York: Springer-Verlag.

Woodward, B. (1996). Patient Privacy in a Computerized World, *1997 Medical and Health Annual 1997*, Chicago: Encyclopedia Britannica, Inc., 256-259.

Chapter 12

Tutorial on Record Linkage Slides Presentation

Martha E. Fair and Patricia Whitridge, Statistics Canada

Acknowledgements

- Ted Hill
- Dr. Newcombe
- Pierre Lalonde
- Dores Zuccarini
- Maureen Carpenter

Overview of Subject

- Development and uses
  - » Future - into the 21st century
  - » Present
  - » Past
- How all the individual topics fit together

Outline (1)

- What you will learn in this tutorial -
  - » An overview of record linkage and its applications - future, present, past
  - » Record linkage vocabulary
  - » Deterministic and probabilistic linkage details
  - » Sample project - birth-death linkage

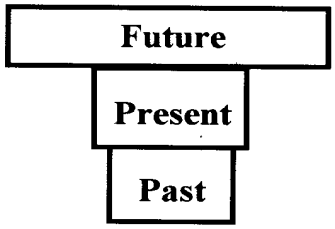
Introduction

- Definition of record linkage
- Statistical uses of record linkage
- Administrative uses of record linkage
- Deterministic linkage
- Probabilistic linkage

Outline (2)

- Getting the data ready for linkage - pre-processing
  Basic operations in a typical record linkage project
  - Searching - looking for the correct linkage
  - Decision making
  - Grouping
- Post-processing of files after linkage

Record Linkage



Outline (3)

- Tricks of the trade
- Examples of applications in health, business, and agriculture
- References where to get more information
- Glossary of terms
- Question period - interest of audience

Methodologies for the 21st Century

- Acquiring, generating, distributing and applying statistical knowledge strategically in a timely fashion
- Innovation
  - » Products
  - » Technologies
  - » Ways in which we generate and use data

Slides Presentation (cont'd)

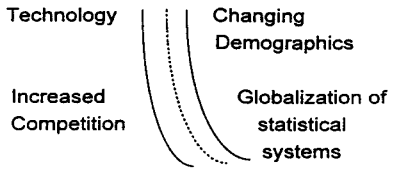
Methodologies for the 21st Century

- Work units change
- Budget reductions
- Redefine our business and our workplace
- Customizing of products
- Rethink and re-image our relationships

Three Main Indicators of Success of a Statistical System

- Adaptability of system in adjusting product line to evolving needs
- How effective is the system in exploiting existing data to meet client needs?
- How credible is the system in terms of statistical quality of its outputs and its non-political objectivity?

The Information Highway and Change



Some Attributes of Health Data in the Information Age

- Comprehensiveness
- Inclusiveness
- Linkage over time - longitudinal
- Patient-oriented
- Complete
- Accurate
- Secure

Building Bridges

Generalized systems	Data integration
Data control	Data access
Data analysis	Dissemination
Small area studies	Collaborative
Events	People
Cross-sectional	Longitudinal

Counting People in the Information Age

- Address list development
- Use of administrative records
- Matching and elimination of duplicate records
- Methods for hard-to-enumerate populations

Fixing the Potholes

- Coding standards and definitions
- Data quality
- Hardware and software incompatibilities
- Complexity of data
- Unnecessary duplication of data
- Timeliness

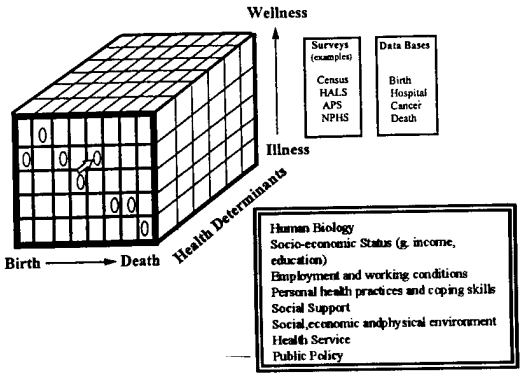
Record Linkage -- Today's Situation (1)

- Shift from paper-based systems to electronic
- Optical imaging of source documents
- Generalized systems
- Suite of software products
- Commercial softwares

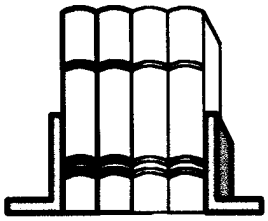
Slides Presentation (cont'd)

Record Linkage --Today's Situation (2)

- Rapid changes
  - markets
  - customer values - meeting our customers changing needs
  - technologies
- Optimal balance between
  - cost
  - quality
  - timeliness



Early Concepts of Record Linkage (1)



Book of Life

Preparing for the Journey Ahead -- the Life Cycle of Events (2)

- Longitudinal files
- Study people or business over time - birth to death
- Understand reasons that lead to different outcomes
- Determinants

Early Concepts of Record Linkage (2)

1. Dunn, H.L. (1946). Record Linkage. *Am J Public Health*, 36, 1412-1416.
2. Newcombe, H.B., J.M.Kennedy, S.J. Axford, A.P. James (1959) Automatic Linkage of Vital Records. *Science*, 130, 954-959.
3. Fellegi, I.P. and A.B.Sunter (1969). A theory of record linkage. *JASA* 64, 1183-1210.

Framework -- Life Cycle of Events (1)

- Maternal Health
- Birth
- Congenital anomalies
- Health surveillance registries
- Childhood - illness
- Childhood - injuries
- Childhood - cancers

Preparing for the Journey Ahead -- the Lifecycle of Events (1)

- Lifecycle of health events
- Lifecycle of businesses
- Lifecycle of an individual
- Lifecycle of a family/household
- Snapshot of the entity
- Forecast of a nation's demographic and economic future

Framework -- Life Cycle of Events (2)

- Occupational and environmental sources
- Health surveys
- Mental health
- Disease specific registries
- Diet and Lifestyle surveys
- Screening programs
- Aging
- Death

Slides Presentation (cont'd)

Record Linkage in the Toolbox of Software

- Common set of software products in reengineering
- Administrative and statistical programs

Highlights of Record Linkage Developments (4)

- Communication and collaboration
  - » Provinces and states
  - » United States
  - » England - Oxford
  - » Scotland
  - » Northern Ireland
  - » Australia
  - » Other countries

Highlights of Record Linkage Developments (1)

- Experience
- Key technical issues
  - » No unique identifier
  - » Discrepancies in identifiers
  - » Processing the large volume of data with reasonable computer time
- Theory

VOCABULARY

- **Glossary of terms**
- **Define the terms as used in this subject**
- **Deterministic linkage**
- **Probablistic linkage**

Highlights of Record Linkage Developments (2)

- Generalized systems
  - » One file linkage
  - » Two file linkage
- Applications
- Development of national data bases

Criteria for Personal Identifying Information (1)

1. Permanent - should exist at birth and remain unchanged
2. Universal - every member of the population
3. Reasonable - person no objection to its disclosure
4. Economical
5. Simple
6. Available

Highlights of Record Linkage Developments (3)

- Development of related generalized software
  - » Automated coding and text recognition
  - » Geographic coding
  - » Preprocessing of files
  - » Postprocessing of files
- Refinements in methods

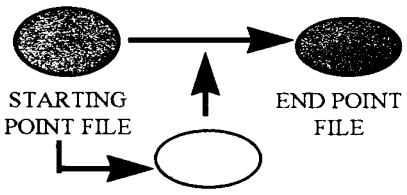
Criteria for Personal Identifying Information (2)

7. Known
8. Accurate
9. Unique

- No identifier or identity set has been devised that is in universal use.
- The efficiency of the record linkage operation depends on how well the items selected for comparison satisfy this standard.

Slides Presentation (cont'd)

Probabilistic Linkage (1)



Vocabulary -- Basic Terms (3)

- Linked pairs
- Possibly linked pairs
- Unlinked/nonlink pairs
- Global weights
- Frequency weights
- Discriminating power
- Specific discriminating power

o Match
o Gray area
o Unmatched

Name:	Date of Birth	Birthplace
Zacharius Orvil Quigley	1897 03 23	Martinique
Zacharius Orvil Quigley	1897 05 25	Martinique

Probabilistic Linkage (2)

Example 1

Name: Martha Fair S.I.N. 123 456 789
John Fair 123 456 789

Example 2

Name: J. Smith Date of Birth: 1900 01 15
J. Smith 1900 01 15

Example 3

Name:	Date of Birth:	Birthplace
Zacharius Orvil Quigley	1897 03 23	Martinique
Zacharius Orvil Quigley	1897 05 25	Martinique

Deterministic Linkages

If there are unique identifiers...

Table A		Table B
#111-222-333	=	#111-222-333
#444-555-666	=	#444-555-666

When is it appropriate to use the deterministic approach?

Vocabulary -- Basic Terms (1)

- Blocks / Pocket identifiers
- Blocking items
- Sort keys

Example 1

Name: Martha Fair S.I.N. 123 456 789
John Fair 123 456 789

Example -- National Death Index Matching Criteria

Agreement on:

1. SSN, Ist Name OR
2. SSN, Last Name OR
3. SSN, father's surname OR
4. If Female, SSN, Last name on user's file = father's surname on NDI OR
5. Month of birth, YoB, First Name, Last name OR
6. MoB, YoB, First Name, Father's surname OR
7. If Female, MoB, YoB, First Name ,Last Name on user's file = father's surname on NDI OR

Vocabulary -- Basic Terms (2)

- Rules
- Correlated items
- Weights
- Total weight
- Thresholds

Example 2

Name: J. Smith Date of Birth: 1900 01 15 Age: 97
J. Smith 1900 01 15 97

Example -- National Death Index Matching Criteria

8. MoB, YoB, first and middle initials, last name OR
9. Month and ± 1 year YoB, first and middle initials, last name OR
10. MoB,± 1 year YoB, first and last name OR
11. MoB, DoB, first and last name OR
12. MoB, DoB, first and middle initials, last name.

- Last name and father's surname spelling - agree on spelling/NYSIIS code
- Optional agreements - Can generate a set of possible and true set of links for resolution

NCHS - 1990

Slides Presentation (cont'd)

Probabilistic Linkages

What do you do if you generate many possible links.....

Table A		Table B
Smith Susan / 40-03-04	????	Smith S / 40-04-03
Joseph Browne / 43-01-12	????	Joe Brown / 43-01-21

Are those links???

Theory of Record Linkage (2)

Goal: Divide Set C into sets:

L: (Links)

Table A				Table B		
Jones	Fred	1938		Jones	Fred	1938
Smith	Susan	1940		Smith	S	1939

U: (Non-links)

Table A				Table B		
Jones	Fred	1938		Smith	S	1939
Smith	Susan	1940		Jones	Fred	1938
Walker	John	1936		Jones	Fred	1938
Walker	John	1936		Smith	S	1939

Kinds of Linkages

One File (Internal)

Record	Table A
1	FRED
2	SUSAN
3	S
4	JOHN
5	SUE

Two File

Table A		Table B
FRED	⟷	FRED
SUSAN	⟷	S
JOHN		

Partitioning Set C

In practice C is split into:

U (Unlinked) , P (Possibly Linked), L (Linked)

P - automatic mapping or

P - manually examined to reset STATUS

Error Types

Type I: L pairs erroneously assigned to U

Type II: U pairs erroneously assigned to L

- Attempt to minimize size of P while controlling error rates

Theory of Record Linkage (1)

Set C has N_a x N_b record pairs (3x2=6)

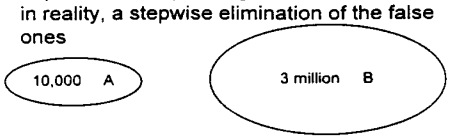
Table A				Table B		
Jones	Fred	1938		Jones	Fred	1938
Jones	Fred	1938		Smith	S	1939
Smith	Susan	1940		Jones	Fred	1938
Smith	Susan	1940		Smith	S	1939
Walker	John	1936		Jones	Fred	1938
Walker	John	1936		Smith	S	1939

Rules and Thresholds (1)

- RULES classify pairs into L or U
  - » RULES use one or more input fields
  - » There are usually several RULES
- RULES produce OUTCOMES of:
  - » Agreement
  - » Disagreement
  - » Partial Agreement
  - » Missing

Linkage Approaches

The process of separating out the true links is, in reality, a stepwise elimination of the false ones



Total pairs =30 billion

Linked pairs = 1000 expected

True unlinked pairs in set C = 30 billion - 1000

Rules and Thresholds (2)

Table A				Table B		
Smith	Susan	1940		Smith	S	1939

RULES (r_i) could be:

- » Surname (r_1)==> Agreement
- » Given Name (r_2)==> Partial Agreement

Birth Date (r_3)==> Disagreement

$$R(a,b) = (r_1(a,b), r_2(a,b), r_3(a,b))$$

Slides Presentation (cont'd)

Rules and Thresholds (3)

Table A				Table B		
Smith	Susan	1940		Smith	S	1939

ODDS RATIO

$$O(a,b) = \frac{P(R(a,b)|(a,b) \in L)}{P(R(a,b)|(a,b) \in U)}$$

Probabilities and Thresholds

- You need to estimate
  - » values for T_L and T_u
  - » and probabilities for all rules and outcomes:
    - $P(r_i(a,b)$ given that (a,b) is in L
    - $P(r_i(a,b)$ given that (a,b) is in U
- Solutions:
  - » Direct estimation
  - » Use of prior information or similar linkages
  - » Iteration

Rules and Thresholds (4)

O (a , b) [ODDS RATIO]
- » large ==> (a , b) is a link
- » small ==> (a , b) is a non-link

Weights (1)

- A **weight** is assigned to each rule that is used in the comparison.
- Logarithms to the base two are often used as in information theory. They may be multiplied by ten to avoid decimal points.
- The weights for all rules are summed to produce a **total weight.**

Thresholds

Partition Set C

- » T_L Lower Threshold
- » T_u Upper Threshold

$O(a,b) < T_L$ **assign (a,b) to U**

$T_L <= O(a,b) <= T_u$ **assign (a,b) to P**

$O(a,b) > Tu$ **assign (a,b) to L**

Weights (2)

OUTCOME Weights

Generic (Independent of Field Value)
Agreement, Disagreement, Partial, Missing...

Frequency (Dependent on Field Value)
Rare values have higher weights (Quigley)
Common values have lower weights (Smith)

Simplifying Assumptions

RULES are independent of one another

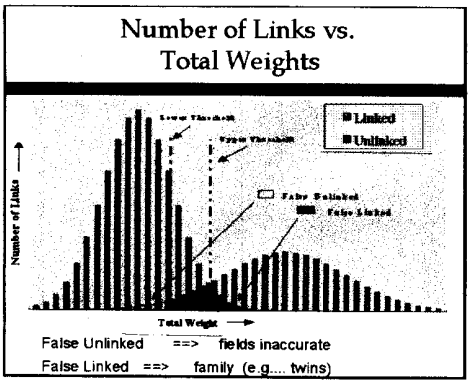
$$P(R) = P(r_1) \times P(r_2) \times \ldots P(r_K)$$

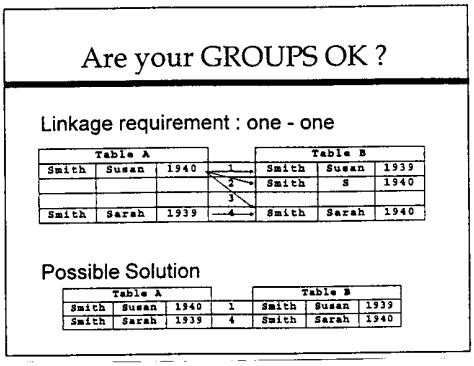
Input tables can be partitioned into POCKETS
- » POCKET fields are "reliable"
- » Pairs not in same POCKET are in set U
- » Reduces number of pairs compared

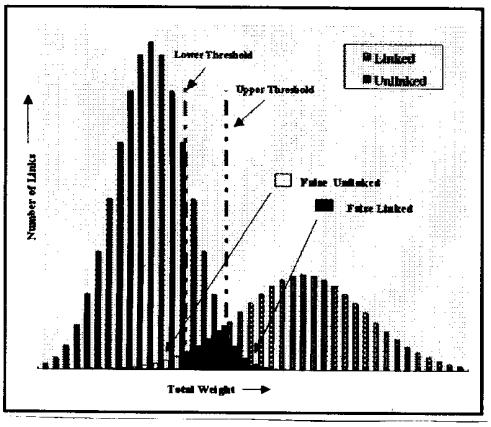
Iterative Estimation in Practice

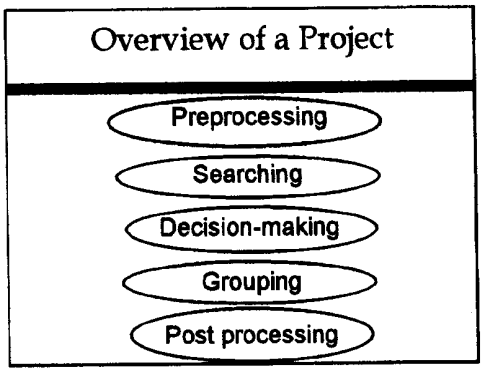
- U Frequency Weights (Agreement, Disagreement)
  - » use frequency of values on one of the input files
- L Weights obtained iteratively
  - » create your pairs
  - » examine them and revise THRESHOLDS
  - » recalculate STATUS to generate "new" Weights

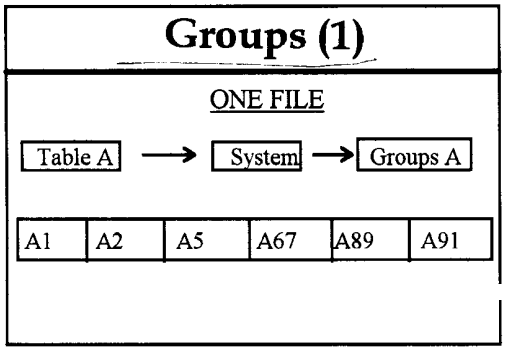
Slides Presentation (cont'd)

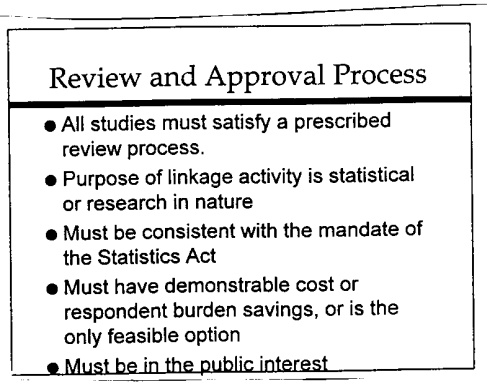


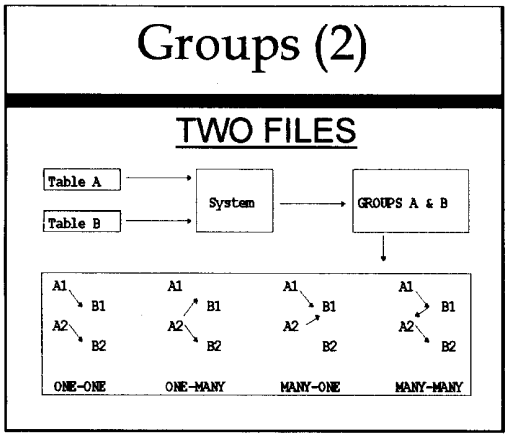


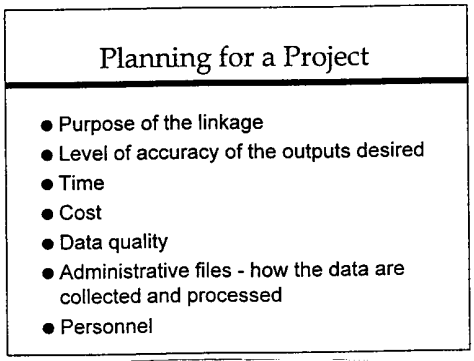






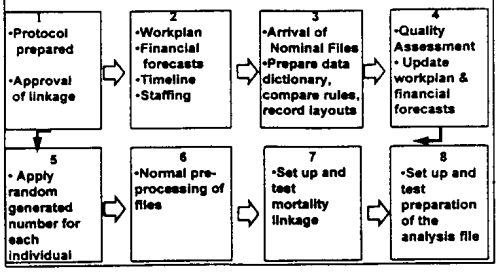






Slides Presentation (cont'd)

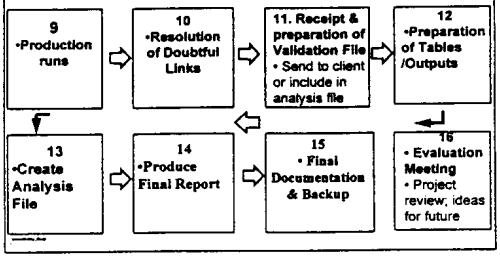
Overview of a Typical Project (1)



Identifying Numbers (1)

- Identifying number
- TIPS - some questions to ask
  - » Is this number unique for each individual e.g..... SSN? Health Insurance Number?
  - » Is there a probability that more than one individual could have this number e.g.... SSN?
  - » Are there alternate entries on the file for this individual i.e.. if the individual changes their name are two records kept?

Overview of a Typical Project (2)



Identifying Numbers (2)

- » Is this number ever re-used by another person e.g..... after the first individual dies?
- » What happens if the individual changes their surname, address etc.. Is this file updated?
- » Is this number common for a family?
- » What happens when the family structure changes?
- » What happens to late registrations?
- » What happens to cancellations/ errors in the file?

An Example -- Birth-Death Linkage Project

- Purpose : To link infant birth and death records
- Outputs desired: Analytical file
- File sizes: About 800,000 birth records being linked to 6,000 death records

PREPROCESSING the Files for Birth -- Death Linkage

- Quality assessment of data items
- Examine items that are common to the two records
- Standardize items such as names, forenames,addresses, geography coding
- Encode and recode fields e.g..... surnames
- Create extra fields as necessary
- Generate duplicate records if required e.g.... maiden and married name for women

TOPIC ONE -- Pre-processing

- **Explain details**
- **Give an example**

Availability and Validity Checks

- Record layout may give the impression data are available for linkage

TIPS

- You may want to add a field to the record indicating whether linkage items are available or blank for the file
  - » 0= blank
  - » 1=available
- Check for valid values, ranges and codes

Slides Presentation (cont'd)

Surnames

- How are surnames assigned in this file
  - » e.g.... in Quebec the legal surname for women are their maiden names
- How are the surnames recorded
  - » Special characters - "
  - » Prefixes
  - » Titles
  - » Surname suffixes
  - » Double barreled names e.g..... Smith-Jones

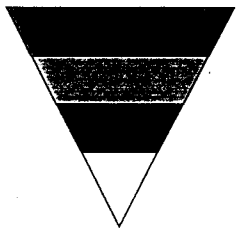
TOPIC TWO -- Searching Phase

- **Explain details**
- **Give an example**

Surnames -- TIPS

- Run quality checks
  - » e.g.. SAS - change all letters to A; leave special characters O'Connor ->A'AAAAAA
  - » - list the frequency distribution of names on the file
  - » check for special names e.g..... nuns, also known as
- Name check
- Special pre-processing of names
- Special characteristics of names

Input File



Blocking variables
Reject Rules
Early cutoff
Global weights
Thresholds
Groupings

Links

Birth Date -- TIPS

- Make sure the dates are year 2000 compatible
- Frequency distributions e.g..... SAS
  - » Look for unlikely dates - particular 1900 instead of blank
  - » Look for the difference between missing and blank
    - – Some use a special code for missing (e.g..... 99)
  - » Look for illogical values of year,month and day

Selecting Blocking Variables -- TIPS

- Blocking variables
  - » Pass 1 - sex code and NYSIIS phonetic code for surname
  - » Pass 2 - sex code and birth date
  - » Pass 3 - birth date only - cases failed pass 2

Geographic Codes -- TIPS

- Examine the codes over time to ensure they are compatible
- Standardize addresses
- Watch out for items that are common to the two files that may be correlated e.g..... place of residence, place of hospital, place of birth

Details of Phonetic Coding of Surnames

- Characteristics
  - » Vowel information is either partially or wholly suppressed because of its instability
  - » Certain consonants with similar sounds are replaced by a standard character
- Examples:
  - » NYSIIS
  - » Soundex
  - » ONCA

Slides Presentation (cont'd)

Examples of NYSIIS Codes

Andersen, Anderson ---> ANDAR
Brian,Brown,Brun ---> BRAN
Capp,Cope,Copp,Kipp--->CAP
Dane,Dean,Dent,Dionne->DAN
Smith,Schmit,Schnidt ---->SNAT
Trueman,Truman ---->TRANAN

SEARCHING Phase

- Objective is to search for pairs that are truly linked
- Possibly apply early rejection rules e.g..... not one item other than the pocket identifiers agree
- Decide on the most efficient order of comparisons e.g..... quick cutoff
- Specify rules and weights to be used in the comparisons

20 Most Common Surnames (1)

Canada			United States		
Rank	Name	%	Name	%	US (Can)
1.	SMITH	0.72	SMITH	0.99	1. (1)
2.	BROWN	0.39	JOHNSON	0.76	2. (18)
3.	WILSON	0.32	WILLIAMS	0.60	3. (16)
4.	MACDONALD	0.30	BROWN	0.56	4. (2)
5.	JOHNSON	0.29	JONES	0.56	5. (12)
6.	MARTIN	0.28	MILLER	0.48	6. (14)
7.	TREMBLAY	0.28	DAVIS	0.44	7. (-)
8.	ANDERSON	0.27	ANDERSON	0.33	8. (8)
9.	CAMPBELL	0.26	WILSON	0.33	9. (3)
10.	TAYLOR	0.25	MOORE	0.29	10. (-)

TOPIC THREE -- Decision-making Phase

- **Weights**
- **Creating comparison rules**
- **Setting thresholds**
- **Manual resolution - optional**

20 Most Common Surnames (2)

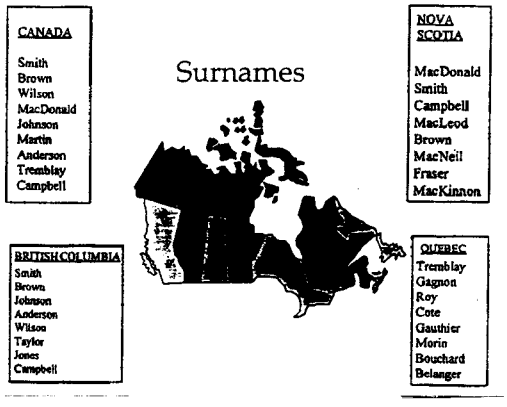
Canada			United States		
Rank	Name	%	Name	%	US (Can)
11.	ROY	0.24	TAYLOR	0.29	11. (10)
12.	JONES	0.23	THOMAS	0.27	12. (-)
13.	THOMPSON	0.23	WHITE	0.27	13. (17)
14.	MILLER	0.23	MARTIN	0.27	14. (6)
15.	GAGNON	0.21	THOMPSON	0.27	15. (13)
16.	WILLIAMS	0.21	JACKSON	0.26	16. (-)
17.	WHITE	0.20	HARRIS	0.24	17. (-)
18.	JOHNSTON	0.20	CLARK	0.23	18. (-)
19.	LEBLANC	0.19	LEWIS	0.21	19. (-)
20.	YOUNG	0.19	WALKER	0.21	20. (-)

Ref:NCHS - 1990

Using The Discriminating Power of Items (1)

- Agree,disagree,missing
- Agree, disagree, partial agreements
- Agree, disagree, partial agreements with global weights
- Agree, disagree, partial agreements using frequency weights
- Conditional agreements

Surnames



Using the Discriminating Power of Items (2)

- Concatenated comparisons
- Cross comparisons
- User-defined code for comparisons - recognizing degrees of similarity

Slides Presentation (cont'd)

Deciding What Comparison Outcomes to Recognize

- Make a list of the items available on A
- Make a list of items available on B
- Examine the data for additional rules that simulate the logic that one would use manually e.g.....date of death versus date of birth
- Examine geographical and mobility patterns that make sense

Comparing and Cross Comparing Months and Days of Birth

- Watch out for different conventions for recording the month and day

Deciding What to Do with Missing Values

- Second given names may not be present for the individual
- Watch out for things like -
  - » Twin 1
  - » Twin 2
  - » Baby Boy
  - » Baby Girl

Decision Making

- Calculate outcome weights
- Decide which pairs are links

Comparing Surnames

TIPS

- Phonetic coding
- Partial agreements
- String comparators
- Maiden versus married names
- Watch out for titles - Sr. Jr.

Histogram of Weights

- Each pair produces a total weight
- The total number of pairs and distribution of weights can be examined
- Note that the number of non-links far exceeds the number of links
- Linkage is how to get rid of the hay and leave the needles - rather than trying to find the needles in the haystack

Comparing and Cross Comparing Initials

TIPS

- Watch out for baptismal names in the first forename field e.g.... Mary and Joseph
- Forenames may have surnames in them by error

Manual Resolution

- Should it be done?
- How much?
- What will it cost?
- Who should do it?

Slides Presentation (cont'd)

TOPIC FOUR – Grouping Phase

- **Grouping**
- **Mapping**
- **Conflict resolution**
- **Manual resolution**
- **Updates**

Post processing

- Bringing in other files e.g.....histories
- Validation files
- Creating analysis files without names
- Saving rules, weights and other items

Grouping

- Watch out for multiple births - prepared special listings for resolution
- Watch out for special naming conventions

TOPIC FIVE

- **Post processing**

Mapping

- One to one
- One to many
- Many to one
- Many to many

- Conflict resolution
- Manual resolution
- Updates

Documentation of the Process

- Data dictionaries
- Record layouts
- Flow diagrams
- Histograms of weights
- Threshold settings
- Rules and weights used
- Analysis file

Multi-pass Linkages

- Decide on the number of passes required
- Each pass should have different blocking criteria
- Choose blocking items that do not overlap in order to pick up the missing links not achieved on an earlier pass
- Examples: NYSIIS code and sex code
  Birth date and first forename

Errors, Their Sources and Magnitudes

- Blocking information.
  - » Use multi-pass
  - » Use a different file
- Thresholds
- Lack of discriminating power
- Underuse of discriminating power
- Correlated items
- Independent validity check

Slides Presentation (cont'd)

Software

- Specific
- Generalized
- Suite of software

Creation of Statistical Data

- Supplementary surveys
- Release of public use tapes
- Building new data sources
- Creation of patient-oriented, rather than event-oriented statistics

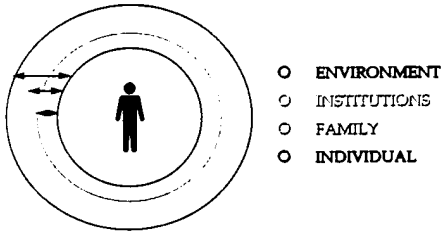
Generalized Record Linkage System

- Version 3 under development
- ORACLE and C compiler
- Unix operating system
- Allows internal or two-file linkages
- Uses weights to determine likelihood pairs of records relate to the same entity

Record Linkage -- Uses in Health Research (1)

- Mortality, cancer and/or birth follow-up of
  - » Cohorts (e.g..... miners, asbestos workers)
  - » Case control studies
  - » Clinical trials
- Building, maintaining and using registries
- Creation of patient-oriented histories
- Follow-up of surveys

The Individual and Society -- Uses of Record Linkage



Record Linkage -- Uses in Health Research (2)

- Occupational and environmental health studies
- Examining factors which influence health care usage and costs
- Regional variations in the incidence of disease

Record Linkage -- Tool for Creation of Statistical Data

- Data quality - e.g..... elimination of duplicates
- Assess data quality
- Coverage - e.g..... reverse record check
- Tracing tool - e.g..... longitudinal studies
- Addition of new variables e.g.. analysis files
- Sampling frame - e.g..... census of agriculture farm register

Files and Facilities

1. **Endpoint files**
   Canadian Birth Data Base
   Canadian Cancer Data Base
   Canadian Mortality Data Base

2. **Generalized systems**
   Record linkage
   Automated coding

Slides Presentation (cont'd)

Use of Record Linkage in Cancer Registries

- Creation of cancer registries
- Maintaining cancer registries
- Death clearance of cancer registries
- Evaluating the quality of registries
- Ascertainment of new death certificate only cases
- Replacing or partially replacing active follow-up of patients
- Carrying out cohort studies
- Follow-up of clinical trials and screening programs

Follow-up of National Breast Screening Program Cohort

OBJECTIVES

To follow-up women to determine the dates and causes of death

To confirm the diagnosis and cancer incidence of the study population

Number of Individuals:
90,000 females

YEARS

Cancer years:
1977-1993

Death years:
1980-1988
1989-1993

ORGANIZATIONS

University of Toronto
Statistics Canada

Advantages of Record Linkage in Cancer Registries

- Reduces respondent burden
- Improves accuracy
- Reduced follow-up costs
- Refines detection and measurement of mortality and cancer rates for particular cohorts

Death Clearance of the Nova Scotia Cancer Registry

OBJECTIVES

- **To calculate survival rates for persons with cancer in Nova Scotia**
- **To add "death certificate only" cases to the Nova Scotia Cancer Registry**

Number of Records:
About 79,000 cancer records relating to about 60,000 individuals were linked to 3.6 million individual deaths

YEARS

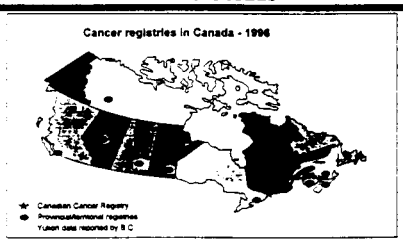
Cancer years:
1969-1988

Death years:
1969-1989

ORGANIZATIONS

- Health Canada
- Nova Scotia Cancer Registry
- Statistics Canada

USE OF RECORD LINKAGE IN BUILDING, MAINTAINING AND USING CANCER REGISTRIES



Canada

Occupational Studies

- Ontario cancer study
  - » Feasibility study in seven Ontario regions
  - » Linkage to Ontario cancer registry

Follow-up of Childhood Cancers

OBJECTIVES

- To examine the problems and risks facing Canadian children with respect to cancer

Number of Individuals:
Approximately 17,000
(including 1985-1991 Ontario cases)

YEARS

Cancer years:
1969-1988

Death years:
1969-1991

ORGANIZATIONS

Health Canada
Statistics Canada
Provincial cancer agencies

Key Elements in a Typical Study

1. **Exposed population**
2. **Comparison group**
3. **Hazard identification**
4. **Endpoints**
5. **Levels of exposure**
6. **Time course of response**
7. **Confounding factors**

Slides Presentation (cont'd)

Canadian Farmers Study

OBJECTIVE

To investigate the mortality and cancer incidence among Canadian farmers

Number of Individuals:
About 326,000 males and females

Data Sources
1971 Census of Agriculture
1971 Census of Population
1971 Central Farm Register
1971-1987 Canadian Mortality Data Base
1971-1986 Canadian Cancer Data Base

ORGANIZATIONS
Health Canada
Statistics Canada

National Dose Registry Study

OBJECTIVE

- To investigate the effects of low level radiation on the Canadian workforce who participate in the National Dose Registry

YEARS
Cancer years: 1969-1988
Death years: 1950-1987

Number of Individuals:
About 255,000 male and female individuals

ORGANIZATIONS
Health Canada
Statistics Canada

Long-Term Medical Follow-up Results (1)

1. Improve Scientific Knowledge of Health Hazards and Risks
   - » Canadian National Dose Registry study
   - » Fluoroscopy study
2. Provide Information to Help Set Safety Standards
   - » Ontario miners study
3. Assist With Health Promotion Activities
   - » National Breast Screening program

Long-Term Medical Follow-up Results (2)

4. Assist Task Forces, Enquiry Boards in the Assessment of Occupational and Environmental Risks
   - Reproductive problems
   - Improve health of mothers and babies
5. Follow-up Populations for Delayed Health Effects
   - Occupational groups
   - Canadian Farm operators study
   - Dow Chemical
   - Esso Imperial study
   - INCO
   - Falconbridge
   - Firefighters

Long-Term Medical Follow-up (3)

6. Improve Computer Methods
   - » Data Collection
   - » Record linkage system
   - » Death clearance of cancer registry files
   - » Discriminating power of partial agreements of names for linking personal records
   - » Creation of pseudo-registry files
   - » Occupational coding from text
   - » Geographic coding using postal code file

Socio-economic Gradients in Mortality

The Use of Health Care Services at Different Stages in the Life Course

- Manitoba Centre for Health Policy and Evaluation
- Statistics Canada
- Mortality and health care utilization described in relation to socioeconomic status
- Measure mortality and use of health care services

RECORD LINKAGE

AGRICULTURE AND BUSINESS APPLICATIONS

OUTLINE

- Introduction
- Population Definition
- Matching Variables
- Challenges
- Census of Agriculture
- Other Agriculture Examples
- General Comments
- Future Directions

Slides Presentation (cont'd)

INTRODUCTION

- Record linkage techniques developed primarily for matching individuals
- Some aspects similar for businesses, some very different
- Special challenges are present in rural areas and for agricultural population
- Incentives to match admin data rather than run new surveys

INTRODUCTION (cont)

- Remember criteria for matching variables:

permanent	available
universal	known
reasonable	accurate
economical	unique
simple	

POPULATION DEFINITION

- Business entities
- Organization: incorporated, family-owned, individual-owned or partnership
- Entities can operate in several provinces
- Structures change over time

MATCHING VARIABLES

- Commonly available:
  - » name (individual or business)
  - » address
  - » phone number
  - » industrial classification (type of business)
- Rarely available:
  - » date of birth (individuals only)
  - » permanent address (eg place of birth)
  - » numeric information (data)

CHALLENGES: Unincorporated Businesses

- Business names rare
- Address confusion: home or business
- Phone number: home or business
- Date of birth not consistently reported, often unknown for partners
- Businesses can be very volatile; same person can go in and out of business in short time span

CHALLENGES: Incorporated Businesses

- Multiple locations possible
- Can involve complex structure of multi-holding corporations
- Different locations, different addresses and phone numbers
- No date of birth information

CHALLENGES: Address Information

- Rural addresses very weak if no mail delivery (eg. Prairie provinces)
- Could pertain to non-resident owners
- Could be different residence and business addresses
- Address not permanent
- Addresses difficult to parse

CHALLENGES: Business Names

- Difficult to parse
- Need to standardize (eg. inc, incorp, corp, lim, ltd, ltee)
- Need to remove articles (eg. the, a, le)
- More than one language (English and French)

Slides Presentation (cont'd)

CHALLENGES: Business Names (cont)

- Different naming customs across country (eg. enregistré, in Quebec)
- NYSIIS codes crucial to many linkages - developed for individual names
- NSKGEN (developed by Statistics Canada) parses business names - generates direct match key (DMK)

CENSUS OF AGRICULTURE: Informatics Environment

- Unix machine
- GRLS (Generalized Record Linkage System, developed at Statistics Canada) - Beta release
- Approximately 280,000 incoming Census records to be matched against 400,000 Farm Register records

CHALLENGES: Structural Problems

- Units to match not always same on both files (eg. businesses vs owners)
- Unincorporated businesses may be owned by many partners
- Individuals may be involved in more than one business

CENSUS OF AGRICULTURE Linkage Process

- Need to link Census farms to Farm Register
- 3 step process: exact match, probabilistic match, then manual resolution
- Exact match - incoming Census farms matched in SQL - includes pre- and post-processors

CHALLENGES: Structural Problems (cont)

- Same name often common in rural communities (clustering)
- Family structure to some businesses, parents and children both involved, difficult to match when passed down

CENSUS OF AGRICULTURE Linkage Process (cont)

- Probabilistic match - remaining unmatched Census farms matched using GRLS, one province at a time, to the Farm Register - includes pre- and post-processors
- Manual resolution - pairs of potential linked farms identified by GRLS manually resolved using extra information

CENSUS OF AGRICULTURE

- Held every 5 years in Canada - most recently May, 1996
- Detailed questionnaire dropped off to every farm (with Census of Population)
- Need to match Census farms with existing Farm Register

CENSUS OF AGRICULTURE Match

- Matching variables: farm name, operator names, phone number, postal code, agricultural region, (dob)
- Rules for full and partial agreements
- Population mix of incorporated and unincorporated farms
- Pockets based on NYSIIS codes of surname (originally tried DMK)

Slides Presentation (cont'd)

CENSUS OF AGRICULTURE Challenges

- Census records are farms, most matching fields at farm operator level
- Many farmers involved in more than one farm; many farms operated by more than one farmer; most farms unincorporated
- No data available to help match or resolve multiple matches

GENERAL COMMENTS

- Experience with GRLS and probabilistic linkage gained from Census of Agriculture
- Some problems can be simplified by changes to questionnaire or Farm Register file
- Re-think process for next Census

CENSUS OF AGRICULTURE Challenges (cont)

- Not all farms have farm name - when present, it drives match - difficult to standardize and parse to use in match
- Rural address information is poor
- Names are clustered and related to address information

FUTURE DIRECTIONS

- GRLS is being constantly improved
- Need to develop and maintain unique, permanent identifiers for businesses (use of Single Business Number will help)
- Matching is becoming more important with budgetary restraint

CENSUS OF AGRICULTURE Challenges (cont)

- Rules and weights adjusted for each province - characteristics very different - difficult to determine rules and weights to optimize discriminating power
- Problems with size of groups and pockets - had to subdivide, or even remove all records with one name

Future Directions and Summary

OTHER AGRICULTURE EXAMPLES

- Income tax files matched to the Farm Register each year
- Lists from growers' associations, marketing boards and provincial organizations also matched to Farm Register each year
- Challenges to link farms always present

Future Directions (1)

- Administrative uses
- Registration file
- Disease registries
- Quality control
- Quality of data sources
- Timeliness
- Customer satisfaction

Slides Presentation (cont'd)

Future Directions (2)

- Relevance of output
- Security of sensitive information
- Reengineering of statistical organizations
- Software developments
- Acquire and maintain information from variety of sources
- Multiple uses of data
- Stewardship of data

Summary

- Record linkage software development
- Quality of data files
- Uniform classification standards
- Analysis of data
- Analysis of data - incorporating uncertainty due to linkage

Summary

- Be useful
- Expand our horizons
- Ask the right questions
- Know the issues and be aware of priority information needs
- Build the right data and information
- Harmonize concepts and outputs
- Confidentiality protection

Address Information

Martha Fair
Chief, Occupational and Environmental Health Research Section
Health Statistics Division
Statistics Canada
R. H. Coats Building, Stn. 18R
Tunney's Pasture
Ottawa,Ontario
K1A 0T6

Phone: (613) 951-1734
Fax: (613) 951 -0792
E-mail: fairmar@statcan.ca

Appendix A -- Glossary of Terms

There are various terms used in record linkage. Some of these have been defined in: Newcombe, H.B. (1988). *Handbook of Record Linkage Methods for Health and Statistical Studies*, Administration and Business. Oxford, U.K. Oxford University Press, pp. 103-106.

The terms used in that book are as follows:

Blocking. -- The use of sequencing information (e.g., the phonetically coded versions of the surnames) to divide the files into "pockets." Normally, records are only compared with each other where they are from the same "pocket," i.e., have identical blocking information. The purpose is to avoid having to compare the enormous numbers of record pairs that would be generated if every record in the file initiating the searches were allowed to pair with every record in the file being searched.

Denominator. -- This usually refers to the denominator in a FREQUENCY RATIO, i.e., the frequency of a given comparison outcome among UNLINKABLE pairs of records brought together at random. It may be applied also to one of the two components of any ODDS.

Frequency Ratio. -- The frequency of a given comparison outcome among correctly LINKED pairs of records, divided by the corresponding frequency among UNLINKABLE pairs brought together at random. The comparison outcome may be defined in any way, for example as a full agreement, a partial agreement, a more extreme disagreement, or any combination of values from the two records that are being compared. The FREQUENCY RATIO may be specific for the particular value of an identifier when it agrees, or for the value of the agreement portion of an identifier that partially agrees, or it may be non-specific for value.

General Frequency. -- A weighted mean of the frequencies of the various values of an identifier among the individual (i.e., unpaired) records of the file being searched. It is non-specific for value. Value-specific frequencies are also obtained from the same source.

Global Frequency. -- The frequency of a comparison outcome among pairs of records, when that outcome is defined in terms that are non-specific for the value of the identifier. The outcome may be a full agreement, a partial agreement, or a more extreme disagreement. The record pairs may be those of a LINKED file, or they may be UNLINKABLE pairs brought together at random. Only in the special case of the full agreement outcomes are the global and the general frequencies numerically equal, but they always remain conceptually different. The difference is that a global frequency, although value non-specific, always reflects the full definition of the non-agreement portion of that definition. A general frequency cannot do this because it is based on a file of single (i.e., unpaired) records.

Global Frequency Ratio. -- The ratio of the global frequency for a particular comparison outcome among LINKED pairs of records, divided by the corresponding frequency among UNLINKABLE pairs. It is equivalent to the global ODDS. GLOBAL FREQUENCY RATIOS for agreement outcomes and partial agreement outcomes are often subsequently converted to this value-specific counterparts during the linkage process. The conversion is accomplished by means of an adjustment upwards where the agreement portion of the identifier has a rare value, and an adjustment downwards where the value is common.

Linkage. -- In its broadest sense, RECORD LINKAGE is the bringing together of information from two or more records that are believed to relate to the same "entity." For an economic or social study, the "entities" in question might be farms or businesses. For a health study, the "entities" of special interest are usually individual people or families. It is in the latter sense that the word is used throughout this book.

Linked. -- In line with the above definition of "record linkage," LINKED pairs of records are pairs believed to relate to the same individual or family (or other kind of entity). Record pairs brought together and judged not to relate to the same individual or family may be referred as "UNLINKABLE" pairs. For short, the two sorts of pairs are sometimes called "LINKS" and "NON-LINKABLE," respectively. As used here, the term implies that some sort of decision has been reached concerning the likely correctness of the match.

Matched. -- This word is variously used in the literature on record linkage. In this book, however, it is given no special technical meaning and merely implies a pairing of records on the basis of some stated similarity (or dissimilarity). For example, early in a linkage operation, records from the two files being LINKED are normally matched for agreement of the surname code. The resulting pairs may also be called "candidate pairs" for linkage, but this emphasis is most appropriate in the later stages when the numbers of competing pairs have diminished. Pairs of records will frequently be spoken of as "correctly matched," "falsely matched," or "randomly matched."

Numerator. -- This usually refers to the numerator in a FREQUENCY RATIO, i.e., the frequency of a given comparison outcome among pairs of records believed to be correctly LINKED. It may be applied also to one of the two components of any ODDS.

Odds. -- This word is used in its ordinary sense but is applied in a number of situations. As relating to a particular outcome from the comparison of a given identifier it is synonymous with the FREQUENCY RATIO for that outcome. As relating to the accumulated FREQUENCY RATIOS for a given record pair it refers to the overall RELATIVE ODDS. It is also applied to the overall ABSOLUTE ODDS.

Outcome. -- This refers to any outcome or result from the comparison of a particular identifier (or concatenated identifiers) on a pair of records, or the comparison of a particular identifier on one record with a different but logically related identifier on the other. It may be defined in almost any way, for example as an AGREEMENT, a PARTIAL AGREEMENT, a more extreme DISAGREEMENT, any other SIMILARITY or DISSIMILARITY, or the absence of an identifier on one record a s compared with its presence or absence on the other. An outcome may be specific for a particular value of an identifier (e.g., as it appears on the search record) or for any part of that identifier, especially where there is an agreement or partial agreement; it may be non-specific for value; or it may even be specific for a particular king of DISAGREEMENT defined in terms of any pair of values being compared.

Value. -- An identifier (e.g., an initial) may be said to have a number of different "values" (e.g., initial "A," initial "B," and so on). Surnames, given names, and places of birth have many possible values. Other identifiers tend to have fewer values that need to be distinguished from each other.

Weight. -- In the literature, this term has been widely applied to the logarithms of various entities, such as:

- a FREQUENCY RATIO for a specified outcome from the comparison of a given identifier;
- the product of all the FREQUENCY RATIOS for a given record pair;
- the NUMERATOR of a particular FREQUENCY RATIO;
- the DENOMINATOR of a particular FREQUENCY RATIO;
- any estimate of such a numerator or denominator, not obtained directly from a file of matched pairs of records.

The use of the logarithm is merely a convenience when doing the arithmetic; it does no affect the logic except to make it appear more complicated. The term "WEIGHT" has therefore been employed sparingly in this book. Instead, reference has been made directly to the source frequency or FREQUENCY RATIO, or to the estimates of these, wherever possible.

Chapter 13

A Checklist for Evaluating Record Linkage Software

Charles Day, National Agricultural Statistics Service

From the 1950's through the early 1980's, researchers and organizations undertaking a large record linkage project had little choice but to develop their own software. They often faced the choice of using less accurate methods or expending dozens of staff years to create proprietary systems. For example, in the late 1970's, the U.S. National Agricultural Statistics Service spent what is conservatively estimated as 50 staff years to develop a state-of-the-art system. Happily, today's record linkage practitioners no longer need do this any more than they need to write their own word processing software, nor should they attempt it. Powerful, flexible, relatively inexpensive software that implements all but the most sophisticated methods is available in the form of generalized packages that can stand alone or software components that can be integrated into a surrounding application. There is no longer any reason for anyone but researchers into the theory of record linkage to attempt to write record linkage software from scratch.

The Record Linkage Workshop and Exposition featured six vendor representatives who exhibited their software on site. This checklist is provided as an aid in evaluating the record linkage software they sell, along with other products that may enter the market. While the authors have endeavored to make the checklist as complete as possible, there may still be important characteristics for your application that the checklist does not cover. There is no substitute for a thorough analysis of your individual needs. Comments on the checklist are welcome. Please email them to *cday@nass.usda.gov* .

****

General

1.1 Is the software a generalized system or specific to a given application?

1.2 Is the software a:

Complete system, ready to perform linkages "out of the box?"

Set of components, requiring that a system be built around them? If so, how complete are the components?

Part of a larger system for performing integrated mailing list functions?

1.3 What types of linkages does the software support?

Unduplication (one file linked to itself)?

Linking two files?

Simultaneously linking multiple files?
Linking one or more files to a reference file (e.g., geographic coding)?

1.4 Can the software be used on the following computers:

Mainframes?
Mini-computer?
Workstation?
IBM-compatible microcomputer?
Macintosh?

1.5 Can the software run under the following operating systems:

MS/PC DOS?
OS/2?
Windows 3.1/95?
Windows NT?
UNIX?
VMS?
Mac OS?
Novell NetWare?
Mainframe OS (e.g., IBM MVS)?

1.6 For PC based systems, what level of processor is required? How much memory? How much hard drive space?

1.7 Can the system perform linkages interactively (in real time)? Can it operate in batch mode?

1.8 How fast is the software on the user's hardware and files the size of the user's files? If the software is interactive, is its performance adequate?

1.9 If the software is to be used as part of a statistical estimation system, are the methods used in the software statistically defensible?

1.10 Is the vendor reliable? Can the vendor provide adequate technical support? Will they continue to exist for the projected life of the software? If this is in question, is a software escrow available? Is the user prepared to support the software him/herself?

1.11 How well is the software documented? Can a new user reasonably be expected to sit down with the manual and begin using the software, or will training be necessary? Does the vendor provide training? At what cost?

1.12 What features does the vendor plan to add in the near future (e.g., in the next version)?

1.13 Is there a user group? Who else is using the software? What features would they like to see added? Have they developed any custom solutions (e.g., front ends, comparison functions) they would be willing to share?

1.14 Is other software, such as database packages or editors, needed to use the system?

1.15 Does the system provide security and data integrity protection features?

1.16 How many and what type of staff personnel will be required to develop a system from the software? To run the system? What type of training will they need and will the vendor provide that training?

Linkage Methodology

2.1 What record linkage method is the software based on?

Fellegi-Sunter?
Information-Theoretic methods?

2.2 How much control does the user have over the linkage process? Is the system a "black box," or can the user set parameters to control the linkage process?

2.3 Does the software require any parameter files? If so, is there a utility provided for generating these files? How effectively does it automate the process? Can the utility be customized?

2.4 Does the user specify the linking variables and types of comparisons?

2.5 What kinds of comparison functions are available for different types of variables? Do the methods give proportional weights (that is, allow degrees of agreement)?

Character-for-character?
Phonetic code comparison (Soundex or NYSIIS variant)?
Information theoretic string comparison function?
Specialized numeric comparisons?

Distance comparisons?
Time/Date comparisons?

Ad hoc methods (e.g., allowing one or more characters different between strings)?

User-defined comparisons?
Conditional comparisons?

2.6 Can the user specify critical variables that must agree for a link to take place?

2.7 How does the system handle missing values for linkage variables?

Computes a weight like any other value?
Uses a median between agreement and disagreement weights?
Uses a zero weight?
Allows user the option to specify treatment?

2.8 Does the system allow array-valued variables (e.g., multiple values for phone number)? How do array-valued comparisons work? What is the maximum number of values in an array?

2.9 What is the maximum number of linking variables?

2.10 How does the software block records? Do users set blocking variables? Can a pass be blocked on more than one variable?

2.11 Does the software support multiple linkage passes with different blocking and different linkage variables?

2.12 Does the software contain or support routines for estimating linkage errors?

2.13 Does the matching algorithm use techniques that take advantage of dependence between variables?

Fellegi-Sunter Systems

3.1 How does the system determine m- and u-probabilities? Can the user set m- and u-probabilities? Does the software provide utilities to set m- and u-probabilities.

3.2 How does the system determine weight cutoffs? Are they set by the user? Does the software provide any utilities for determining weight cutoffs?

3.3 Does the software allow linkage weights to be fixed by the user? What about weights for missing values?

Data Management

4.1 In what file formats can the software use data?

Flat file?

SAS Dataset?

Database? If yes, what kind of database?

Dbase?
Fox Pro?
Xbase?
Informix?
Sybase?
ORACLE?
Other database package?

4.2 What is the maximum file size (number of records) that the software can handle?

4.3 How does the software manage records? Does it use temporary data files or sorted files? Does it use pointers?

4.4 Can the user specify subsets of the data files to be linked?

4.5 Does the software provide for "test matches," of a few hundred records to test the specifications?

4.6 Does the software provide a utility for viewing and manipulating data records?

Post-linkage Functions

5.1 Does the software provide a utility for review of possible links? If so, what kind of functionality is provided for? What kind of interface does the utility use, character-based or GUI? Does the utility allow for review between passes, or only at the end of the process? Can more than one person work on the record review simultaneously? Can records be "put aside" for later review? Is there any provision for adding comments to the reviewed record pairs in the form of hypertext? Can pairs of groups of records be updated? Can the user “back up” or restore the possible links before committing to decisions? Can a “master” record be created which combines values from two or more records for different fields?

5.2 Does the software provide for results of earlier linkages (particularly reviews of possible links) to be applied to the current linkage process?

5.3 Does the software provide a utility for generating reports on the linked, unlinked, duplicate, and possible link records? Can the report format be customized? Is the report viewed in character mode, or is the report review done in a graphical environment? Can the report be printed? If so, what kind of printer is required?

5.4 Does the software provide a utility for extracting files of linked and unlinked records? Can the user specify the format of such extracts?

5.5 Does the software generate statistics for evaluating the linkage process? Can the user customize the statistics generated by the system?

Standardization

6.1 Does the software provide a means of standardizing (parsing out the pieces of) name and address fields?

6.2 Does the software allow for partitioning of variables to maximize the use of the information contained in these variables (for example, partitioning a phone number into area code, exchange, and the last four random digits)?

6.3 Can name and address standardization be customized? Can different processes be used on different files?

6.4 Does address standardization meet U.S. Postal Service standards?

6.5 Does standardization change the original data fields, or does it append standardized fields to the original data record?

6.6 How well do the standardization routines work on the types of names the user wishes to link?

6.7 How well do the standardization routines work on the addresses the user will encounter? (E.g., how well does it handle rural addresses? Foreign addresses?)

Costs

7.1 What are the purchase and maintenance costs of the software itself, along with any needed additional software (e.g., database packages), and new or upgraded hardware.

7.2 What will be the cost of training personnel to use the system.

7.3 What are the projected personnel and (in the case of mainframe systems) computer-time costs associated with running the system.

7.4 Is the cost of developing a system for the intended purposes using the software within the available budget?

Empirical Testing

8.1 What levels of false match and false nonmatch can be expected with the system? Are these levels acceptable?

8.2 How much manual intervention (e.g., possible match review) will the system require.

8.3 How rapidly can typical match projects be completed using the system?

Chapter 13

Software Demonstrations

MatchWare Product Overview

Matthew A. Jaro, MatchWare Technologies, Inc.

Probabilistic linkage technology makes it feasible to link large data files and achieve results governed by mathematical principles which adhere to statistically valid standards. The problem addressed by this methodology is that of matching two data files under conditions of uncertainty. The objective is to identify and link records which represent a common entity whether that entity is an individual, a family, an event, a business, an institution, or an address. As an alternative the goal might be to unduplicate a single data file or to group records by categories of commonality. Each field participating in the linkage comparison is subject to error which is measured by the probability that the field agrees given a record pair matches versus the probability of chance agreement of its values. Thus, when one calculates the likelihood of a correct match or link while allowing for incomplete and/or error conditions within the records, the process is said to be probabilistic. I. P. Fellegi and A. B. Sunter pioneered record linkage theory in the late 1950s. The first practical implementation of probabilistic linkage methodology in the United States was originally designed, programmed, and tested by Matt Jaro on behalf of the U. S. Census Bureau in 1985, while conducting research into establishing a model to support census coverage undercount evaluation and analysis.

Probabilistic record linkage methodology is imperative if computers are to consistently and effectively replicate the evaluation and judgment process of human clerks attempting to link common records. The ideal goal is to have the computer emulate the intuitive thought process of a human being as they might review, judge, evaluate, measure, and score linkage qualifications of records representing commonality.

MatchWare's development, systems design, and programming staff rigorously and strictly adhere to ANSI-C programming language standards for all software implementations. As a result, MatchWare software has achieved an exceptional level of cross-platform portability and can be integrated into a wide range of application solution specific systems. Following are the products currently offered by the company:

AutoStan is an intelligent pattern recognition parsing system which conditions records into a normalized/standardized fix fielded format. AutoStan optimizes the performance of any linkage or matching system which utilizes consumer or business names and/or address data as identifiers during a match comparison. AutoMatch is a state-of-the-art software implementation of probabilistic record linkage methodology for matching records under conditions of uncertainty. AutoMatch simulates the thought process a human being might follow while examining and identifying data records representing a common entity or event. AutoMatch's comparative algorithms manage a comprehensive range of data anomalies and utilize frequency analysis methodology to precisely discriminate weight score values.

AutoStan and AutoMatch are stand-alone, self-contained software systems which include numerous support utilities and require no other ancillary software. Both systems are generalized and support a wide range of mission critical record linkage applications. AutoStan and AutoMatch adhere to widely accepted standards of statistical methodology to ensure valid results and the highest levels of data integrity. Users have ready access to Rule/Table Portfolios in order to calibrate the software for their particular requirements. MatchWare/CL is a callable library (API) version of AutoStan and AutoMatch functionality in executable module form. MatchWare/CL utilizes AutoStan and AutoMatch Rule/Table Portfolios, weight scoring formulae, and statistical algorithms. MatchWare/CL is compatible with any database manage-

ment system or user interface, and has been integrated into a variety of application solution specific systems.

Both AutoStan and AutoMatch are generalized and support a wide range of mission critical health data registry, geocoding, and database marketing applications.

For more information, contact Max Eveleth Jr., Executive Vice President, MatchWare Technologies, Inc., 153 Port Road - 2nd Floor, Kennebunk, ME 04043-5135; Phone: (207) 967-2225; Fax: (207) 967-8362; or e-mail: *meveleth@matchware.com* .

■ :- and J-ARGUS: Software Packages for Statistical Disclosure Control

Anco J. Hundepool, Agnes Wessels and Lars van Gemerden, Statistics Netherlands

In recent years, Statistics Netherlands has developed a prototype version of a software package, ARGUS, to protect microdata files against statistical disclosure. The launch of the SDC-project within the 4th framework of the European Union had enabled us to make a new start with the development of software for Statistical Disclosure Control. More information on the SDC-project can be found at *http://www. cbs.nl/sdc.*

This prototype has served as a starting point for the development of :-ARGUS, a software package for the SDC of microdata files. The aim is to produce a data file for which the risk of disclosure has been minimized and which can be supplied to researchers and other users. The basic principle of :-ARGUS is that frequency tables of combinations of identifying variables are inspected. If the frequency in a cell is too low, it means that a certain combination does not occur frequently enough in the population and that the corresponding records, therefore, can easily be identified by an intruder. Techniques used in :-ARGUS to solve these problems are global recoding (using less detailed code lists) and local suppression (imputing missing values in these combinations).

This SDC-project, however, also plans to develop τ-ARGUS -- software devoted to the SDC of tabular data. τ-ARGUS takes the dominance-rule as a starting point to identify the unsafe (primary) cells, although other rules could be used, as well. Global recoding is applied to reduce most of the unsafe cells and optimization techniques are used to find a optimal set of secondary cells, which must be suppressed to protect the primary unsafe cells.

Both :- and τ-ARGUS have been developed for Windows 95 PC's. However, we have developed ARGUS using Borland C++, which raises the possibility of easily generating modules (the parts of ARGUS accessing large datafiles) to be used on other platforms like UNIX.

Further information can be obtained from Anco Hundepool, Department for Statistical Methods, Statistics Netherlands, P.O. Box 4000,2270 J.M. Voorburg, The Netherlands; tel: +31-70-3375038; fax: +31-70-3375990; or e-mail: *argus@cbs.nlofahnl@cbs.nl*.

■ OX-LINK: The Oxford Medical Record Linkage System Demonstration of the PC Version

Leicester E. Gill, University of Oxford, UK

The micro-computer version of OX-LINK is being used to match a dataset containing 150,000 hospital discharge and vital records. The matching and linking process is undertaken in three stages:

- The creation of an ONCA header, which is attached to every record on the dataset.
- Sorting the file on the keys which are stored in the ONCA header.
- Running OX-LINK to create a file of potential match pairs. A number of output files are produced which are used for verification of the match by clerical staff. The threshold weight matrix can be edited using Microsoft EDIT, and the whole of this stage can be rerun to demonstrate the changes in acceptance weight.

For more information, write to:

L. E. Gill
University of Oxford
Unit of Health-Care Epidemiology
Institute of Health Sciences,
Old Road, Headington, Oxford, OX37LF

or e-mail: *leicester.gill@clinical-epidemiology.ox.ac.uk* or *lester@pgme.warwick.ac.uk* .

■ Software for Record Linkage of Primary Care Data

John R. H. Charlton, Office of National Statistics, UK

The UK Royal College of General Practitioners collected data on all consultations in sixty practices in England and Wales over a one-year period 1991/92. In addition, socio-economic data were collected by survey from all patients registered with these practices. Each practice was sent a copy of its own data and the data from all the practices were combined into one dataset containing information on about 1.5 million consultations and about half a million patients.

The software demonstrated was written so that individual practices could easily access their own data, without specialised database software, or knowledge of the data structures and codes. Later, a modified program was written so that the Royal College of General Practitioners could extract data from the combined data from all practices. An anonymized version of the dataset was made available to other researchers and a further modified version of the program was produced for use with this dataset.

The program has two main functions. Firstly, to enable researchers to link different parts of the dataset, particularly patients and diseases, and secondly, to provide data summaries such as frequencies and rates. It is based on the Paradox database software and written in PAL, the language provided with Paradox for DOS. An installation program is provided to convert the ASCII files provided into the Paradox tables used by the program. The program can be run under either DOS or Windows.

For more information, contact Judith Charlton, 195 Warren Road, Orpington, Kent, BR6 6ES, U.K.; e-mail: *100025.1356@compuserve.com* .

GRLS -- Record Linkage

Kathy Zilahi, Statistics Canada

This product addresses the problem of trying to link records where no unique identifiers exist. Our Generalized Record Linkage System (GRLS) was developed to enable such problem linkages to be successfully accomplished. GRLS improves both the quality and the ease of your linkage.

Features

Based on statistical decision theory, GRLS breaks a linkage operation into three steps:

- Search: Using comparison rules and associated linkage weights, the files are matched and a database of potential links is created.
- Decide: Linkage weights are refined and by using threshold weights, the potential links are divided into sets of possible and definite links.
- Group: Records which pertain to the same entity (person, business, etc.) are grouped together (the output of GRLS).

The GRLS record linkage system:

- provides a convenient framework for testing linkage parameters;
- allows concurrent users for each linkage project;
- allows background or interactive linkage;
- eliminates confusion (and paper!) with on-line help;
- makes your final linkage fast, cheap and accurate.

Applicability

GRLS handles one-file (internal) and two-file linkages such as:

- unduplicating mailing address lists (one-file);
- bringing hospital admission records together to build "case histories" (one-file);
- epidemiology studies: e.g., linking a file of workers exposed to potential health hazards, to a mortality database for the purpose of detecting health risks associated with particular occupations (two-file).

Platform Specifications

GRLS uses a client-server architecture, where a PC is the client and a UNIX box is the server. The ORACLE relational database management system Version 7.3 with SQL*PLUS, PL*SQL, PRO/C, FORMS 4.5 runtime, GRAPHICS 2.5 runtime and a "C" compiler are also required. With ORACLE Version 7.3, distributed processing can easily be achieved by using either a remote or local host from a mainframe, mid-range computer, or PC.

Contact Information

For more information, contact Ted Hill, by phone: (613) 951-2394; fax: (613) 951-0607; or e-mail: *tedhill@statcan.ca*; or Bonnie Rideout, by phone: (613) 951-1714; fax: (613) 951-0607; or e-mail: *bburges@statcan.ca* .

Appendix: List of Attendees at the Record Linkage Workshop and Exposition March 20-21, 1997 Arlington, VA

Rebecca Adamson
Mathtech., Inc.

Yahia Ahmed
Internal Revenue Service

Elizabeth Ahuja
Department of Veterans Affairs

Lois Alexander
Consultant

Mark E. Allen
California Cancer Registry

Richard Allen
National Agricultural Statistics Service

Wendy Alvey
Internal Revenue Service

Otto Andersen
Statistics Denmark

Christy Anderson
Naval Health Research Center
San Diego, CA

Maxine Anderson-Brown
Bureau of the Census

Maria Rosario Araneta
Naval Health Research Center
San Diego, CA

Catherine Armington
Consultant

John Armstrong
Elections Canada

Faye Aziz
Social Security Administration

Paula C. Baker
The Ohio State University

A. John Bass
The University of Western Australia

Thomas Belin
UCLA School of Medicine

Julie Bernier
Statistics Canada

Jean-Marie Berthelot
Statistics Canada

A. Richard Bolstein
George Mason University

Kara Broadbent
National Agricultural Statistics Service

Mr. Alan Broder
White Oak Technologies

William Buczko
Health Care Financing
Administration

Frederick Buhr
Health and Family Services
Madison, WI

Dave Burhop
Mental Health and Mental Retardation
Richmond, VA

Robert Burton
National Center for Education Statistics

Carol Caldwell
Bureau of the Census

Maureen Carpenter
Statistics Canada

Mark Carrozza
University of Cincinnati

Roma Chappell
Office for National Statistics, UK

John Charlton
Office for National Statistics, UK

Cynthia Z.F. Clark
Bureau of the Census

Melvin E. Cole, III
Bureau of the Census

Larry Cook
University of Utah

Abbate Corrado
Italian Statistical Institute

Bob Cote
Canadian Institute for Health Information

Brenda G. Cox
Mathematica Policy Research, Inc.

Robert Creecy
Bureau of the Census

Catherine Cromey
Statistics Canada

John L. Czajka
Mathematica Policy Research, Inc.

Martin H. David
University of Wisconsin-Madison

Pierre David
Statistics Canada

Charles Day
National Agricultural Statistics Service

J. Michael Dean
University of Utah

Virginia A. deWolf
Bureau of Labor Statistics

Edma Diller
University of Utah

Cathryn Dippo
Bureau of Labor Statistics

Katarzyna Doerffer
AECL
Chalk River, Ontario CANADA

Patricia Doyle
Bureau of the Census

Judith A. Droitcour
General Accounting Office

Catherine Eginard
Eurostat
LUXEMBOURG

William Eilerman
Dept. of Housing &
Urban Development

M. Nabil El-Khorazaty
Research Triangle Institute

Timothy Evans
Bureau of the Census

Martha Fair
Statistics Canada

Ivan P. Fellegi
Statistics Canada

Charles M. Fleming
National Agricultural Statistics Service

John L. Fox
Wisconsin Bureau of Public Health

Gerhard Fries
Federal Reserve Board

Dave E. Galdi
Bureau of the Census

Gerald Gates
Bureau of the Census

Jane Gentleman
Statistics Canada

Leicester Gill
University of Oxford, UK

Robert E. Gillette
Treasury Department

Garofalo Giuseppe
Italian Statistical Institute

Frank Grabowiecki
Statistics Canada

Wayne B. Gray
Clark University

Nicholas Greenia
Internal Revenue Service

Robert Guernsey
Bureau of Labor Statistics

George Hanuschak
National Agricultural Statistics Service

Linda Hardy
National Science Foundation

Lin Hattersley
Office of National Statistics, UK

Marta Haworth
Office for National Statistics, UK

Sigurd Hermansen
Westat, Inc.

Thomas N. Herzog
Dept. of Housing and Urban Development

Shiu Man Ho
University of Maryland at Baltimore

John Horm
National Center for Health Statistics

Christian Houle
Statistics Canada

Elizabeth Huang
Bureau of the Census

Larry Huff
Bureau of Labor Statistics

Anco Hundepool
Statistics Netherlands

Alice Hung
City of Houston Planning & Development

Donsig Jang
Mathematica Policy Research, Inc.

Matthew Jaro
MatchWare Technologies, Inc.

Paul Jelfs
Australian Institute of Health & Welfare

Barry W. Johnson
Internal Revenue Service

Sandra Johnson
National Highway Traffic Safety Administration

Vickie L. Kee
Bureau of the Census

Steve Kendrick
Scottish Health Service

Arthur Kennickell
Federal Reserve Board

Tim Kerns
University of Maryland at Baltimore

Jay Jong-IK Kim
Bureau of the Census

Jeong Kim
Bureau of the Census

Karl Kim
University of Hawaii

Nancy Kirkendall
Office of Management and Budget

David Klein
RAND Corporation

Matthew A. Koch
Research Triangle Institute

David Koepke
University of Chicago

Selma Kunitz
Kunitz and Associates, Inc.

Tony LaBillois
Statistics Canada

Aarno Laihonen
Statistics Finland

Eric Langlet
Statistics Canada

Michael Larsen
Stanford University

Eric Larson
General Accounting Office

Robin Lee
Internal Revenue Service

Charlene A. Leggieri
Bureau of the Census

Larry Lie
Bureau of Labor Statistics

A. Russell Localio
Pennsylvania State University

Marcella Loftus
Shared Medical Systems

Susan B. Long
Syracuse University

Steven Macdonald
Centers for Disease Control

Traci Mach
The Ohio State University

Steven Machlin
Agency for Health Care Policy & Research

Joann S. Mack
Health Care Financing Administration

Jennifer Madans
National Center for Health Statistics

Kent H. Marquis
Bureau of the Census

David J. McDonell
National Agricultural Statistics Service

Patricia McGuire
General Accounting Office

Elspeth McVey
Office for National Statistics, UK

Ramesh Menon

Scott Meyer
Statistics Canada

Gordon Mikkelson
Bureau of Labor Statistics

Eva Miller
New Jersey Department of Education

Kimberly S. Miller
University of Pennsylvania

Nash J. Monsour
Bureau of the Census

Richard Moore
Bureau of the Census

Chris Moriarity
National Center for Health Statistics

William A. Morrill
Mathtech, Inc.

Kirk Mueller
Bureau of Labor Statistics

Edward Mulrow
National Opinion Research Center

Cam Mustard
Statistics Canada

Patricia Nechodom
University of Utah

Randall Neugebauer
Bureau of the Census

Lawrence Nitz
University of Hawaii

Fiona O'Brien
Scottish Health Service

Karen O'Conor
Internal Revenue Service

Philip Parsons
Ontario Cancer Treatment & Research Fdn

Roberta Pense
National Agricultural Statistics Service

Michael Pergamit
National Opinion Research Center

Jay Pfeiffer
Florida Department of Education

Rich Pinder
Los Angeles Cancer Surveillance Program

Timothy Pivetz
Bureau of Labor Statistics

Jessica Pollner
Price Waterhouse, LLP

Edward Porter
Bureau of the Census

Adam Probert
Statistics Canada

Jonathan Price

Shruti Rajan
The Urban Institute

James Reading
University of Utah

Martha Farnsworth Riche
Bureau of the Census

Kenneth W. Robertson
Bureau of Labor Statistics

Leslie Roos
Statistics Canada

Wendy Rotz
Internal Revenue Service

Donald Rubin
Harvard University

John M. Ryan
Talbot County, MD Health Department

Peter Sailer
Internal Revenue Service

Susan R. Sama
Dept. of Labor & Industries
Olympia, WA

Douglas A. Samuelson
InfoLogix

Tony Santiago
Standard & Poors Corporation

Fritz Scheuren
Ernst and Young, LLP

Mark A. Schipper
Energy Information Administration

Karen Schlangen
Anteon Corporation

Abdul Hannan Shaikh
Center for Women and Child Development
BANGLADESH

Tiefu Shen
Illinois Department of Public Health

Carolyn Shettle
National Science Foundation

Heidi Shierholz
Bureau of Labor Statistics

Gerald Silverstein
Department of Treasury

Eleanor Singer
University of Michigan

Cotty Smith
Bureau of the Census

Michael G. Smith
Link Soft Technologies, Inc.

Edward J. Sondik
National Center for Health Statistics

Edward J. Spar
Council of Professional Associations
on Federal Statistics

Philip M. Steel
Bureau of the Census

Donald D. Stockford
Department of Veterans Affairs

Lynne Stokes
University of Texas at Austin

Latanya Sweeney
Massachusetts Institute of Technology

Kenneth H. Szeflinski
Internal Revenue Service

Yves Thibaudeau
Bureau of the Census

Robyn Thoelke
National Agricultural Statistics Service

John Tibert
York University

John R. Tucker
National Research Council

Dennis Utter
National Highway Traffic
Safety Administration

John Van Voorhis
University of Chicago

Leah Vaughn
San Francisco, CA

Lydia Voti
University of Miami

Jenny B. Wahl
St. Olaf College

Claus Wall
Institute for Clinical Evaluation Sciences
North York, Ontario CANADA

David Wallace
Statistics Canada

Katherine Wallman
Office of Management and Budget

Michael Weber
Internal Revenue Service

Paula Weir
Energy Information Administration

Sandra West
Bureau of Labor Statistics

Sandy West
Investment Company Institute

Michael Westland
Statistics Canada

Andrew A. White
National Research Council

Patricia Whitridge
Statistics Canada

Bruce Whyte
Argyll & ClydeHealth Board
Paisley, SCOTLAND

Brian Wiersema
University of Maryland

Leon Willenborg
Statistics Netherlands

Marianne Winglee
Westat, Inc.

Alice Winkler
Bureau of Labor Statistics

William E. Winkler
Bureau of the Census

Michael Wolfson
Statistics Canada

David Woodrow
Standard & Poors Corporation

Tommy Wright
Bureau of the Census

Marek Wysocki
Statistics Canada

David Yu
University of Chicago

Elaine Zanutto
Harvard University

Alvan Zarate
National Center for Health Statistics

Laura Zayatz
Bureau of the Census

Kathleen Zilahi
Statistics Canada