Utilizing Privacy Preserving Record Linkage to Link Data from Two Federal Statistical Agencies

Cordell Golden, National Center for Health Statistics

Lisa B. Mirel, National Center for Science and Engineering Statistics

Jason Meyer, HealthVerity

Rui Wang, Mathematica, Inc

Federal Committee on Statistical Methodology (FCSM)

Research and Policy Conference

October 22, 2024

Outline

- Need for Privacy Enhancing Technologies (PETs)
- National Center for Health Statistics (NCHS) Data Linkage Program's Exploration of PETs
- National Secure Data Service (NSDS) Demonstration Project: Utilizing Privacy Preserving Record Linkage (PPRL) to Link Data from Two Federal Statistical Agencies
  - Goals & Data Sources
  - Data Flow
  - Analytic Opportunities & Limitations
  - Current Status & Next Steps
  - Lessons Learned

Need for Privacy Preserving Technologies (PETs)

- With the passage of the Foundations for Evidence-Based Policymaking Act in 2018, increasing access to data to make informed policy decisions has come to the forefront

- Providing secure access to these data has posed challenges

- Privacy enhancing technologies, such as privacy-preserving record linkage (PPRL), offer tools and methods to mitigate privacy risks with confidential or restricted data, thus enabling new opportunities for:
  - Linkage
  - Dissemination
  - Analysis

- The National Secure Data Service (NSDS) Demonstration project provides the opportunity to assess these tools and methodologies to inform future initiatives

NCHS Data Linkage Program: Exploration of PETs

NCHS Data Linkage Program:

- Links data from health surveys to administrative data sources to support high quality research and program evaluation

- Conducts linkages using deterministic and probabilistic methods that compare clear-text identifiers from two data sources
  - Concerns over data privacy and security limit data sources available and potential linkage opportunities

- Has conducted evaluations of selected PPRL applications to assess performance and accuracy
  - Evaluations to date have been based on *PPRL simulations* conducted within NCHS's secured computing environment
  - Several key roles involved in the PPRL process, including:
    - Data owners
    - Trusted third party (optional)
  - For the evaluations, NCHS served in multiple roles

NSDS Demonstration Project:
NCHS-NCSES PPRL

NCHS-NCSES PPRL

NCHS-NCSES PPRL:

Utilizing Privacy Preserving Record Linkage to Link Data from Two Federal Statistical Agencies (NCSES/NCHS) (https://www.americasdatahub.org/award-pprl1/)

- Supports evidence building opportunities and efforts to incorporate privacy enhancing technologies into the NSDS

- Provides an opportunity for a *real-world* application of PPRL
    - External linkage partners (two federal statistical agencies)
    - Linkage conducted outside of both agencies' secured boundaries
    - Utilization of trusted-third party (FedRAMP authorized PPRL tool)

NCHS-NCSES PPRL: Goals & Data Sources

Goals:

1. Develop data sharing agreement to link data from two federal statistical agencies
2. Establish processes for utilizing a commercial PPRL tool: **health**verity
3. Inform analytic initiatives
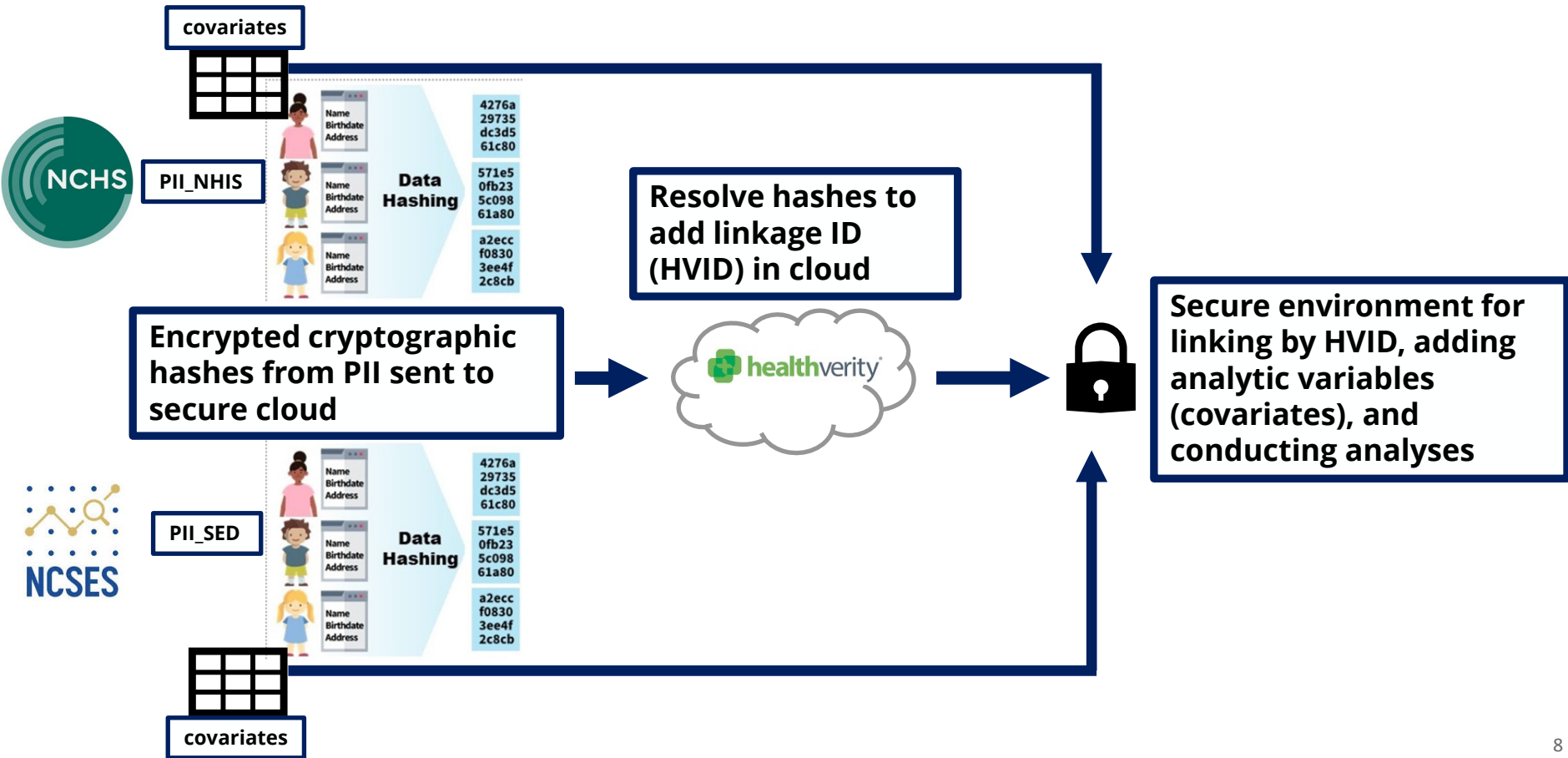
Data Sources:

National Center for Health Statistics (NCHS)

National Health Interview Survey (NHIS)

National Center for Science and Engineering Statistics (NCSES)

Survey of Earned Doctorates (SED)

NCHS-NCSES PPRL: Data Flow

covariates

PII_NHIS

Resolve hashes to add linkage ID (HVID) in cloud

Encrypted cryptographic hashes from PII sent to secure cloud

PII_SED

covariates

Secure environment for linking by HVID, adding analytic variables (covariates), and conducting analyses

NCHS-NCSES PPRL: NHIS and SED

NHIS participants who earned a research doctorate from an accredited U.S. institution

NCHS-NCSES PPRL: Evaluation of Disability Questions

NHIS	SED
Do you have difficulty: • seeing, even when wearing glasses or contact lenses? • Hearing, even when using your hearing aid (s) • Walking or climbing stairs • Communicating, for example, understanding or being understood • Remembering or concentrating • Self care, such as washing all over or dressing?	What is the usual degree of difficulty you have with: • Seeing word or letters in ordinary newsprint (with glasses or contact lenses) • Hearing what is normally said in conversation with another person (with hearing aid) • Walking without human or mechanical assistance or using stairs • Lifting or carrying something as heavy as 10 pounds • Concentrating, remembering, or making decisions because of a physical, mental or emotional condition
➢ No difficulty	➢ None
➢ Some difficulty	➢ Slight
➢ A lot of difficulty	➢ Moderate
➢ Cannot do at all	➢ Severe
	➢ Unable to do

NCHS-NCSES PPRL: Analytic Opportunities and Limitations

- **Opportunities:**
  - Compare composite estimates from each survey based on disability questions for those who linked
  - Directly compare questions and responses by the same individuals
- **Additional analyses:**

  Examination of selected health-related factors for NHIS participants who earned a research doctorate, including:
  - Sleep time, sleep quality
  - Frequency of physical activity
  - Life satisfaction
  - Mental health
- **Limitations:**
  - Timing when questions are asked
  - Small subset of population

NCHS-NCSES PPRL: Current Status & Next Steps

Objective	Tasks
1. Develop data sharing and governance agreement	✓ Review and sign software license agreement ✓ Obtain approval to conduct linkage in a commercial cloud environment with FedRAMP ✓ Obtain final signatures on data sharing agreement
2. Utilize a PPRL tool	❑ Standardize PII fields to create encrypted cryptographic hashes
3. Create an analytic dataset	❑ Link/analyze data in secure environment

NCHS-NCSES PPRL: Lessons Learned to Date

- Establish and engage key stakeholders and champions at both data sharing institutions early in the process

- Develop understanding of processes and technical requirements to use commercial PPRL tool in cloud environment

- Allow sufficient time for finalizing data sharing agreements

Conclusion

NCHS-NCSES PPRL:

- Supports the Evidence Act and the establishment of a National Secure Data Service

- Informs future linkages across the federal government by:
  - Providing a prototype for data sharing agreements for federal statistical agencies
  - Informing the development of a secure computing environment for use of PPRL tools and analysis of resulting linked data

- Serves as a model for improving the efficiency, availability, quality, accessibility, and interoperability of data sharing in the federal government while preserving privacy

Questions?

Cordell Golden

Chief, Data Linkage Methodology and Analysis Branch
National Center for Health Statistics
Centers for Disease Control and Prevention
CGolden@cdc.gov

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348   www.cdc.gov