USING SYNTHETIC DATA TO REDUCE DISCLOSURE RISK IN LOCAL HEALTH SURVEYS

Wen Qin Deng

New York City Department of Health and Mental Hygiene

2024 FCSM Research and Policy Conference

October 23, 2024

NYC Health

Project Team Members

Stephen Immerwahr
Tashema Bholanath
Wen Qin Deng
Nneka Lundy De La Cruz
Fangtao He
Jingchen (Monika) Hu

Acknowledgement:

Amber Levanon-Seligson
Steven Fernandez
Sungwoo Lim

NYC Health

TALK ROADMAP

01

BACKGROUND

02

DISCLOSURE RISK ASSESSMENT

03

MITGATION SOLUTIONS & RESULTS

04

SUMMARY & TAKEAWAYS

01

BACKGROUND

Health

PUBLIC-USE DATA: USEFULNESS AND CHALLENGES

Public-use data files are extremely valuable

Disclosure risks may exist in the public release of record-level survey data

e.g., potential linkage to administrative database (vaccine, etc.)

PROJECT OBJECTIVES

- Systematic methods to:
  - EVALUATE DISCLOSURE RISKS
  - IMPLEMENT MITIGATION SOLUTIONS
  - EVALUATE UTILITY AND RISK REDUCTION

NYC COMMUNITY HEALTH SURVEY (CHS) OVERVIEW

Annual cross-sectional health surveillance survey of ≈ 10,000 NYC adults

Monitors progress towards citywide health initiatives and other core surveillance efforts

Collects information including health status, mental health, healthcare access, chronic diseases, health and risk behaviors, and social determinants of health

NYC Health

02

DISCLOSURE RISK ASSESSMENT

APPROACH OVERVIEW

- Assume intruder knows a combination of identifying variables of each record

- Evaluate disclosure risk of all confidential survey records
  - **Core variables** – demographic variables
  - **Key variables** – demographic and health-related variables that are easily knowable

- Identify "high-risk" survey records using identifying categorical variables and sampling weights
  - **Weighted populations (Weighted N) and 95% Confidence Intervals (CIs)**

NYC Health

IDENTIFYING HIGH-RISK RECORDS

Core + Key variable

(one key at a time)

- Age Group **x** Sex **x** Race/ethnicity **x** Borough **x** **Key Variable A**
- 25 key variables identified elevated risk of re-identification

Weighted N less than 100 in the lower bound of 95% CIs are flagged as high-risk

- **Weighted N method** - i.e., the estimated population of these records in this combination are less than 100 in NYC
- **4%-24%** (of all observations) with elevated risk of re-identification

03

MITIGATION SOLUTIONS & RESULTS

NYC Health

DATA SYNTHENSIS IN RECENT LITERATURE

Dirichlet Process Mixture of Multinomial Distributions Model

Hu et al. (2014), Drechsler & Hu (2021)

- Bayesian nonparametric procedure
- Multinomial distribution; Latent class model
- Easy implementation, Synthesis order doesn’t matter
- R Package: NPBayesImputeCat

Classification and Regression Trees

Reiter (2005), Drechsler & Hu (2021)

- Nonparametric, based on a machine learning algorithm
- Regression trees
- Easy implementation, Application to categorical and continuous variables
- R Package: CART

NYC Health

MITIGATION SOLUTIONS

Suppression

- Applied to key variables for <u>a subset</u> of high-risk records
- **Method:** set values to missing
- Introduces additional missing values in dataset, *and thus not considered*

Partial Synthesis – **DPMPM**

- Applied to key variables for <u>a subset</u> of high-risk records
- **Method:** synthesize new value using nonparametric Bayesian models

Partial Synthesis – **CART**

- Applied to key variables for <u>a subset</u> of high-risk records
- **Method:** synthesize new value using classification and regression models

OVERVIEW OF OUR APPROACH

MITIGATION SOLUTION

CONFIDENTIAL DATA

Partial data synthesis with DPMPM – a Bayesian latent class model

SUBSET OF HIGH-RISK RECORDS

SYNTHETIC DATA

Determine a threshold (e.g., 5% of all records)

A randomized process

EVALUATE DATA UTILITY

PROTECTED PUBLIC-USE DATA

95% CI Overlap and MSE of prevalence estimates of selected health measures

RISK RESULTS AFTER DPMPM SYNTHESIS

Before Synthesis

4% to **24%** high-risk observations of all observations

4% - 24%

After Synthesis

At most **21%** of the dataset remains classified as high-risk (i.e., at least **79%** protection)

Synthesis-at-most-5% method

21%
(At most)

Note: among 25 key variables selected in the 2021 CHS

DATA UTILITY RESULTS AFTER DPMPM SYNTHESIS

With synthesis-at-most 5%

94%

overlap in the 95% confidence intervals of important health measures, on average

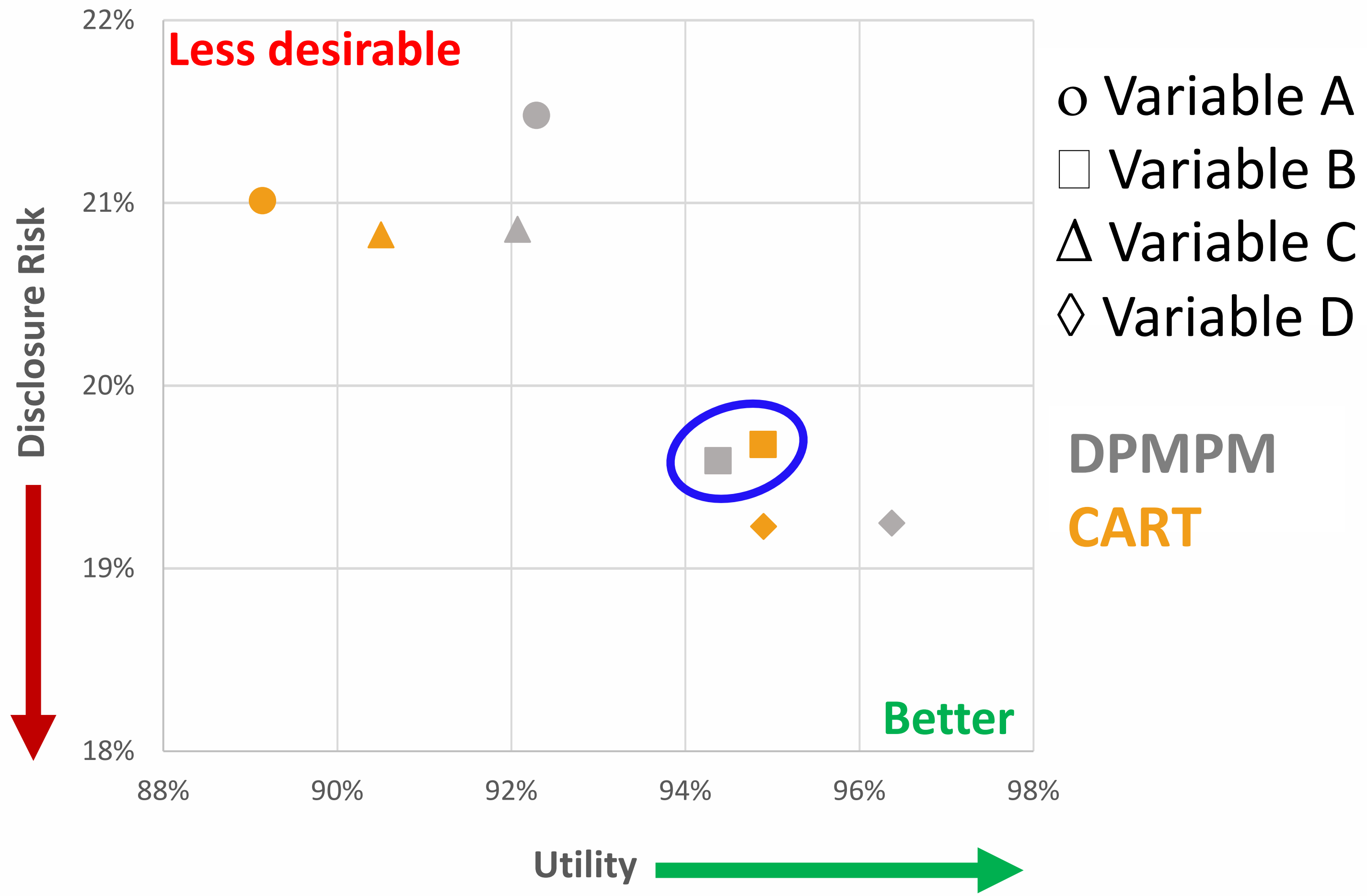
Best in balancing risk reduction and utility preservation

SYNTHESIS	Thresholds: 5%, 10%, 20%
95% CONFIDENCE INTERVAL OVERLAP	Cutoffs: 50%, 75%, 90%
HEALTH MEASURES	10 commonly reported

NYC Health

RESULTS COMPARISON: DPMPM VS CART

- **Disclosure risk (*y*-axis)**
  - % of high-risk after mitigation
  - Smaller means lower risk
- **Utility (*x*-axis)**
  - 95% CIO of health outcomes before and after mitigation
  - Larger means higher utility
- Utility-risk trade-off
- Final choice
  - **DPMPM (overall higher utility** at the price of slightly higher **disclosure risks**)



04

SUMMARY AND TAKEAWAYS

NYC Health

SUMMARY AND KEY TAKEAWAYS

1. **Weighted N** is a useful approach to quantify risk
2. Considerations in choosing the **appropriate synthesis approach**
3. Any mitigation presents a **risk-utility trade-off**
4. Multiple considerations in **setting parameters**

NYC Health

REFERENCES

Drechsler, J. (2011), Synthetic Datasets for Statistical Disclosure Control, Springer: New York.

Grant-Chapman, H. and Vallee, H. Q. (2022), Making government data publicly available: guidance for agencies on releasing data responsibly, Center for Democracy and Technology.

Hu, J., Reiter, J. P., and Wang, Q. (2014), Disclosure risk evaluation for fully synthetic categorical data, Privacy in Statistical Databases, J. Domingo-Ferrer (ed), 185-199.

Drechsler, J. and Hu, J. (2021), Synthesizing geocodes to facilitate access to detailed geographical information in large-scale administrative data, Journal of Survey Statistics and Methodology, 9(3), 523–548.

Little, R. J. A. (1993), Statistical analysis of masked data, Journal of Official Statistics, 9(2), 407-426.

Reiter, J. P. (2005), Using CART to generate partially synthetic public use microdata, Journal of Official Statistics, 21, 441-462.

Reiter, J. P. and Mitra, R. (2009), Estimating risks of identification disclosure in partially synthetic data, The Journal of Privacy and Confidentiality, 1, 99-110.

Simon, G., Shortreed, S. M., Coley, R. Y., Iturralde, E. M., Platt, R., Toh, S., and Ahmedani, B. (2020), Toolkit for assessing and mitigating risk of re-identification when sharing data derived from health records, Sentinel.

Snoke, J., Raab, G. M., Nowok, B., Dibben, C., and Slavkovic, A. (2018), General and specific utility measures for synthetic data, Journal of Royal Statistical Society, Series A, 181, 663-688.

NYC Health

Thank You!

Contact:

Wen Qin Deng

wdeng@health.nyc.gov

Health