Disclaimer: The views expressed in this presentation are those of the author and do not necessarily reflect the views of the National Center for Science and Engineering Statistics or the National Science Foundation

Current Approaches to Privacy Protection with Blended Data: Health

Lisa B. Mirel

May 22, 2023

CNSTAT Workshop: Approaches to Sharing Blended Data in a 21st Century Data Infrastructure

National Center for Science and Engineering Statistics

Social, Behavioral and Economic Sciences

National Science Foundation

Data Linkage

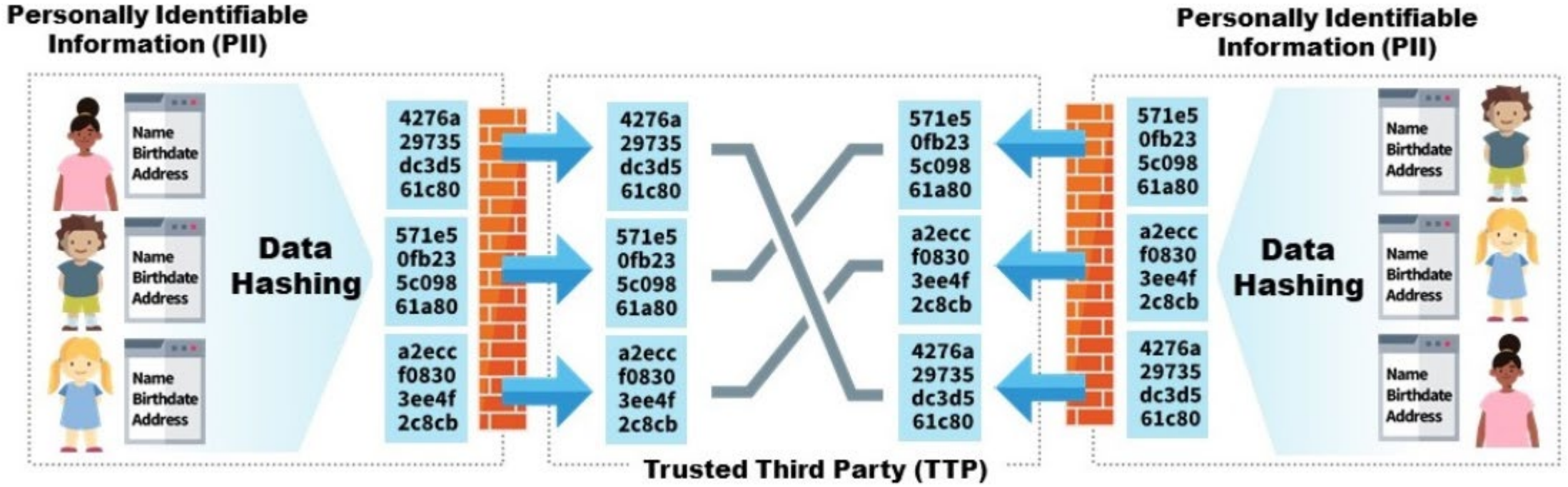
- Linking (blending) data is a powerful and efficient mechanism for producing policy-relevant information

- Privacy concerns impact the
  - Sources used for linkage
  - Access to blended data

Safeguarding Privacy with Blended Data

- This talk will focus on two areas
  - Safeguarding privacy through privacy-preserving record linkages (PPRL): assessing the quality of PPRL compared with results obtained through clear text matching
  - Increasing access through synthetic data: partially synthetic linked mortality files and the creation of fully synthetic linked data files with verification metrics

Privacy Preserving Record Linkage



Source: Clinical and Community Data Initiative | Overweight & Obesity | CDC

Privacy preserving techniques have increased the potential to expand data sharing while reducing privacy concerns

PPRL is a method that can be used to link de-identified data using hashing algorithms and tokens

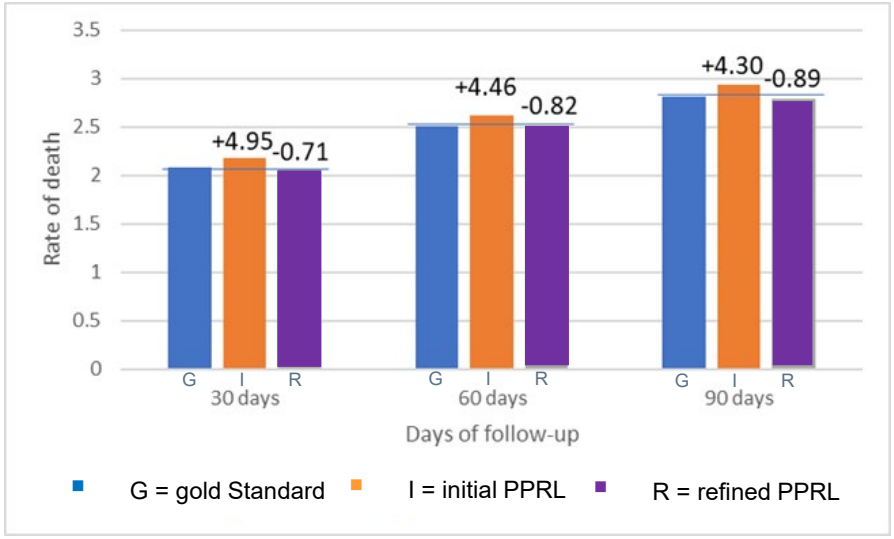
Case Study: Clear Text Matching vs. PPRL

- How do results from PPRL compare with results from clear text matching (matching using direct identifiers)?
- National Hospital Care Survey data linked to death certificate information from the National Death Index
  - Assess PPRL compared with “gold standard” (clear text matching)
    - Initial PPRL and refined PPRL (removed tokens that resulted in links with false positive rate > 50%)
  - Determine the quality of PPRL linked data (recall and precision)
  - Compare estimates in secondary analysis

Results

- Depending on the selection of tokens from PPRL (initial or refined):
  - Precision: 93.8% to 98.9%
  - Recall: 98.7% to 97.8%
- Impact of PPRL links on secondary data analysis was minimal

Death Rates by Linkage Approach and Follow-Up Period and Percent Difference from Gold Standard



Mirel, Lisa B. et al. 2022. *'A Methodological Assessment of Privacy Preserving Record Linkage Using Survey and Administrative Data*'. Stat Journal of the IAOS. 1 Jan: 413–21. DOI: 10.3233/SJI-210891

Summary and Next Steps

- PPRL can be an effective record linkage technique that produces results similar to clear text matching
  - High precision and recall estimates
- Explore accuracy of PPRL
  - When using sources with less complete personally identifiable information
  - For subpopulations
- Assess the use of PPRL to expand data linkage activities
- Establish a trusted third party or honest broker

Increasing Access: Synthetic Data

- Once sources are linked, there is an increased re-identification risk because of the amount of detail added
- Most linked products are available through a restricted access environment (e.g., Federal Statistical Research Data Centers)
- Methods are being developed to create data files that are more accessible while maintaining analytic utility and continuing to protect privacy

Creating Synthetic Data Files

- Partially synthetic: noise is added to the file; there is a one-to-one mapping between real individuals and the partially synthetic records

- Fully synthetic: there is no direct mapping between a synthetic record and real individual records

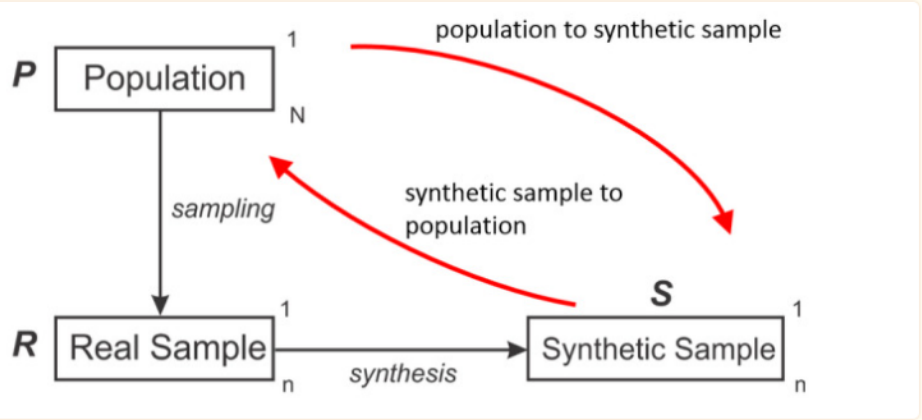
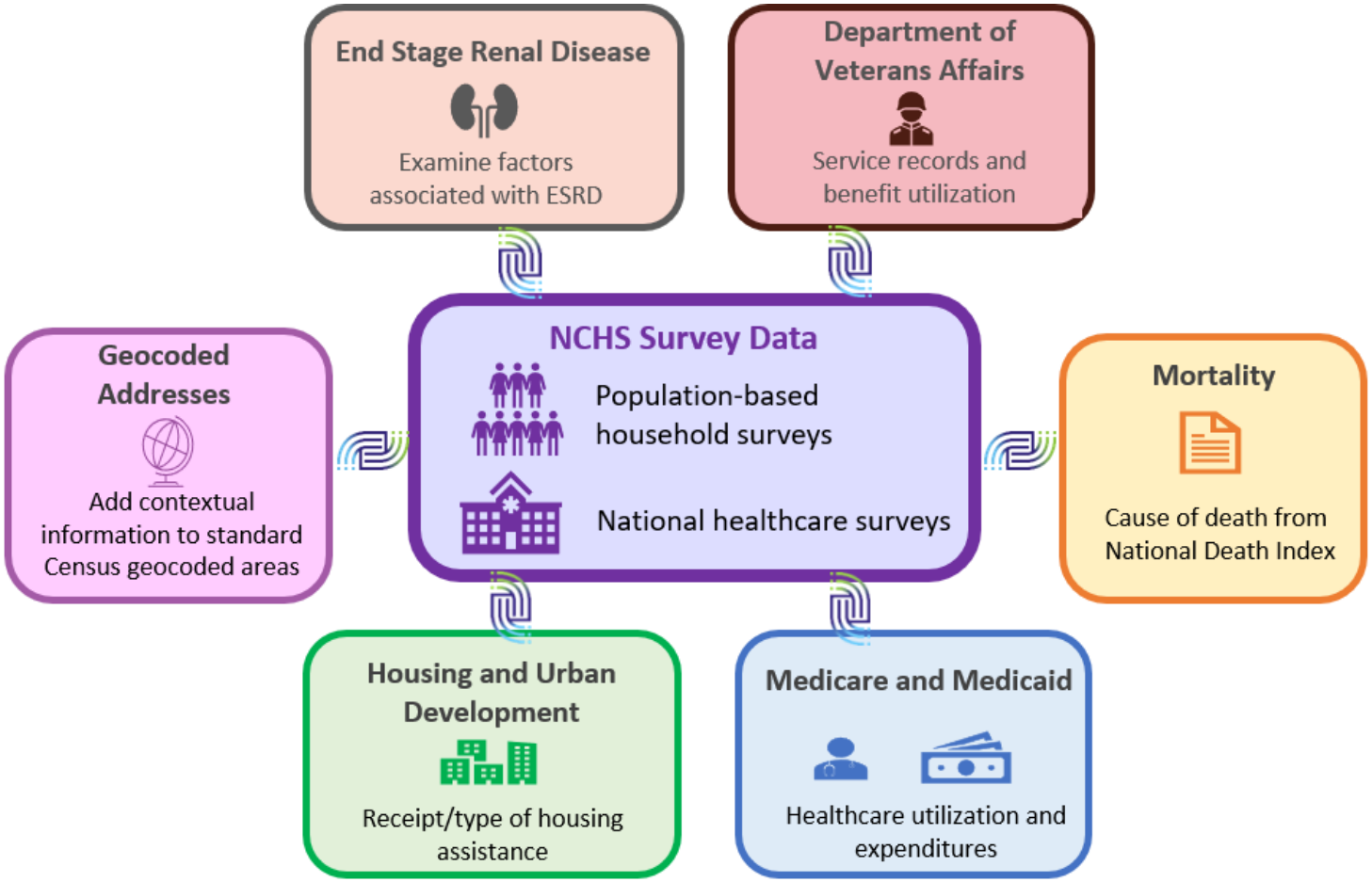


Figure 1

The relationships between the different datasets under consideration. Matching between a synthetic sample record and someone in the population goes through the real sample and can occur in 2 directions.

El Emam K, Mosquera L, Bass J. 2020. *Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation*. Journal of Medical Internet Research 16 Nov;22(11):e23139. DOI: 10.2196/23139. PMID: 33196453; PMCID: PMC7704280.

National Center for Health Statistics (NCHS): Data Linkage Program

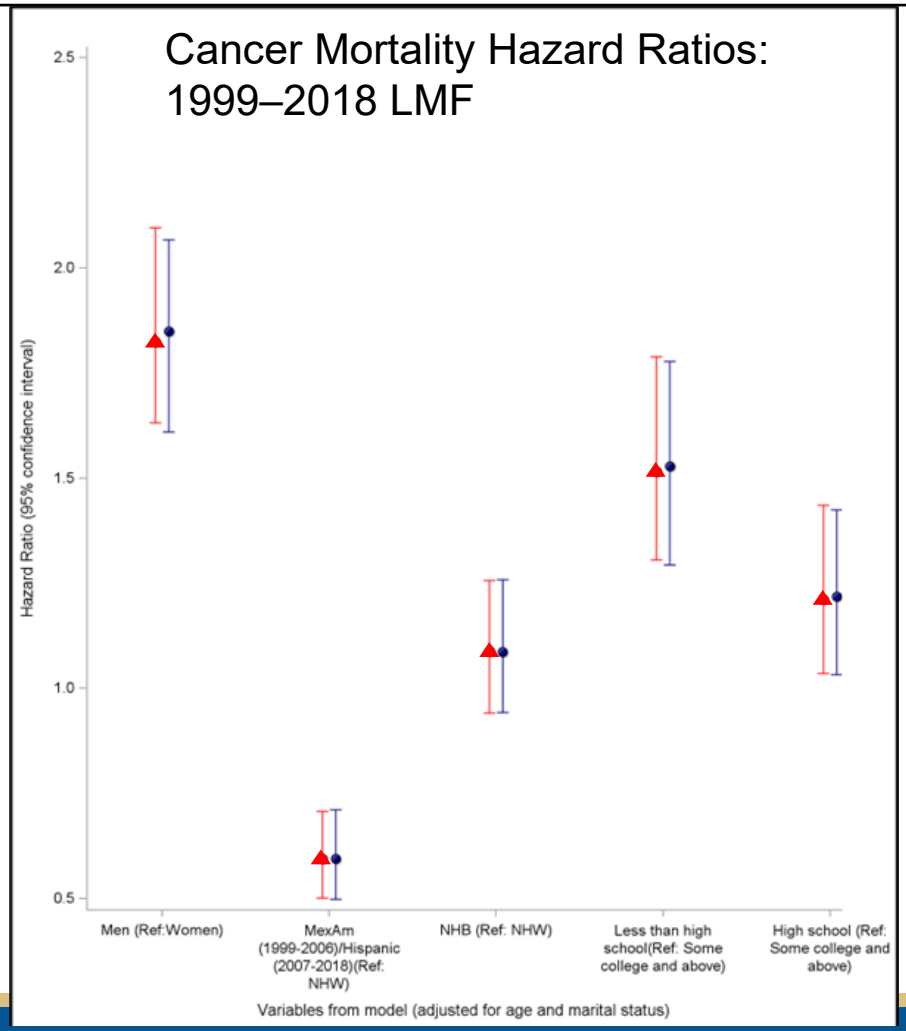


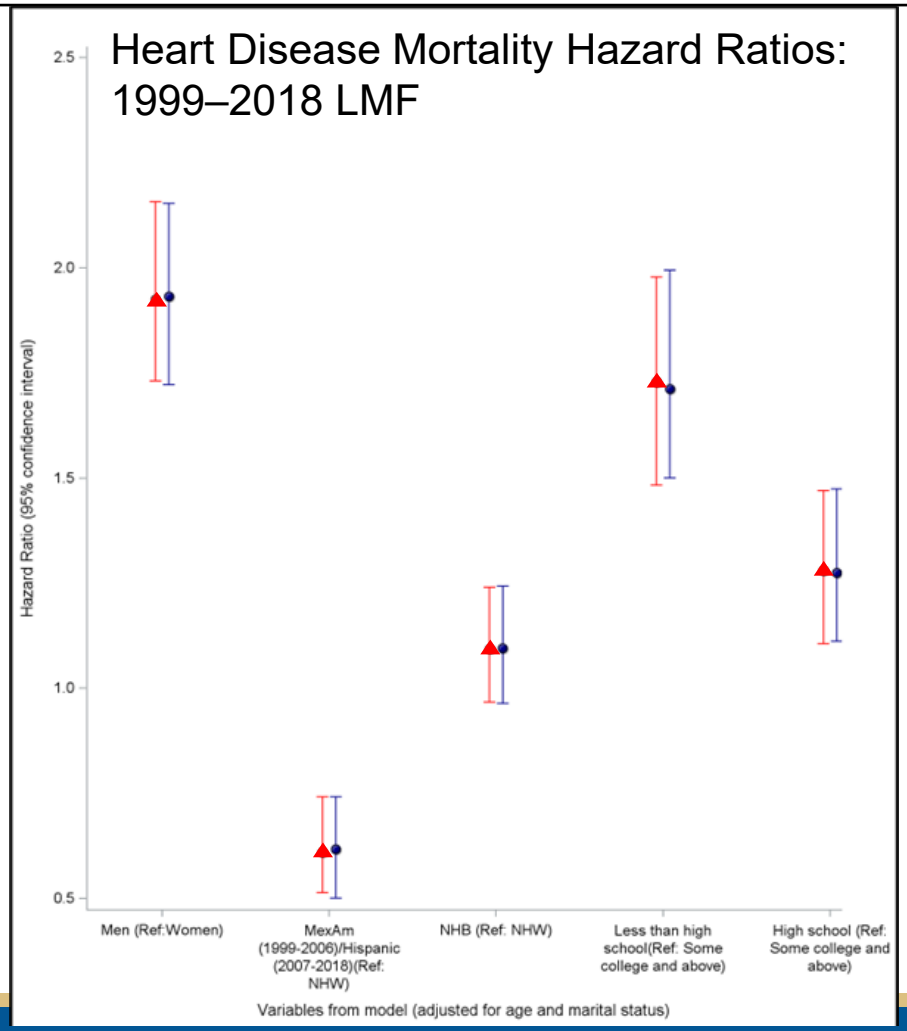
Partially Synthetic Linked Mortality Files (LMF)

File details	Restricted use	Partially synthetic public use
Availability	Requires an approved data use agreement or research data center proposal	Available for public download from data linkage website
Dates	Exact date of death, birth date, and interview date	Date of death represented by quarter and year or by person months of follow up, depending on survey
Cause of death	Detailed underlying and multiple code of death information	Most common underlying cause of deaths and two indicators for multiple cause of death, diabetes and hypertension
Participants	Both adults and children	Only adults
Perturbation	No perturbation	Perturbed information for cause of death or follow up time for select decedents; vital status is not perturbed

Comparing LMF Hazard Ratios for Restricted and Public-use (Partially Synthetic) Data

Mirel LB, et al.2021 *Comparative analysis of the National Health and Nutrition Examination Survey public-use and restricted-use linked mortality files*. National Health Statistics Reports; no 155. Hyattsville, MD: National Center for Health Statistics. DOI: https://doi.org/10.15620/cdc:104744.





- Restricted use LMF
- Public-use LMF

Increasing Access: Fully Synthetic Linked Data

- Data Linkage Program at NCHS is piloting innovative methods to create public-use synthetic linked data files, which are based on the true restricted data
- Create publicly available linked synthetic datasets
  - Conduct comparison analyses
  - Establish a verification system
- Develop an interactive data visualization tool to further increase accessibility and utility of evidence building linked data

Methods for Creating Fully Synthetic Linked Files

- SimPop package in R for sample design variables
- Synthpop package: classification and regression trees for all other variables
- Assess creation of multiple implicates for variance estimation
- Conduct disclosure risk assessment

Preliminary Results

Compared percent distributions for synthetic and true variables

- Initial assessment shows close alignment between true and synthetic data

Evaluated logistic regression model coefficients and confidence intervals

- Creating multiple implicates improves alignment of standard errors

Still early in the process but synthetic generation methods seem to be working well

Important to note that can't test every model

Proposed Verification Metrics

Decision agreement

- Are the signs of the coefficients the same (+ or -) and do both have p-values < 0.05 or > = 0.05?

Estimate agreement

- Does the synthetic data estimate fall within the confidence interval of the true data?

Confidence interval percent overlap

- What is the percent overlap of the confidence intervals?

Opportunities and Challenges

- Opportunities:
  - A tiered access approach of both publicly available and restricted data resources creates opportunities to expand data access and research potential
  - Researchers and policymakers can utilize synthetic linked data to address emerging public health threats and develop hypotheses
- Challenges:
  - Ensuring synthetic data methodologies produces analytically sound estimates
  - Communicating the importance of using implicates and verifying the results
  - Evaluating and improving synthetic data generator for sustainability as new data are linked

Final Thoughts

- Safeguards are needed to blend data and analyze blended data
  - Privacy enhancing technologies allow for combining sources without sharing identifiers, but a trusted third party is needed
  - Synthesizing data allows for tiered access, but ensuring its utility and accuracy is imperative
- When done correctly, privacy enhancing technologies can support vast research opportunities

Contact: lbmirel@nsf.gov