National Center for Health Statistics
Data Linkage

Privacy Preserving Techniques: Case Studies from the Data Linkage Program

Lisa B. Mirel

Board of Scientific Counselors

May 19, 2021

Introduction

- Linking data is a powerful mechanism to provide policy relevant information in an efficient way

- NCHS currently links data from several population-based and establishment surveys to administrative data sources, but privacy concerns impact:
  - Data that can be released
  - Sources used for linkage

- Two case studies attempt to address these concerns:
  - Synthetic data creation for public release
  - Privacy preserving record linkage

Data Sources for Case Studies

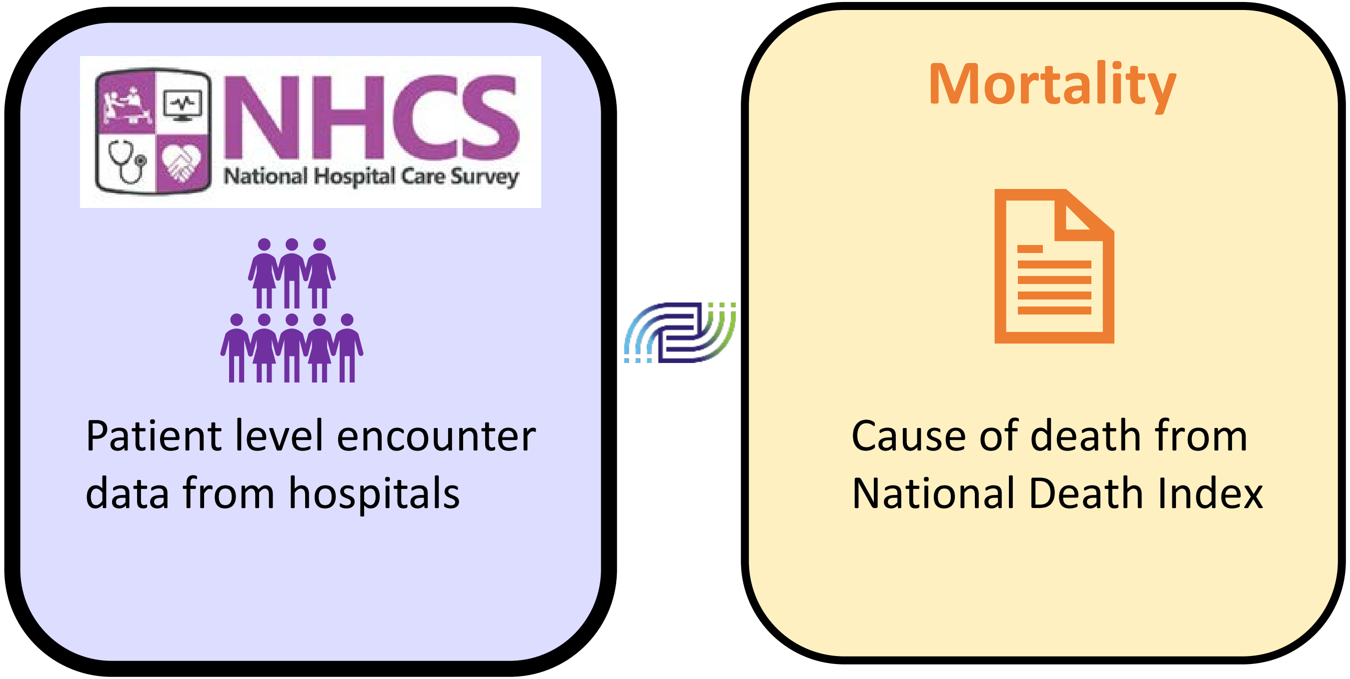
NHCS
National Hospital Care Survey

National Hospital Care Survey (NHCS): An establishment survey that, when fully implemented, will provide nationally representative statistics, from claims and electronic health records (EHRs), on health and health care utilization from hospital inpatient stays and emergency department visits

NDI
National Death Index

National Death Index (NDI): A centralized database of U.S. death records gathered from states' vital statistics offices

Data Sources: Linked NHCS-NDI Data (cont.)



Patient level encounter data from hospitals

Mortality

Cause of death from National Death Index

Case Study I: Synthetic Data

Synthetic Data

- Explore the feasibility of developing a fully synthetic linked mortality file for public release
  - Model occurrence and date of death
  - Decision tree for cause of death

- Research question: how do the estimates from a fully synthetic linked mortality file compare to the original restricted use file?

Synthetic Data Methodology: Occurrence of death and date of death

- Occurrence of death and date of death modeled using Cox proportional hazards model:
  - Linear predictor of death risk
  - Baseline survival rates for each day

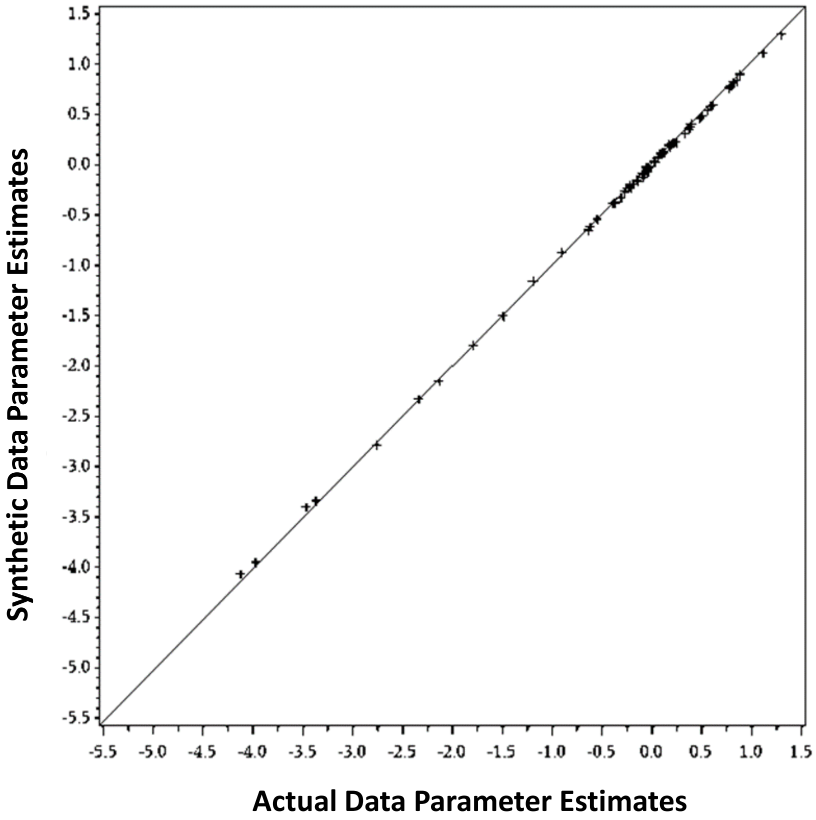
Synthetic Data:
Major and Contributing Causes of Death

- Modeled using classification tree (CART):
  - Contributing causes were modeled contingent on major causes
  - Predictors were similar to those used in survival model for date and occurrence of death

Results

- Comparative analysis shows very similar results for all-cause mortality when compared to the restricted-use data

- Some differences are noted in cause-specific estimates



Resnick, D., Cox, CS., Mirel, LB. (2021). Using synthetic data to replace linkage derived elements: a case study. Health Services and Outcomes Research Methodology. 1-18. 10.1007/s10742-021-00241-z.

Next Steps: Synthetic Data

- PCORTF FY 21 funding to create fully synthetic linked data files
  - Convene stakeholders meeting
  - Develop a methodology and implement
  - Create a validation method for researchers

Opportunities and Challenges

- Opportunities:
  - Increase data accessibility, which could expand data user community
  - Support evidence-based policymaking
  - Balance needs of researchers while protecting against disclosure risk
- Challenges:
  - User communication
  - Create statistically valid datasets with distributions similar to the true data
  - Create methods and processes for validation

Case Study II: Privacy Preserving Record Linkage

Privacy Preserving Record Linkage

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

- Privacy preserving record linkage or “PPRL” is a method that can be used to link de-identified data using hashing algorithms and tokens

- Research question: how do results from PPRL compare to results from the standard linkage algorithm?

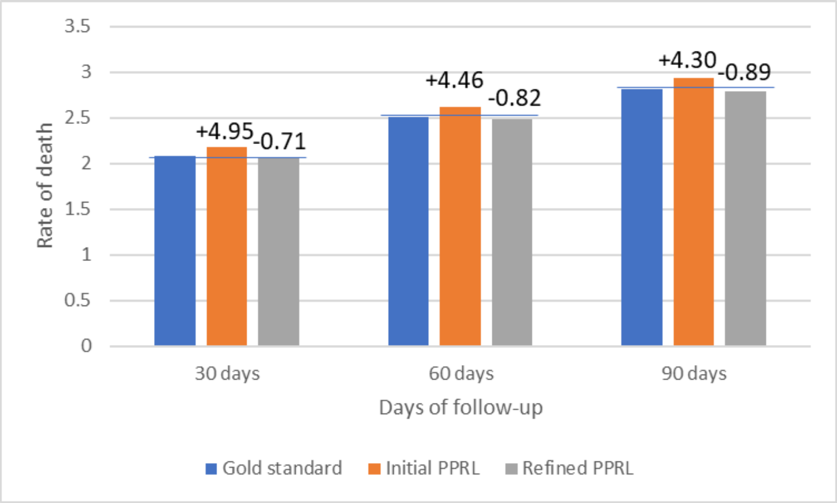
Privacy Preserving Record Linkage (cont.)

- NCHS case study: National Hospital Care Survey data linked to death certificate information from the National Death Index
  - Assess PPRL compared to “gold standard” (standard linkage algorithms)
    - Initial PPRL and refined PPRL
  - Determine the quality of PPRL linked data (sensitivity and specificity)
  - Compare estimates from two methods in secondary analysis

Results

- Depending on the selection of tokens from PPRL:
  - Sensitivity ranged from 98.7% to 97.8%
  - Specificity ranged from 99.7% to 99.9%
- 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



2016 NHCS Patients Linked to 2016-2017 NDI by Linkage Approach and Demographics

Characteristic	% Linked Gold Standard	% Linked Initial PPRL	% Linked Refined PPRL
Overall	5.1	5.4	5.0
Age			
<18	0.3	0.3	0.3
18-64	2.9	3.1	2.9
65+	20.0	20.8	19.9
Sex			
Male	5.9	6.2	5.8
Female	4.5	4.7	4.4

Note: Percent with non-missing SSN is highest in the 65+ group, which also happens to be the group that is most likely to link to the NDI.

Summary

- PPRL can be an effective record linkage technique that produces results similar to the standard linkage algorithm
  - High sensitivity and specificity estimates
  - Age- and sex-specific linkage rates show minimal differences for these subpopulations

Next Steps: PPRL

- Explore accuracy of PPRL when using sources with less complete PII
- Evaluate the impact of PPRL on health outcomes beyond mortality
- Assess the use of PPRL to expand NCHS data linkage activities

Opportunities and Challenges

- Opportunities:
  - Expand linkage beyond traditional federal data sources without sharing PII
  - Increase opportunities for NCHS, CDC and HHS to link data to address emerging public health threats

- Challenges:
  - Ensure PPRL methodology is safe and secure when sharing hashed tokens
  - Continue to evaluate and calibrate PPRL results for linkage accuracy

Discussion

- Synthetic data
  - Are there unique considerations with public use synthetic data products from NCHS?
    - User communication
    - Validation process
    - Privacy concerns
- PPRL
  - How do we extend this work to other sources (e.g., private sector, emerging public health data)?
    - Priority sources
    - Privacy concerns

Thank you!

National Center for Health Statistics
Data Linkage