Advancing the Data Modernization Initiative through Data Linkages

Lisa B. Mirel

NCHS Board of Scientific Counselors

May 26, 2022

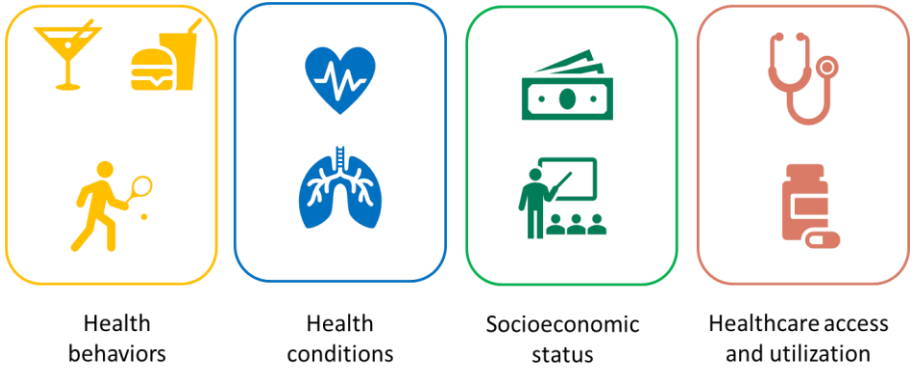
Overview

- NCHS Data Linkage Program
- Data Modernization Initiative
  - Linking to New Sources of Data
  - Evaluating Linkage Methodologies
  - Implementing Privacy Preserving Techniques
    - Privacy Preserving Record Linkage
    - Synthetic Data
- Summary
- Questions for the Board

NCHS Data Linkage Program: Sources

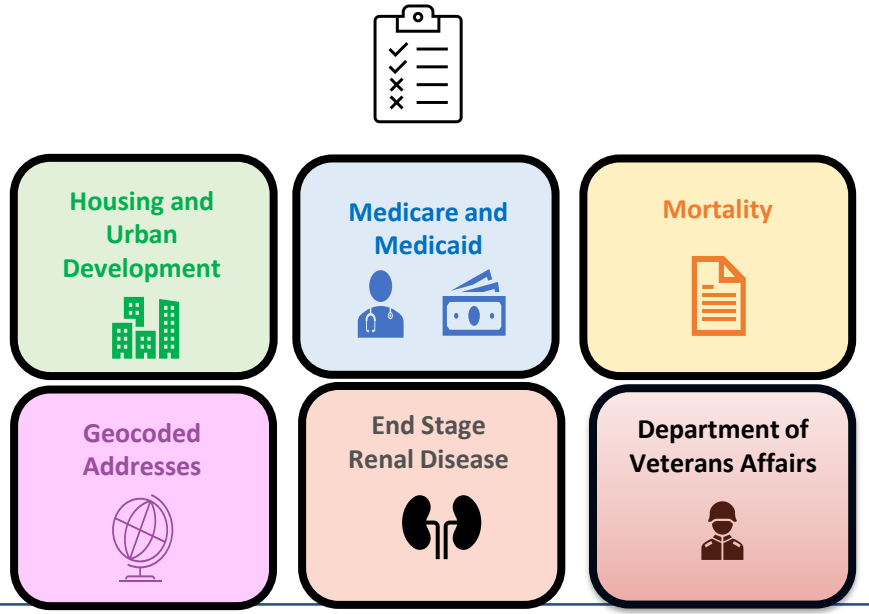
Survey Data

Sampling frame
Known inference



Administrative Data

Program participation/vital status
Not meant for research purposes



Linking NCHS Survey Data to Two New Sources of Data

Veterans Affairs (VA) data

Transformed Medicaid Statistical Information System (T-MSIS)

Both of these new data sources:

- Use enhanced linkage methodologies
- Will be available for researchers in summer 2022
- Can be used to answer key policy questions, such as:
  - ✓ VA: What are the health characteristics, outcomes, and health care utilization for Veterans within and outside the VA health system?
  - ✓ T-MSIS: How do health policy changes affect the health of Medicaid recipients?

T-MSIS PII data quality assessment will be included in linkage documentation.

Evaluating Linkage Methodologies

- NCHS Data Linkage Program developed state-of-the-art methodologies to link data (including machine learning)

- Comparison of NCHS enhanced algorithm to open-source linkage software

  - Exploring Match*Pro
  - Could be utilized enterprise wide

Initial Match*Pro Comparison to Standard Methods

- Random sample was compared
- Preliminary results are promising for total number of possible pairs
  - P(>0.95)→ concordance 87%
  - P(>0.85)→ concordance 89%

		Gold Standard		
Match*Pro P(>0.95)		Did not link	Linked	Total
	Did not link	n/a	988	988
	Linked	10	6,908	6,918
	Total	10	7,896	7,916

		Gold Standard		
Match*Pro P(>0.85)		Did not link	Linked	Total
	Did not link	n/a	792	792
	Linked	22	7,104	7,126
	Total	22	7,896	7,918

Next Steps for Comparison Analysis and Future Use

- Explore off diagonals
- Assess impact on secondary analysis
- Resolve computing capacity issues for larger datasets

		Gold Standard		
Match*Pro P(>0.85)		Did not link	Linked	Total
	Did not link	n/a	792	792
	Linked	22	7,104	7,126
	Total	22	7,896	7,918

Implementing Privacy Preserving Techniques: Evaluating Privacy Preserving Record Linkage (PPRL)

- PPRL could expand linkage opportunities to new sources of data
- Assessed how PPRL compared to standard linkage methodologies (using NHCS-NDI linkage for comparison)

✓ Precision ranged from 93.8% to 98.9%

✓ Recall ranged from 98.7% to 97.8%

A methodological assessment of privacy preserving record linkage using survey and administrative data

Cite

Article type: Research Article

Authors: Mirel, Lisa B.[a,*] | Resnick, Dean M.[b] | Aram, Jonathan[a] | Cox, Christine S.[c]

Affiliations: [a] Data Linkage Methodology and Analysis Branch, Division of Analysis and Epidemiology, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, USA | [b] Statistics and Methodology Department, NORC at the University of Chicago, Bethesda, MD, USA | [c] Health Care Programs Department, NORC at the University of Chicago, Bethesda, MD, USA

Correspondence: [*] Corresponding author: Lisa B. Mirel, Data Linkage Methodology and Analysis Branch, Division of Analysis and Epidemiology, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, USA. Tel.: +1 301 458 4087; E-mail: LMirel@cdc.gov.

Journal IAOS: DOI 10.3233/SJI-210891

Implementing Privacy Preserving Techniques: PPRL Next Steps

- Assess different PPRL tools, using NCHS linked data as gold standard
  - Supported with OS-PCORTF FY 22 funding
  - Open source and commercial products
  - Evaluating sub-populations and linkage quality
- Collaborate with ASPE to inform other HHS OPDIVs considering PPRL implementation going forward
- Participate in CDC's PPRL Community of Practice

Implementing Privacy Preserving Techniques: Synthetic Linked Data

Piloting innovative methods to create public-use linked files that increase data accessibility while maintaining analytic utility and protecting privacy

File Development

- Conducted expert interviews: 9 federal partners, 3 non-federal
- Focus on health equity and SDOH
- Selected variables, now testing
- Three proposed files
  - NHIS linked to HUD, CMS Medicare (two files: one for ages 18+, one for ages 65+)
  - NHCS linked to HUD and NDI (all ages)

File Dissemination

- Data will be available for researchers in summer 2023
- Creation of a validation service
  - Analyses will be run on synthetic data and then validated on the true linked data
  - Homomorphic encryption vs. manual validation
- Interactive data tables

Summary: DMI and Data Linkages

- Strengthening the exploration of new technologies for data linkage
- Creating opportunities to:
  - Expand linkages beyond traditional federal data sources
  - Create files that protect privacy and increase data accessibility
- Developing new resources to address emerging public health issues
- Supporting the objectives of the Foundations Evidence Building for Policymaking Act of 2018 and for CDC to respond to public health emergencies

Questions for the Board:

- What are some innovative strategies we should consider for outreach about the new linked files?
- Are there concerns with synthesizing variables that have been brought in at the zip code level (e.g., percent of zip code residents that are uninsured) for a person level file?
- Do you have suggestions on how to set up a sustainable validation service for the synthetic data?

Questions?

Thank you!