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

Data Linkage Overview

Lisa B. Mirel

April 26, 2023

CNSTAT: Approaches to Improve the Measurement of Law Enforcement Suicide

National Center for Science and Engineering Statistics

Social, Behavioral and Economic Sciences

National Science Foundation

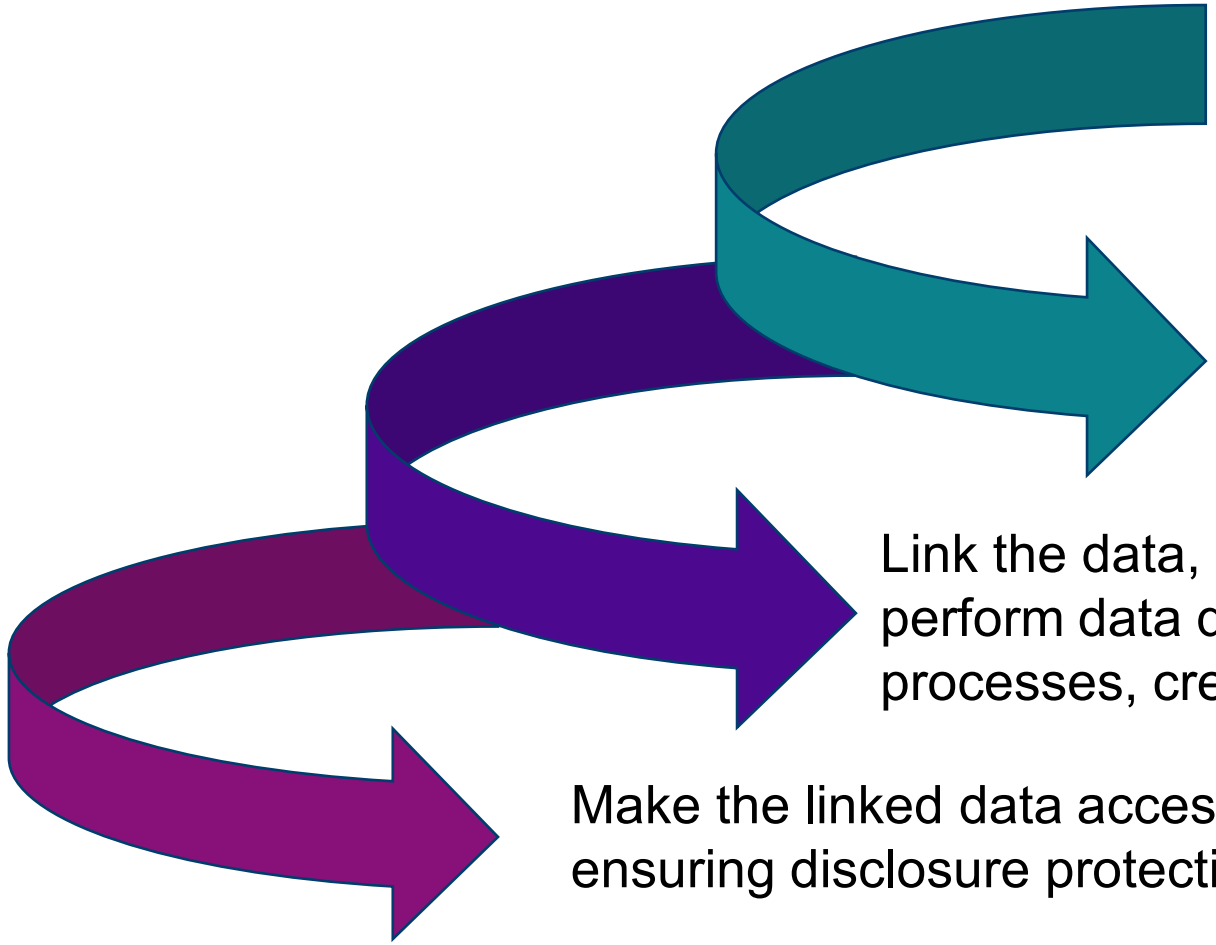
Data Linkage

- Linking data is a powerful and efficient mechanism for producing policy-relevant information
  - Brings together information to create a new, richer resource
  - Allows for the construction of longitudinal events with passive follow-up

Linking Survey and Administrative Data

- Survey data are collected from a targeted group to get information on factors like health status, well-being, access to benefits, etc.
- Administrative data are often collected for programmatic purposes
- Combining these data creates opportunities to answer key policy-relevant questions that would not be possible with each data source alone

Linkage Lifecycle



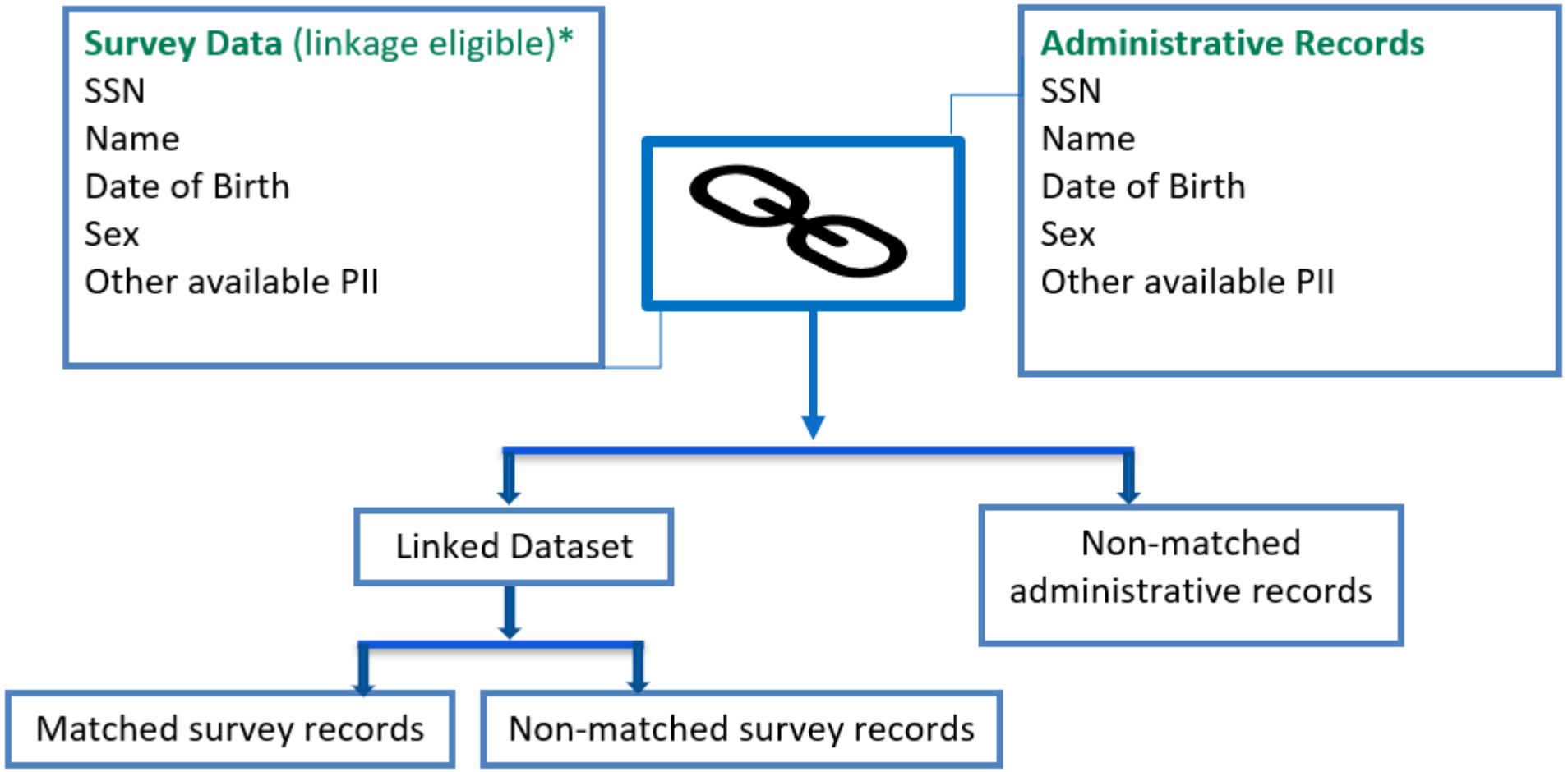
Determine feasibility of linking: coverage, data quality of linkage variables, questions to be answered

Determine data ownership and data sharing agreements, requirements and limitations on use

Link the data, manage data security, perform data quality checks, document processes, create curated linked files

Make the linked data accessible while ensuring disclosure protections

Example of a Linkage Process



*To be considered eligible for data linkage, linkage consent must be granted and participants must provide at least two of the following three identifiers: valid social security number (SSN), valid date of birth (month, day, and year) or valid name (first and last).

Example: Linkage Methodology

- Linkage occurs in two passes

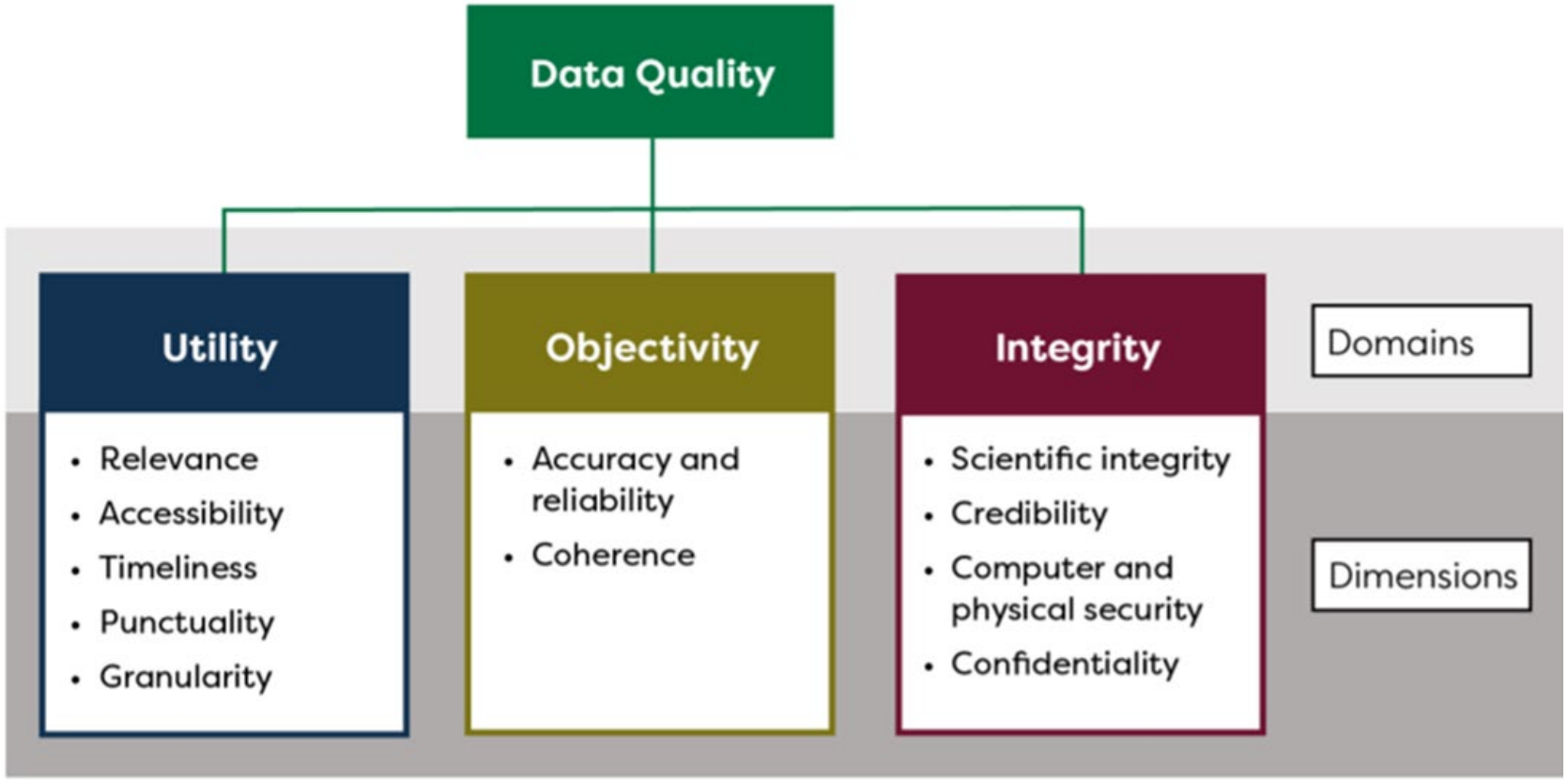
1. Deterministic match using Social Security Number (SSN)
   - Identifier fields such as name, state of residence, and date of birth are compared for validation
   - This dataset, based on the deterministic match, becomes the “truth deck” used later to estimate type I and type II errors

2. Probabilistic matching techniques used to identify likely pairs using other identifiers (not SSN)
   - Pair scores are calculated on the agreement status of the identifiers such as name, state of residence, and date of birth
   - SSN is not used to score pairs; instead, it is used to measure linkage accuracy (when available)

Factors to Consider

- Linkage eligibility (consent, sufficient personally identifiable information)
- Linkage error
- Analytic considerations
  - Data quality
  - Coverage
  - Data limitations and inference
  - Timeliness

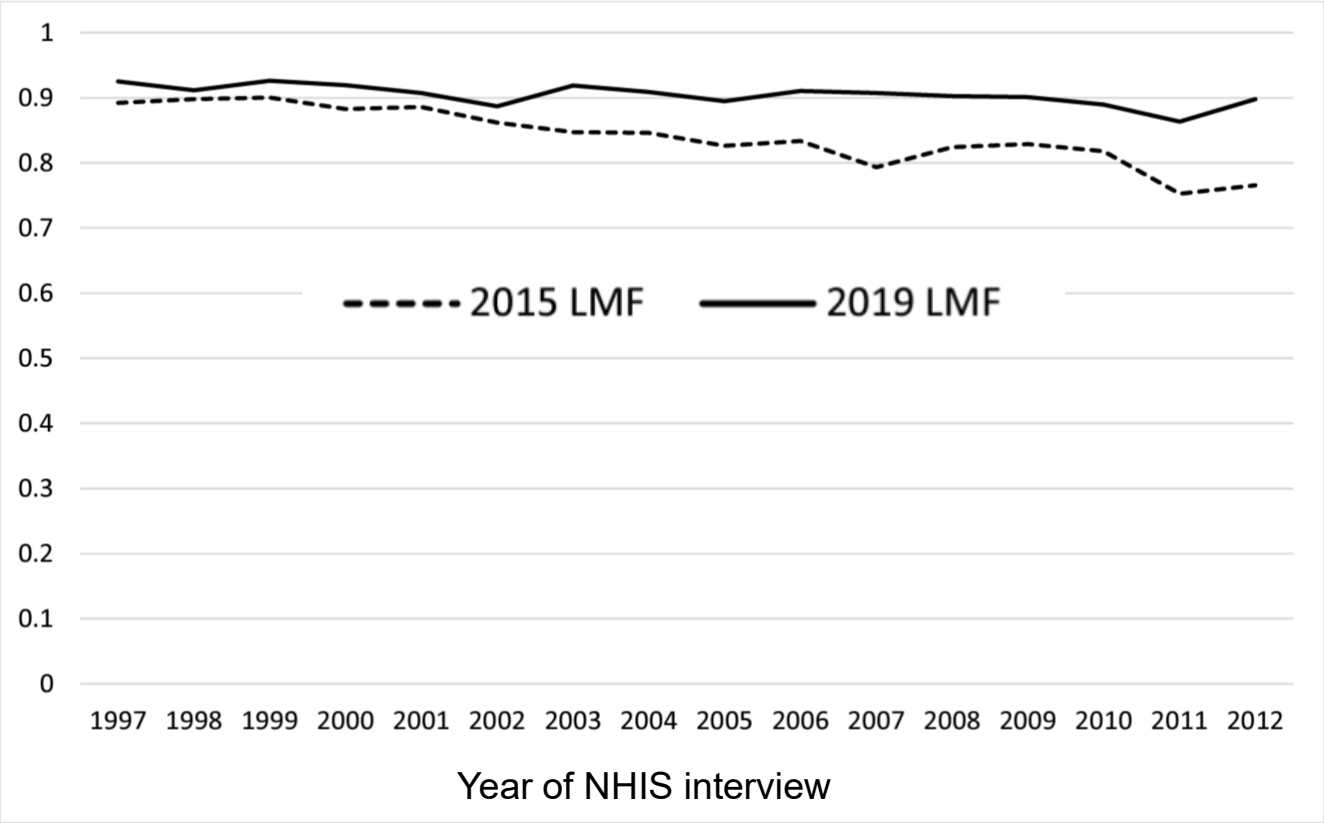
Transparency of Linked Data Quality is Essential for Proper Inference



Federal Committee on Statistical Methodology. 2020. A Framework for Data Quality. FCSM 20-04, September 2020.

Accuracy and Reliability: Example from the NCHS Linked Mortality Files (LMF)

Figure 1. Kappa statistics for concordance of mortality status with MEPS for the 2015 and 2019 LMFs, NHIS 1997-2012



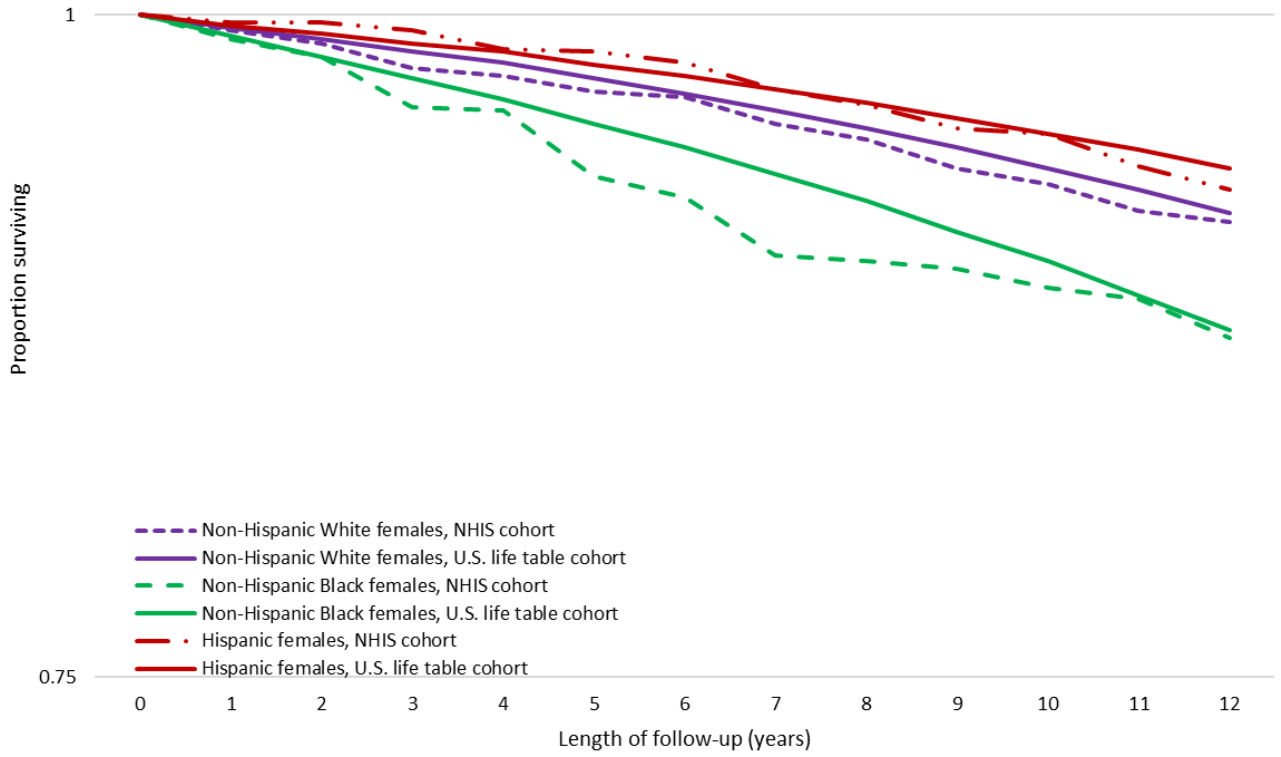
Compared concordance of external source of mortality status to linkage results, based on two different linkage methodologies

Concordance improved with the new linkage methodology (2019 LMF)

NCHS: National Center for Health Statistics; MEPS: Medical Expenditure Panel Survey; NHIS: National Health Interview Survey

Integrity/Coherence: Example from the NHIS LMFs

Figure 2. Survival curves for females, aged 50-59 years, by race/ethnicity and sex: 2006 NHIS LMF and U.S. life table cohorts



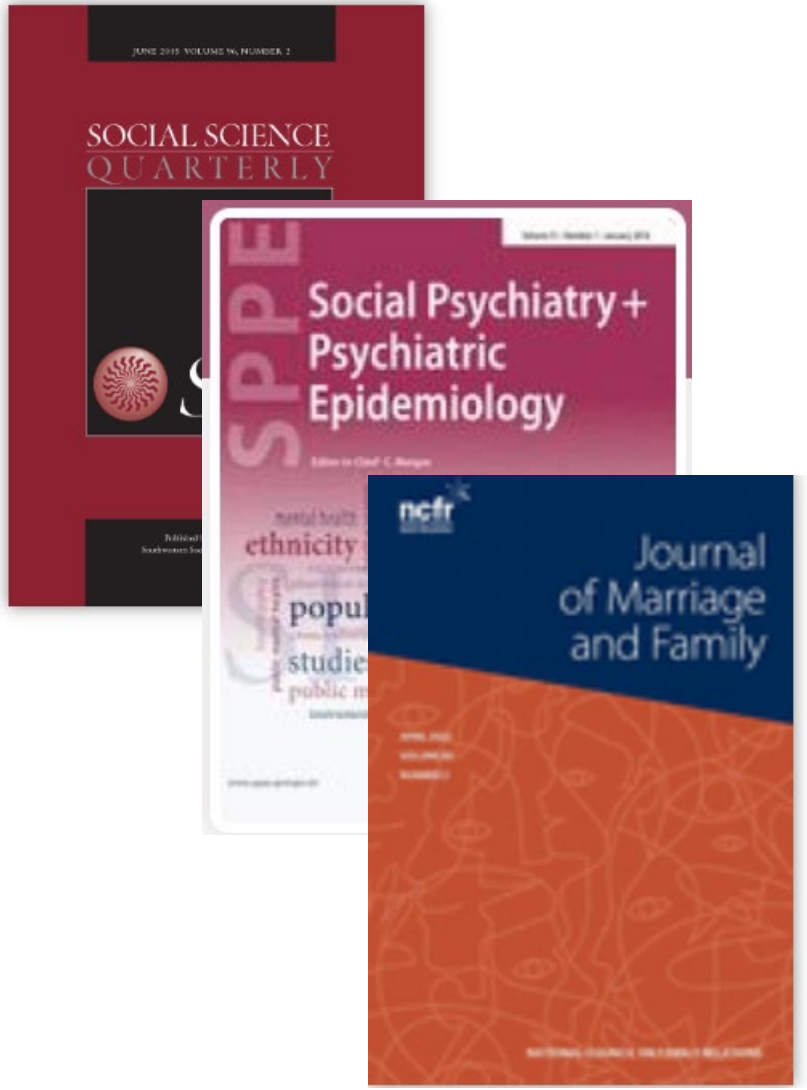
NHIS: National Health Interview Survey

Compared life expectancy models for national and linked data populations

Alignment of estimates support robust analyses using the linked data

Use of Linked Data to Examine Suicide Deaths

- Suicide in the City: Do Characteristics of Place Really Influence Risk?
- Adult Suicide Mortality in the United States: Marital Status, Family Size, Socioeconomic Status, and Differences by Sex
- Psychological Distress as a Risk Factor for All-Cause, Chronic Disease, and Suicide-Specific Mortality
- Family and Household Formations and Suicide in the United States



Data Linkage: Issues and Opportunities

Agreements and Data Sharing

Issues:

- Who owns the linked data?
- Where will the data reside?
- Where will the linkage occur?

Opportunities:

- Common data sharing model
- National Secure Data Service
- Federal Statistical Research Data Centers (FSRDCs)
- Agency-specific secure data access facility

Data Linkage: Issues and Opportunities

Agreements and Data Sharing

Issues:
- Who
- Wher
- Wher

Opportu
- Comi
- Natio
- Fede (RDC

Linkage Methods

Issue:
- Many methods require PII exchange

Opportunities:
- Assess Privacy Preserving Record Linkage (PPRL) tools that encrypt PII
- Validate PPRL tools against standard methodologies
- Utilize PPRL tools to expand data sources used in linkages

Data Linkage: Issues and Opportunities

Agreements and Data Sharing

Issues:

- Who
- Wher
- Wher

Opportu

- Comm
- Natio
- Fede (RDC

Linkage Methods

Issue:

- Many m

Opportuni

- Assess (PPRL)
- Validate method
- Utilize F in linkag

Quality of Linked Data

Issue:

- Need for methodologic standards and assessment tools for data quality

Opportunities:

- Incorporate deterministic and probabilistic methods to improve linkage quality
- Use external sources to assess data quality and benchmark
- Make use of federal efforts for data quality assessments and metadata resources

Data Linkage: Issues and Opportunities

Agreements and Data Sharing

Issues:
- Who
- Wher
- Wher

Opportu
- Comm
- Natio
- Fede (RDC

Linkage Methods

Issue:
- Many m

Opportuni
- Assess (PPRL)
- Validate method
- Utilize F in linkag

Quality of Linked Data

Issue:
- Need for assessm

Opportuniti
- Incorpora methods
- Use exte and benc
- Make use assessm

Data Accessibility

Issue:
- Linked data are primarily available through FSRDCs

Opportunities:
- Standard application process (SAP)
- Create more publicly available linked data based on synthetic data with verification/validation options
- Develop interactive dashboards for linked data systems that maintain privacy protections but expand access to potential new users

Successful Linkages Rely on Several Factors

- Support and adequate resources from both entities
- Consensus on data management responsibilities
- Agreement on secure access
- Commitment to high quality data standards
- Mutual understanding on why sources are being integrated
- Investigation of the strengths and limitations of the data and documentation of potential bias and error

Final Thoughts

- Continue to identify and integrate the data needed to answer key policy questions
- Utilize innovative technologies
- Explore alternative data sources for linkages

References

- Denney, J.T., et al., Suicide in the City: Do Characteristics of Place Really Influence Risk? Soc Sci Q, 2015. 96(2): p. 313329.
- Denney, J.T., et al., Adult Suicide Mortality in the United States: Marital Status, Family Size, Socioeconomic Status, and Differences by Sex. Soc Sci Q, 2009. 90(5): p. 1167.
- Hockey, M., et al., Psychological distress as a risk factor for all-cause, chronic disease- and suicide-specific mortality: a prospective analysis using data from the National Health Interview Survey. Soc Psychiatry and Psychiatr Epidemiol, 2021: p. 1-12.
- Denney, J.T., Family and Household Formations and Suicide in the United States. Journal of Marriage and Family, 2010. 72(1): p. 202-213.

Contact: Lbmirel@nsf.gov