Discussion
Data Driven Insights: The Utility and Policy Building of Integrated Data from Federal Statistical Agencies

Lisa B. Mirel

FCSM 2023

October 26, 2023: 10.30 AM

NATIONAL CENTER FOR SCIENCE AND ENGINEERING STATISTICS
NATIONAL SCIENCE FOUNDATION

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

Linking Data

- 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 data with passive follow-up

Factors to Consider in Data Linkage

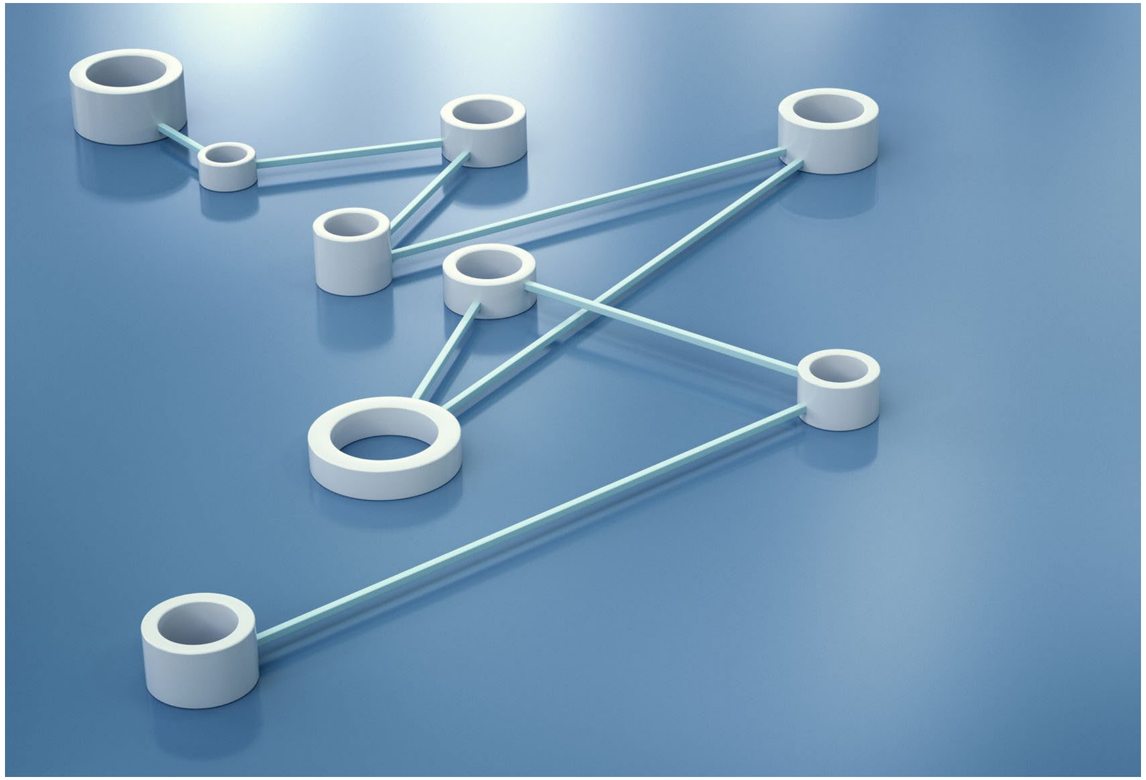
- Linkage eligibility (e.g., consent, sufficient personally identifiable information)
- Linkage error
- Analytic considerations
  - Data quality
  - Disclosure risks
  - Coverage
  - Data limitations and inference
  - Timeliness

Three Diverse Linkage Talks

- **National Center for Science and Engineering Statistics**
  - Linking the Policy and Utility of Linked Data in Science and Engineering (Finamore)
- **National Center for Health Statistics**
  - Does the Decade Matter? Examining the Impact of Using Geocodes from Different Decades in the Analysis of Merged Survey and Contextual Data (Parker)
  - Using Linked Data to Train and Validate Machine Learning Prediction Models (Davy)

Linking the Policy and Utility of Linked Data in Science and Engineering

- Recent NCSES data linkage efforts
  - Development of a policy-driven research and linkage framework
  - Insights on the way forward in this critical field of data integration
  - What lessons can be learned from others?

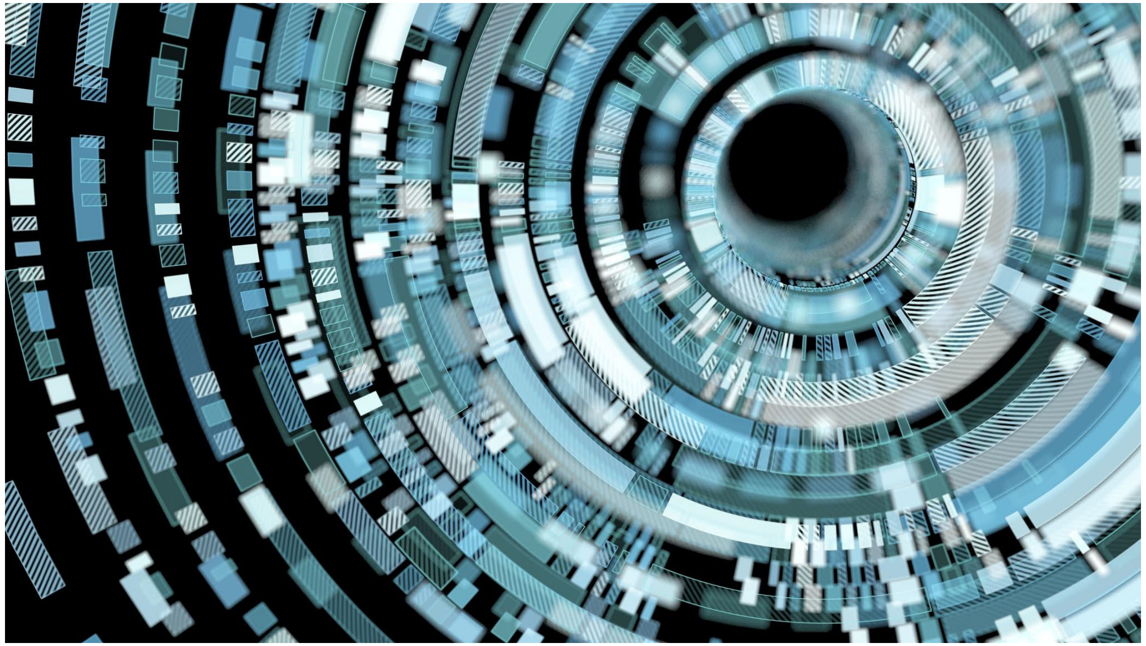


Does the Decade Matter?

- Geocoded survey data allow researchers to merge survey data with geographic information from other sources to add contextual information
- Geocodes change with each decennial Census
- Time period of the geocodes should be carefully considered because they may impact results due to real change over time and/or administrative changes

Using Linked Data to Train and Validate Machine Learning Prediction Models

- When data linkage is not possible, machine learning (ML) prediction models can be used to predict outcomes, such as morbidity and mortality
- ML models require high quality, accurate training data and a validation source
- ML models perform reasonably well for predicting all-cause mortality



Takeaways and Common Themes

- Establishing a robust linkage program requires many analytic and practical considerations
- By augmenting information, linked survey and administrative information facilitate richer analyses
- Quality assessments of linked data are important for inference and training datasets

Discussion Points

- What are key policies that need to be considered when establishing a robust linkage program?
- Are linkages, whether entity to entity or geocoded data, being considered to replace information that could be collected through survey questions?
- If linked data were to be used as a training data set for a ML model how often would they need to be updated?
    - What skills are needed to keep the training dataset up to date?

Discussion Points cont.

- Are considerations being given to utilize linkages to inform data collection (e.g., adaptive design)?
- What information is missing from the surveys discussed that would benefit from linking to another source to support evidence-based policymaking?

Final Thoughts

- Continue to identify and integrate the data needed to inform key policy questions
- Utilize innovative technologies and data sources
- Evaluate alternative data sources for linkages and uses of linked data

Lisa B. Mirel

Email address: lbmirel@nsf.gov

https://ncses.nsf.gov

@NCSESgov