National Center for Health Statistics
Data Linkage

The NCHS Data Linkage Program: Connecting Data Across Agencies

Lisa B. Mirel

Director, Data Linkage Program

Committee on National Statistics

September 27, 2021

National Center for Health Statistics (NCHS)

- Nation’s principal health statistics agency

- One of 13 federal statistical agencies

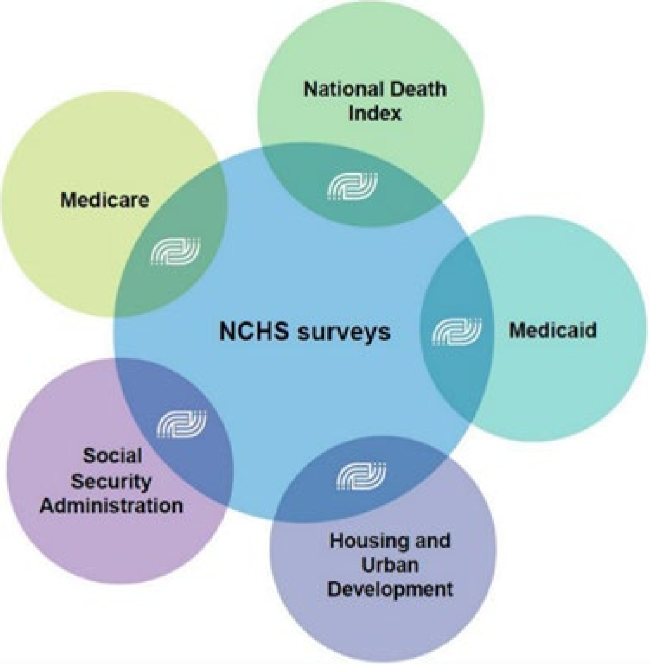
- **Mission:** To provide statistical information that will guide actions and policies to improve the health of the American people. As the Nation’s principal health statistics agency, NCHS leads the way with accurate, relevant, and timely data.

Policy Questions Require Evidence-Based Answers

- Many pressing policy questions require complex, detailed data
  - *Do Social Security Disability Insurance beneficiaries have access to care during the waiting period before Medicare entitlement?*
  - *Are there adverse health effects associated with the mandatory folic acid fortification policy for grain products?*
  - *How effective are health and housing policies in reducing lead exposure?*
  - *How likely are women and children who receive federal assisted housing to participate in Women, Infant, Children supplemental nutrition program?*

Data Linkage as a Solution

- Linking data is a powerful mechanism to provide policy relevant information in an efficient way
  - Health survey data are collected to monitor health status, health behaviors and health care access
  - Administrative data are collected for programmatic purposes
- Combining these types of existing data efficiently creates opportunities to answer key health and policy relevant questions



NCHS Surveys Included in Linkages

National Health Interview Survey (NHIS)

A nationally representative, cross-sectional household interview survey that serves as an important source of information on the health of the civilian noninstitutionalized population of the United States

National Health and Nutrition Examination Survey (NHANES)

A nationally representative, cross-sectional sample of the U.S. civilian noninstitutionalized population, which includes an interview in the household followed by an examination in a mobile examination center

National Health Care Surveys

A family of data collection efforts that gather information about providers of health care services and the patients they serve across the spectrum of health care settings from ambulatory and hospital care to long term care settings

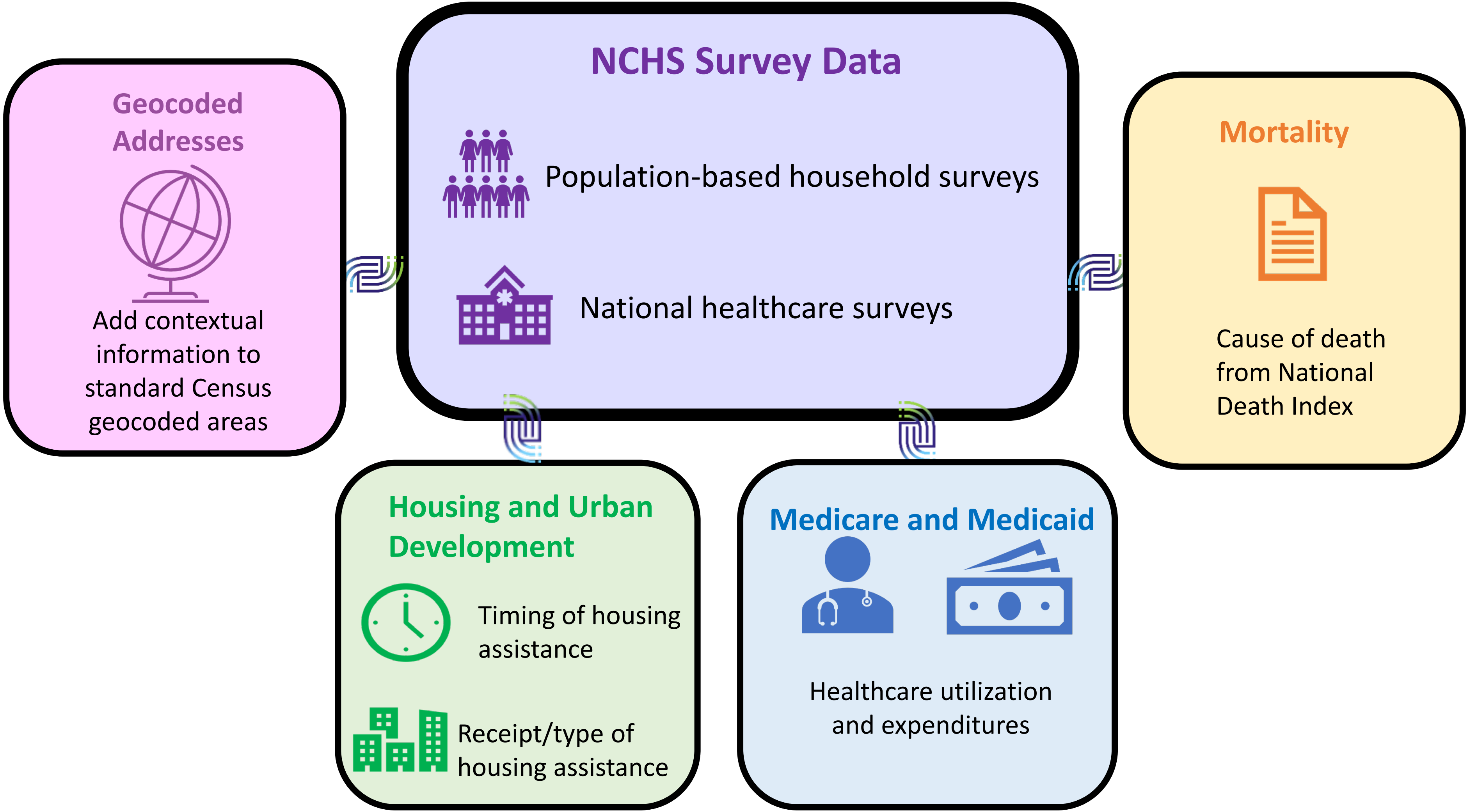
NCHS Surveys Collect Information on

Health behaviors

Health conditions

Socioeconomic status

Healthcare access and utilization



NCHS Survey Data

Population-based household surveys

National healthcare surveys

Geocoded Addresses

Add contextual information to standard Census geocoded areas

Mortality

Cause of death from National Death Index

Housing and Urban Development

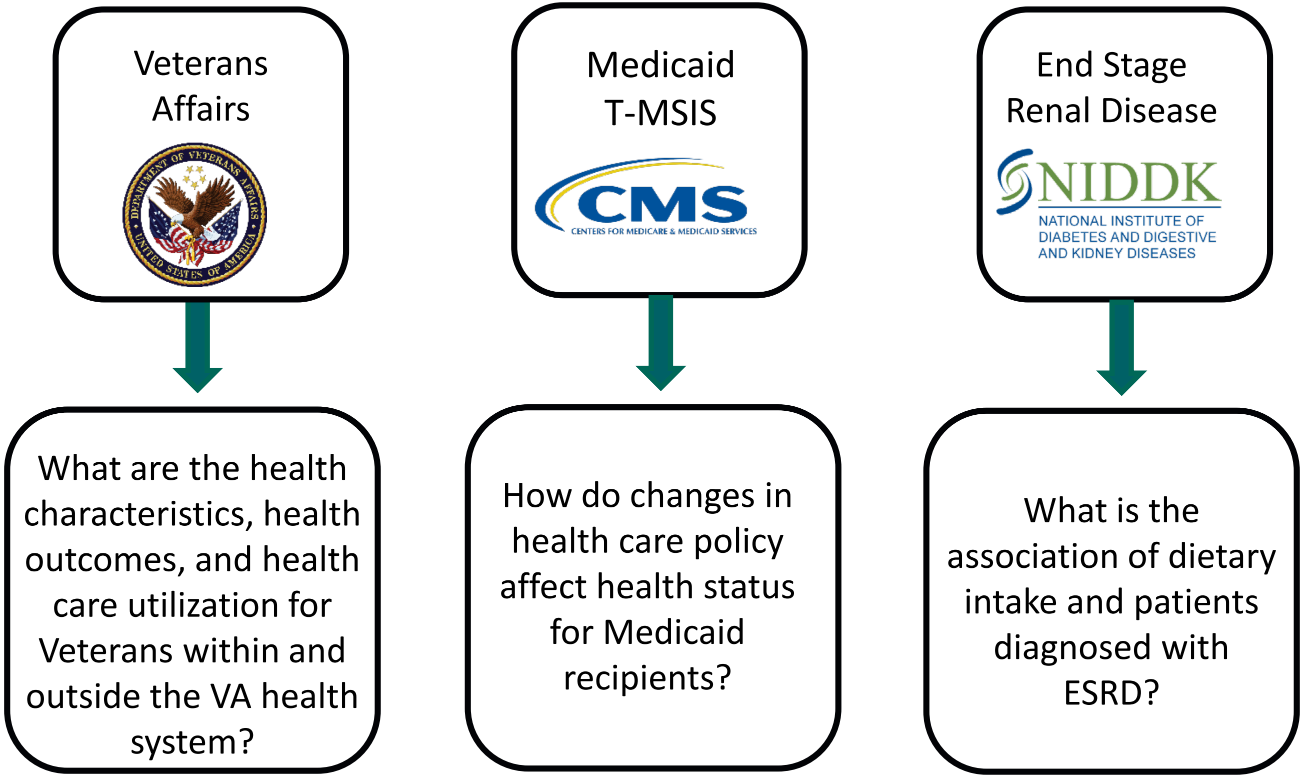
Timing of housing assistance

Receipt/type of housing assistance

Medicare and Medicaid

Healthcare utilization and expenditures

Upcoming NCHS Data Linkages



Veterans Affairs

What are the health characteristics, health outcomes, and health care utilization for Veterans within and outside the VA health system?

Anticipated release: early 2022

Medicaid T-MSIS

How do changes in health care policy affect health status for Medicaid recipients?

Anticipated release: early 2022

End Stage Renal Disease

What is the association of dietary intake and patients diagnosed with ESRD?

Anticipated release: Fall 2021

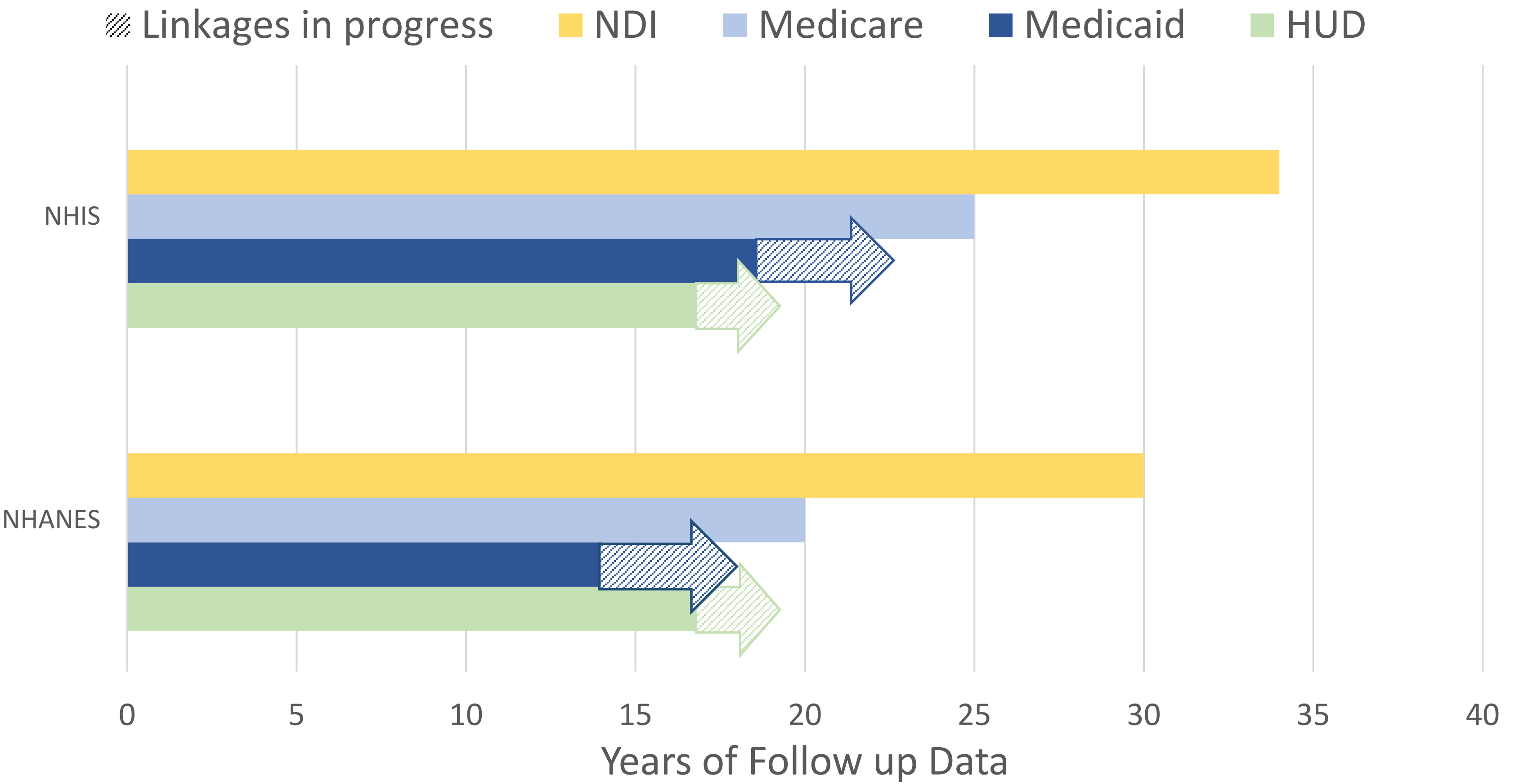
NCHS is Creating Resources that Support Evidence Building

Link NCHS survey data with the National Death Index and other health related administrative data

Provide documentation on linkage methodology, linkage quality and analytic guidelines

Release the curated data files for multiple research projects and to replicate the findings of other researchers

Availability of Linked NHIS and NHANES Data



Examples of Research using NCHS Linked Data

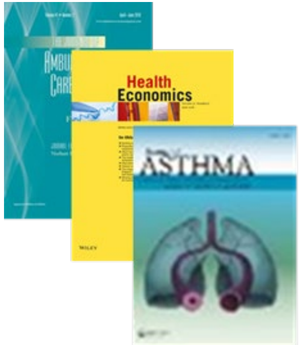
*Over **1,000** publications based on NCHS linked data*

Linked Mortality Data

- Deaths Associated with Underweight, Overweight, and Obesity
- Diet quality and all-cause mortality
- Educational Differentials in US Adult Mortality

Linked NCHS-CMS Data

- Characteristics of those who chose Medicare Advantage upon Medicare enrollment
- Health service use among the previously uninsured
- Concordance between survey reported childhood asthma and linked Medicaid



Linked NCHS-HUD Data

- Housing assistance and blood lead levels
- Cigarette smoking and adverse health outcomes among adults receiving federal housing assistance
- Housing assistance associated with insurance rates and unmet medical need

Challenges and Opportunities

Agreements and Data Sharing

Issue: Who owns the linked data? Where will the data reside? Where will the linkage occur?

Opportunity: Common data sharing model, HHS Data Council Subcommittee, HHS Protect

Challenges and Opportunities

Agreements and Data Sharing	Linkage Methods
Issue: Who owns the linked data? Where will the data reside? Where will the linkage occur? **Opportunity:** Common data sharing model, HHS Data Council Subcommittee, HHS Protect	**Issue**: Current methods require PII exchange **Opportunity:** Assess Privacy Preserving Record Linkage (PPRL) methods which mask PII, validating against standard methodologies, potential to expand data sources

Linkage Methods: PPRL

- NCHS Data Linkage Program has assessed PPRL software and algorithms to increase linkage activities across health care spectrum

- PPRL can be an effective record linkage technique that produces results similar to the standard linkage algorithm

  - Evaluations and careful implementation can reduce threats to scientific integrity, credibility, and accessibility
  - Potential to use of PPRL to expand NCHS data linkage activities

- Results of this work can assist researchers in evaluating the results of PPRL created data sources

Challenges and Opportunities

Agreements and Data Sharing

Issue: Who owns the linked data? Where will the data reside? Where will the linkage occur?
Opportunity: Common data sharing model, HHS Data Council Subcommittee, HHS Protect

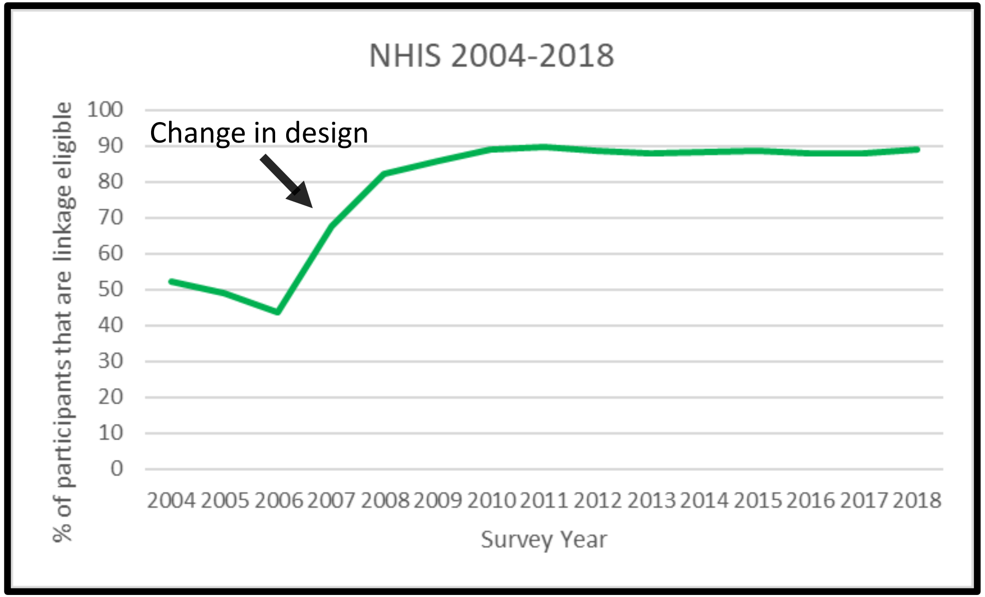
Linkage Methods

Issue: Current methods require PII exchange
Opportunity: Assess Privacy Preserving Record Linkage (PPRL) methods which mask PII, validating against standard methodologies, potential to expand data sources

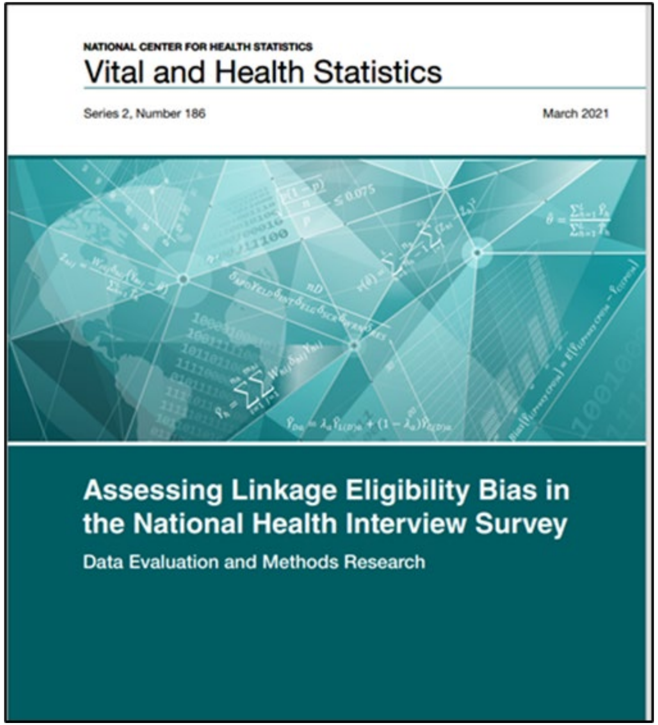
Quality of Linked Data

Issue: PII quality and completeness in surveys, need for methodologic standards and assessment tools.
Opportunity: Incorporate deterministic and probabilistic methods, utilize machine learning techniques to improve linkage efficiencies, use external sources to assess data quality, federal efforts to improve linked data metadata

Quality of Linked Data: Linkage Eligibility



Note: NHIS is the National Health Interview Survey; this graph only includes sample adults/ sample children



NATIONAL CENTER FOR HEALTH STATISTICS

Vital and Health Statistics

Series 2, Number 186

March 2021

Assessing Linkage Eligibility Bias in the National Health Interview Survey

Data Evaluation and Methods Research

Quality of Linked Data: Improve Accuracy

- Expanding use of machine learning techniques to improve linkage accuracy and efficiency
- Utilize deterministic and probabilistic methods
  - Pass 1: Deterministic match using survey-collected SSN (other identifiers used to validate). This data set becomes the "truth deck."
  - Pass 2: Probabilistic match using other identifiers. SSN is not used to create the probabilistic matches. Used to measure linkage accuracy.

$$A_i = Log_2\left(\frac{m_i}{u_i}\right)$$

$$D_i = Log_2\left(\frac{(1-m_i)}{(1-u_i)}\right)$$

- Estimate type I and type II errors

Challenges and Opportunities

Agreements and Data Sharing

Issue: Who owns the linked data? Where will the data reside? Where will the linkage occur?

Opportunity: Common data sharing model, HHS Data Council Subcommittee, HHS Protect

Linkage Methods

Issue: Current methods require PII exchange

Opportunity: Assess Privacy Preserving Record Linkage (PPRL) methods which mask PII, validating against standard methodologies, potential to expand data sources

Quality of Linked Data

Issue: PII quality and completeness in surveys, need for methodologic standards and assessment tools.

Opportunity: Incorporate deterministic and probabilistic methods, utilize machine learning techniques to improve linkage efficiencies, use external sources to assess data quality, federal efforts to improve linked data metadata

Data Accessibility

Issue: Linked data primarily are only available through RDCs

Opportunity: Create more publicly available linked data based on synthetic data and validation server, develop interactive dashboards for linked data systems that maintain privacy protections but expand access to potential new users

Data Accessibility: Linked Synthetic Data

- NCHS Data Linkage Program is piloting innovative methods to create public-use linked data files to improve accessibility and reduce disclosure risk
  - Conduct meeting(s) with researchers to identify key variables for synthetic datasets
  - Create publicly available linked synthetic datasets and establish a verification system
  - Develop an interactive data visualization tool to further increase accessibility and utility of evidence building linked data

Lessons Learned

- Successful data sharing relies on several factors

  - Support and adequate resources from both entities
  - Consensus on data management responsibilities
  - Agreement on secure access

Other Uses of Linked Data

- Linked data can be used to inform sampling, follow-up, and data quality
  - Augmenting data collection with administrative records (MCBS)
  - Informing the NHANES feasibility follow up study (change of address information and linked mortality files)
  - Assessing data quality (pregnancy check box on the death certificate)

More Information

NCHS Data Linkage Program:

datalinkage@cdc.gov

www.cdc.gov/nchs/data-linkage

Lisa Mirel: LMirel@cdc.gov

Subscribe to our ListServ (updates on program including release dates): Send an email message to **list@cdc.gov.** Leave the subject line blank. In the body of the message, type or paste: SUBSCRIBE NCHS-DATA-LINKAGE-PROGRAM lastname, firstname where 'lastname, firstname' is your last and first name.

National Center for Health Statistics
Data Linkage