NATIONAL CENTER FOR HEALTH STATISTICS

National Center for Health Statistics Data Linkage Program: Generating Synthetic Data to Support Tiered Access

Cordell Golden
Chief, Data Linkage Methodology and Analysis Branch

Privacy and Public Policy Conference
September 13, 2024

National Center for Health Statistics (NCHS)

- Principal health statistics agency in U.S.

- One of 13 principal federal statistical agencies

- **Mission:** To provide timely, relevant, and accurate health data and statistics that inform and guide programs and policies to improve our nation's health

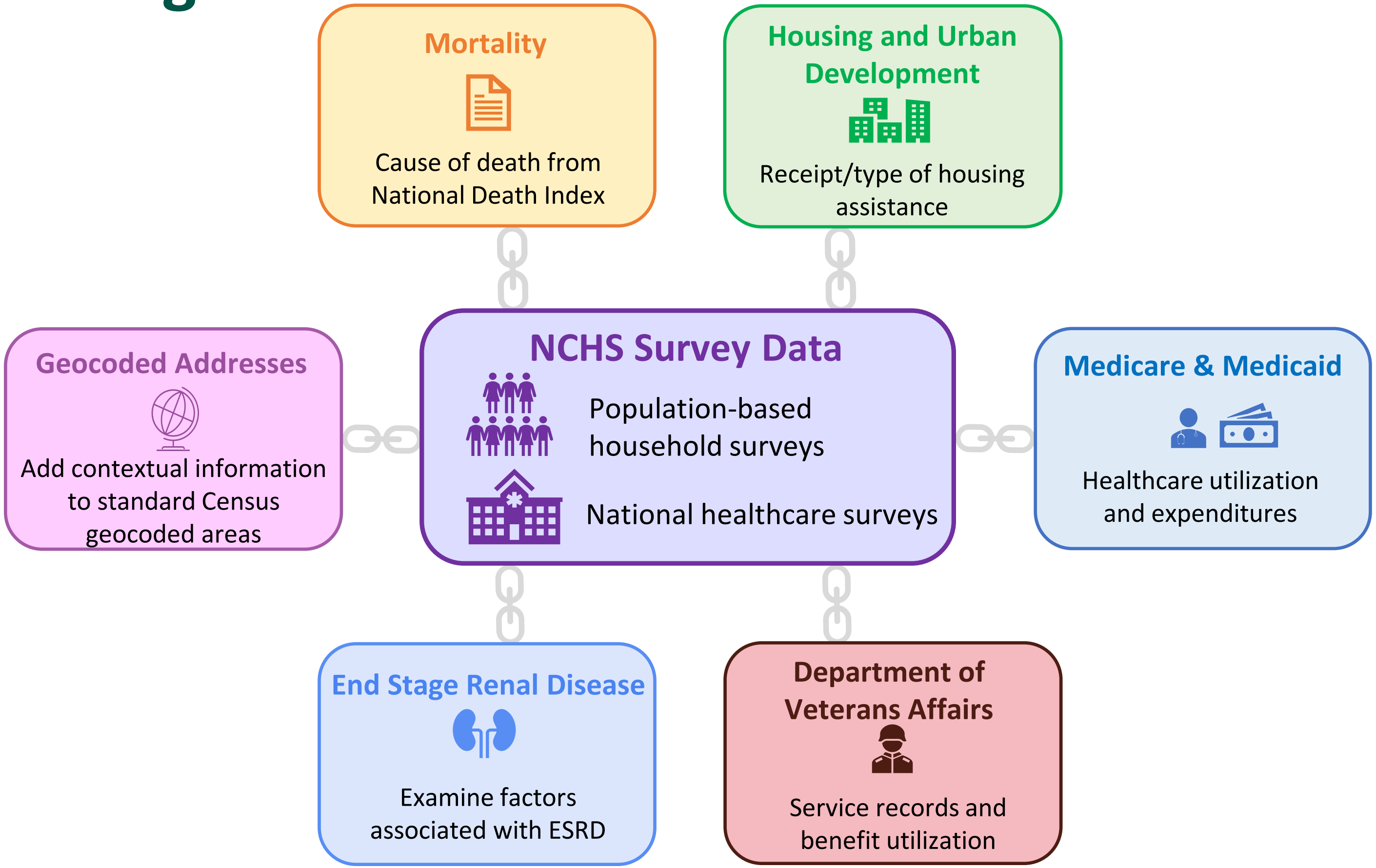
NCHS Data Linkage Program: Overview

- Create linked data files that support high quality research and program evaluation

- Utilize state of the art linkage methodologies and provide documentation and support for analyzing linked data files

- Explore innovative methods for maintaining researcher access to linked data

NCHS Linkages



Mortality

Cause of death from National Death Index

Housing and Urban Development

Receipt/type of housing assistance

Geocoded Addresses

Add contextual information to standard Census geocoded areas

NCHS Survey Data

Population-based household surveys

National healthcare surveys

Medicare & Medicaid

Healthcare utilization and expenditures

End Stage Renal Disease

Examine factors associated with ESRD

Department of Veterans Affairs

Service records and benefit utilization

Pilot Project: Synthetic Linked Data

Background

- NCHS links data from several NCHS population-based and health care provider surveys to health-related administrative data to create policy-relevant data resources in an efficient manner

- Privacy concerns impact linked data accessibility, and thus utilization
    - Nearly all NCHS linked data files are available only through the NCHS Research Data Center (RDC) Network

- To minimize this barrier to access, the NCHS Data Linkage Program is engaged in a pilot project to create public-use synthetic data files containing linked survey and administrative data

Pilot Project: Synthetic Linked Data Files

Project Objectives:

- Create downloadable public-use fully synthetic linked data files
    - Supports tiered access to federal data and Evidence Act requirements
    - Increases accessibility to NCHS linked data resources

- Create processes for users to verify synthetic results
    - Develop visualization tools that incorporate verification metrics to further increase accessibility and utility of linked data

- Project funded through ASPE’s Patient Centered Outcomes Research Trust Fund (PCORTF)

Pilot Project: Synthetic Linked Data Files

Accomplishments thus far:

1) Established collaborations with subject matter experts
   - Georgia Tech Research Institute for application of data synthesis techniques
   - Dr. Jerry Reiter (Duke University) for data synthesis and disclosure risk expertise
2) Conducted interviews with data users and subject matter experts to identify variables and populations of interest
3) Identified and applied appropriate synthetic data generation methodology
   - Incorporates sample design variables for National Health Interview Survey (NHIS)
4) Developed synthetic linked NHIS-HUD-CMS file
   - Multiple implicates for variance estimation
5) Conducted utility and disclosure assessments
6) Obtained approval from NCHS Disclosure Review Board
7) Developed test version of verification service (R-Shiny app) for selected regression models

Synthetic Linked 2018 NHIS-HUD-CMS file

Data Sources	Contextual Data	Survey Data (Demographic/SES)	Survey Data (Health-related)	Linked Data
Survey Data: 2018 NHIS **Administrative Data:** 2018 HUD 2018 Medicare **Contextual Data:** AHRQ SDOH Database	**AHRQ SDOH**(ages 18+): Percentage of households in Zip Code Tabulation Area (ZCTA) with: - Internet coverage - Medicaid (age < 64) - Income-to-poverty ratio < 1.0 - No health insurance	• Sex • Race/ethnicity • Age • Rental status • Education level • Employment status • Poverty status • Marital status	• Number of chronic conditions: - Diabetes - Obesity - Hypertension - Cancer - Asthma • Subjective health status • Flu vaccine • Smoking status • Serious psychological distress • Disability status • Type of health insurance • Usual place of care	**HUD** (ages 18+): • Receipt of housing assistance at time of interview • Receipt of housing assistance 2- and 5-years preceding interview and any time after interview **CMS Medicare**(ages 65+): • Months of FFS and MA enrollment • Medicare/Medicaid enrollment • Number of hospitalizations • Number of emergency visits • Total Medicare payments • Vital status

Synthetic Linked 2018 NHIS-HUD-CMS file

- Representative of 2018 NHIS participants who were eligible for linkage to HUD administrative data
- 25 synthetic data implicates – intended to be analyzed together not as individual datasets
- Each implicate contains 22,426 records
  - Records do not map to individual records in original data
- 51 variables
  - Original synthesized variables
  - Selected recodes of original variables to support logistic regression analyses
  - Universe/subpopulation variables to subset selected analytic cohorts (e.g., ages 65 and older, income to poverty ratio < 2.0)

Data Dissemination Plan

Downloadable public-use fully synthetic linked data files

- Publish analytic guidance for data users
    - Data development methodology
    - Codebooks
    - Sample code and guidance on multiple implicate analyses
- Provide verification process for data users to confirm accuracy of results
- Develop data visualizations that incorporate verification metrics

Proposed Verification Metrics

Decision Agreement	1. Are the signs of the coefficient estimates the same (+ / -)? 2. Is the p-value in the same direction, above or below, with reference to 0.05 for both coefficient estimates?
Estimate Agreement	3. Is the synthetic coefficient estimate contained within the confidence interval of the original coefficient estimate?
Percent Overlap for Confidence Intervals	4. What is the percentage overlap of the confidence intervals?

Proposed Verification Request Process

1. Verification request process for data users:
   - Select independent and dependent variable(s) for logistic regression model
   - Email request to Data Linkage mailbox

2. Return verification report to users

3. Assess user feedback to inform future efforts
   - Types of analyses requested
   - Synthetic data performance based on verification metrics
   - User requested enhancements

Next Steps

- Disseminate synthetic linked data files and finalize verification process
- Compile and evaluate user feedback
- Expand visualizations to support analysis using synthetic linked data files
- Repeat process for other data sources:
  - National Hospital Care Survey linked to National Death Index

Acknowledgments

GTRI

- Mark Bolding
- Richard Boyd
- Jordan Chandler
- Austin Himschoot
- James Jun

Duke University

- Jerome Reiter

NCHS

- Christine Cox
- Orlando Davy
- Roberto del Pozzo
- Kimberly Lochner
- Crescent Martin
- Frances McCarty
- Jessie Parker

Thank you!

Cordell Golden

CGolden@cdc.gov

Contact the Data Linkage Program: datalinkage@cdc.gov

Visit our website: www.cdc.gov/nchs/data-linkage

Subscribe to the NCHS Data Linkage Program LISTSERV to receive updates! Email a message to list@cdc.gov. Leave the subject line blank. In the body of the message, type:

- SUBSCRIBE NCHS-DATA-LINKAGE-PROGRAM last name, first name

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

The findings and conclusions in this presentation are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.