Privacy Preserving Techniques: Synthetic Linked Data Files

Cordell Golden, MPS

Chief, Data Linkage Methodology and Analysis Branch

NCHS Board of Scientific Counselors

September 14, 2023

Introduction

- Data linkage is a powerful mechanism for providing policy-relevant information in an efficient way
- NCHS links data from several NCHS population-based and health care provider surveys to health-related administrative data
- 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

Objectives

- Provide an overview of synthetic data file creation, dissemination, and verification plans

- Obtain feedback from the BSC with a particular focus on the planned verification process

NCHS Surveys Linked

National Health Interview Survey (NHIS)

A nationally representative, cross-sectional sample of the US civilian noninstitutionalized population, which includes a household interview that serves as an important source of information on the nation's health

National Health and Nutrition Examination Survey (NHANES)

A nationally representative, cross-sectional sample of the US civilian noninstitutionalized population, which includes a household interview followed by an examination in a mobile examination center (MEC) that serves as an important source of information on the health and nutritional status of adults and children

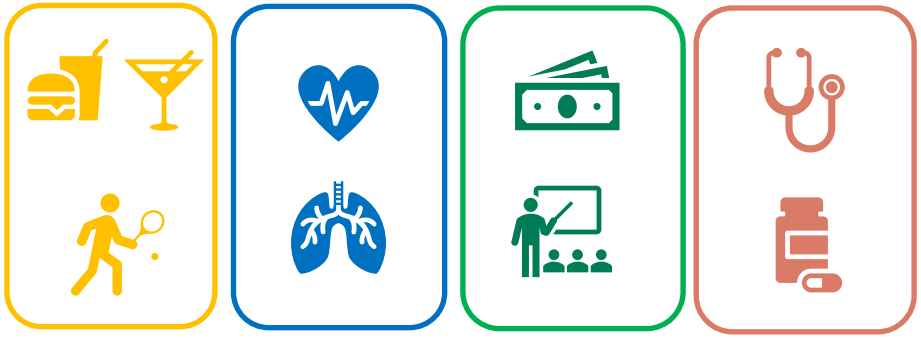
National Hospital Care Survey (NHCS)

NHCS collects data on patient care in hospital-based settings (inpatient, emergency, and outpatient departments) to describe patterns of health care delivery and utilization in the US

NCHS Linked Data Files

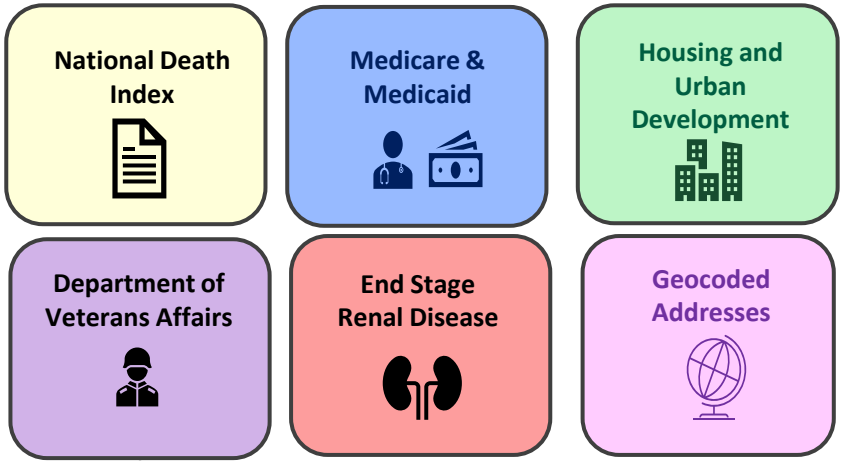
Survey Data

Sampling frame
Known inference



Administrative Data

Program participation/vital status
Not designed for research purposes



Pilot project: Synthetic Linked Data

Pilot Project: Synthetic Linked Data Files

GOALS:

- 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:

- 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
- Conducted interviews with data users and subject matter experts to identify variables and populations of interest
- Identified and applied appropriate synthetic data generation methodology
  - Incorporates sample design variables for NHIS
- Developed preliminary versions of two synthetic linked files:
  - Linked NHIS-HUD/CMS
  - Linked NHCS-NDI
- Conducted preliminary utility and disclosure assessments
  - Single file vs multiple implicates to maintain desired correlations

Proposed File #1:
2018 NHIS linked to HUD (ages 18+) and CMS (ages 65+)

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 - Arthritis • 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

Proposed File #2: 2016 NHCS linked to NDI (all ages)

Data Sources	Survey Data (Demographic/SES)	Survey Data (Health-related)	Linked Data
Survey Data: 2016 NHCS **Administrative Data:** 2016-2017 NDI	• Sex • Age • Imputed Race/Ethnicity (by geography and last name) • Urban/Rural status	• Presence of selected conditions: - Heart Disease - COPD - Mental Health - Septicemia/Sepsis - Diabetes • Month/year of most recent hospital encounter • Number of hospitalizations • Number of ED visits • ED admission within 7 days of hospital discharge • Total ICU days • Total Opioid-related encounters • Hospital discharge status	• Cause of death • Time to death: 30, 60, 90 days, 1 year • Opioid involved mortality • Drug overdose mortality

Current Data Dissemination Plan

Downloadable public-use fully synthetic data files

- Obtain NCHS Disclosure Review Board approval to release
- Publish analytic guidance for data users
  - Data development methodology
  - Codebooks
  - Sample code and guidance on multiple implicate analyses
  - Potential NCHS demonstration publication
- Provide verification process for data users to confirm accuracy of results
- Develop data visualizations that incorporate verification metrics

Verification Metrics being considered for Regression Models

Decision Agreement	• Are the signs of the coefficients the same (+ or -) for the true and synthetic data? • Do both p-values indicate statistical significance ($p \leq 0.05$ or $p > 0.05$)?
Estimate Agreement	Does the synthetic data estimate fall within the confidence interval of the true data?
Percent Overlap for Confidence Intervals	What is the percent overlap of the confidence intervals?

Proposed Verification Request Process

1. Verification request process for data users:
   - Specify type of regression model
   - Select independent and dependent variable(s)
   - 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 (early 2024)
- Collect meaningful user feedback (ongoing)
- Launch data visualization effort (estimated release in late 2024)

Discussion

What are approaches to most effectively communicate information about the synthetic data, including limitations of the synthetic data and verification process?

- *For example, we are planning a NCHS report that will be published on the NCHS website*

Who should we be sure to engage, with respect to optimizing disclosure risk and data release?

- *Should we wait to release the synthetic files until verification processes are ready?*
- *Should we consider limiting the number of user requests for verification as a way to minimize potential disclosure risks?*

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