Leveraging Non-Probability Data at the National Center for Health Statistics

Katherine Irimata, PhD

Paul Scanlon, PhD

Lauren Rossen, PhD

Guangyu Zhang, PhD

National Center for Health Statistics, Centers for Disease Control and Prevention

2024 Joint Statistical Meetings

National Center for Health Statistics

Overview

- The National Center for Health Statistics (NCHS) is the nation's principal health statistics agency, providing data and disseminating statistical information to identify and address health issues
- NCHS accomplishes this through a variety of data sources and methods:

Vital Records

NVSS

National Death Index (NDI)

Population Surveys

National Health Interview Survey

nhanes

NSFG

Provider Surveys

NAMCS National Ambulatory Medical Care Survey

NHAMCS National Hospital Ambulatory Medical Care Survey

NHCS National Hospital Care Survey

NPALS National Post-acute and Long-term Care Study

Panel Surveys

Research and Development Survey (RANDS)

NCHS Rapid Surveys System

Overview

- Non-probability data have been considered in certain applications to expand research, data collection, and reporting of official health statistics
- Recent applications include:

Vital Records

National Death Index (NDI)

Population Surveys

National Health Interview Survey

nhanes

NSFG

Provider Surveys

NAMCS National Ambulatory Medical Care Survey

NHAMCS National Hospital Ambulatory Medical Care Survey

NHCS National Hospital Care Survey

NPALS National Post-acute and Long-term Care Study

Panel Surveys

Research and Development Survey (RANDS)

NCHS Rapid Surveys System

Motivation

- Earlier presentations in this session introduced cutting edge research on combining probability and non-probability data
- NCHS has utilized non-probability data to improve the study of public health including producing more granular estimates and oversampling hard-to-reach populations
- Two applications will be presented:
  - National Hospital Care Survey (NHCS)
  - Research and Development Survey (RANDS)
- This presentation discusses the incorporation of non-probability data in federal statistical agencies and future applications and opportunities

National Hospital Care Survey

Background

- The National Health Care Surveys are a collection of surveys designed to study health care topics across a broad spectrum of health care settings
- One of these surveys, the National Hospital Care Survey (NHCS) collects data on patient care in hospital-based settings to describe patterns of health care delivery and utilization in the US
- Settings include inpatient and emergency departments
- Data is collected from a nationally representative sample of US hospitals sourced from administrative and electronic health records
- https://www.cdc.gov/nchs/nhcs/index.htm

Application

- Prior to 2020, low response rates have resulted in unreliable national estimates from the NHCS
- To improve reliability of the estimates, NCHS worked with NORC to implement a modeling-based weighting methodology that utilized third-party data sources to produce reliable national estimates
- The Premier Healthcare Database, a commercially available non-probability hospital-based database, was used in production of weights
- The Healthcare Cost and Utilization Project nationwide samples (Nationwide Inpatient Sample and Nationwide Emergency Department Sample) were used as benchmarks to calibrate the encounter-level weights for each setting

Methods

- NCHS and NORC developed a statistical modeling methodology for weighting the 2020 NHCS
- Sampling weights for responding NHCS hospitals and hospitals from the Premier database were adjusted based on an estimated propensity of response
- A combined sample was created from the NHCS hospitals and Premier hospitals using a weighting factor of 0.5
- Variance estimation for the combined sample was conducted using stratified jackknife
- A weighted version of NHCS-only data was created by calibrating to key national estimates

Data Access and Documentation

- A public use file of NHCS data (subsample from the restricted-use NHCS-only data) is available: https://www.cdc.gov/nchs/nhcs/2020nhcs.htm
- Technical documentation: https://www.cdc.gov/nchs/data/nhcs/2020-NHCS-PUF-Tech-Doc-508.pdf
- Learn more in the JSM session:

Wednesday, August 7, 2-3:50 PM

Session: Bolstering Health Survey Data with Health-Related Administrative Data: Organizational Infrastructure and Analytic Examples

Presentation Title: Utilizing Additional Data Sources to Improve National Representative Estimates on Patient Care in Hospitals

Speaker: Geoffrey Jackson

Research and Development Survey

Background

- Research and Development Survey (RANDS) is an ongoing series of surveys that are primarily sampled from web-based, recruited, commercial survey panels
- Designed to expand NCHS' methodological research:
  - To supplement NCHS' survey and questionnaire evaluation efforts, including the detection of measurement error
  - To explore ways to integrate data from commercial survey panels with high-quality data collections to produce reliable national estimates
- Estimates from a special series, RANDS during COVID-19, were published to report on the health impacts of the COVID-19 pandemic
- Eight rounds of RANDS and three rounds of RANDS during COVID-19 are publicly available
- https://www.cdc.gov/nchs/rands/index.htm

Background

- While RANDS has primarily included probability sampled surveys, nonprobability surveys were collected for RANDS during COVID-19 Rounds 1 and 2, RANDS 8-10 (RANDS 9 and 10 not yet publicly available)
- RANDS during COVID-19 nonprobability samples were collected by Dynata and used to supplement the sample sizes and used for methodological studies comparing probability and nonprobability samples
- RANDS 8 featured an oversample of gender minorities
- RANDS 9 and 10 featured oversamples of Middle Eastern, North African, and Afro-Caribbean adults

Application

- RANDS 8 included two data collection approaches on gender identity to conduct a question design experiment for non-binary gender measures
- Due to the small size of the gender minority population and the split sample design, the questionnaire was fielded to a probability sample (NORC's AmeriSpeak Panel) and a nonprobability sample with an oversample of gender minority respondents (CINT's Lucid Panel and Community Marketing and Insights' Panel)

	Probability Sample	Nonprobability Sample
Total Respondents	6,857	9,791
Gender Minority Respondents*	109 (1.6%)	685 (7.0%)

*Respondents were identified as gender minorities based on their questionnaire responses.

Methods

- Balancing weight was created to avoid confounding factors that could exist due to different distributions of demographic variables between the two samples
- Age, race/Hispanic ethnicity, education, marital status and metropolitan status were applied to balance the two samples using inverse propensity scores
- Probability sample was treated as a benchmark for the non-probability sample
- Creates a pseudo-sample consisting of two strata: one stratum consisting of AmeriSpeak panelists with unit weights (weight of 1 for each respondent), and a non-probability stratum consisting of panelists from the Lucid and CMI Panels with balancing weights to approximately match the sample proportions of AmeriSpeak panelists on the selected variables

Data Access and Documentation

- Approach will be applied to RANDS 9 and 10 to study differences in responses to questions about race and ethnicity and health outcomes among Middle Eastern, North African, and Afro-Caribbean adults
- RANDS public use files: https://www.cdc.gov/nchs/rands/
- RANDS 8 technical documentation: https://www.cdc.gov/nchs/rands/files/RANDS8-np-technical-documentation.pdf

Other Non-Probability Research at NCHS

Overview

- Research and investigation of the uses of non-probability data is ongoing based on limitations and needs of NCHS data systems and programs
- For example:
  - Improving reliability of estimates (NHCS)
  - Oversampling hard to reach populations (RANDS)
  - Small area estimation

Background

- Currently have a contract to investigate how nonprobability surveys can be used to project estimates in other sources for the purpose of small area/domain estimation
- Research focus has included:
  - Assessing various methods and models for generating estimates of self-rated health, everyday discrimination, and loneliness for gender minority population using data from RANDS 8 and the National Health Interview Survey (NHIS)
  - Evaluation of various statistical learning approaches for prediction
  - Identification of limitations of various approaches and models (e.g., when the selection probabilities for a given non-probability sample are correlated with the outcome variable of interest; when there are limited covariates shared across data sources)

Discussion

Discussion

- Non-probability data should be used with caution, particularly the use and dissemination from a federal statistical agency
  - Non-probability data can introduce bias
  - Removing and/or quantifying bias can be challenging
- However, non-probability data can also improve the availability and usability of important data for decision making and policy
- Purpose and use have been important factors for determining where to consider incorporating non-probability data
- The presented examples demonstrate important applications in which data would not otherwise be available and ongoing research is being used to guide future uses of non-probability data

Contact Information:
Katherine Irimata
kirimata@cdc.gov

Acknowledgements:
Morgan Earp
Yulei He
Geoffrey Jackson
Jennifer Parker
Van Parsons
Priyam Patel
Valerie Ryan
Makram Talih

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