Using Administrative Data to Mitigate Missing Data in Surveys: Design through Survey Data File Preparation

Examples from the National Survey of Early Care and Education (NSECE)

09.05.23

A Rupa Datta

NORC at the University of Chicago

Agenda

01 Mitigation strategies when conducting surveys

02 Applying administrative data sources

03 Introduction to the NSECE

04 Example mitigations in the NSECE

05 Concluding comments

06 Mitigation strategies during analysis

Mitigation strategies when conducting surveys

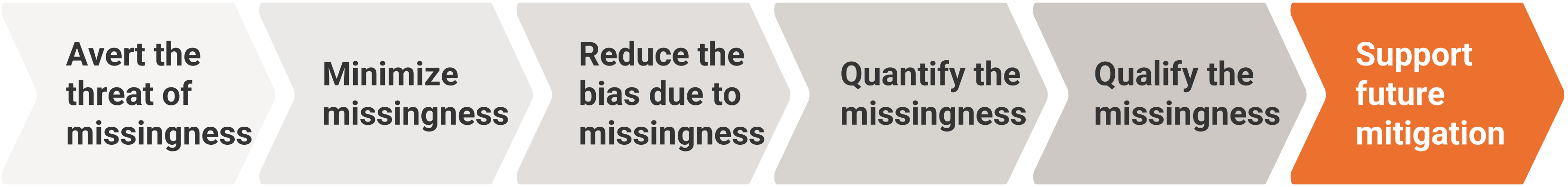
NORC at the University of Chicago

Missing data can impede our survey objectives

- Random missingness may reduce sample sizes and increase standard errors
  - Inaccurate inferences and estimates should not arise
- Systematic missingness can result in biased data and findings
  - Inaccurate or misleading estimates and inferences could arise. 😠

- The more we can know about the missingness, the more we can do to ensure that estimates and inferences from the data are accurate despite the missingness.

Mitigation strategy: How can we feasibly minimize the harm that non-response will do to inferences and estimates?



Avert the threat of missingness

Minimize missingness

Reduce the bias due to missingness

Quantify the missingness

Qualify the missingness

Support future mitigation

Lack of relevant information is often the greatest obstacle to protecting our data from missingness

- **Do we have missingness?**
- **Is it systematic or random?**
- **How might the missingness affect estimates or inferences?**
- **How big is the non-response?**
- **Who are the non-respondents?**
- **Who do we think is at risk of non-response?**

Applying administrative data

NORC at the University of Chicago

Administrative data were generated as a by-product of a non-research activity

- Sources can include
  - Administration of a public program, such as child care subsidies
  - Operations of a commercial activity, such as credit card transactions
  - Other human activity, such as internet searches, social media posts, mapping/route navigation, weather tracking, etc
- Administrative data can vary widely in
  - How well documented and curated the data are
  - The validity, reliability and consistency of the data
  - Their accessibility (legally, financially, logistically)

When and how to use administrative data depend on the characteristics of the data and the intended use

Disadvantages

- May not correspond to any populations of inferential interest
- Data may be poorly documented
- Quality can be variable
- Linkage to research data can introduce new biases and errors
- Access can be extremely challenging
  - Long waits
  - Costly fees or legal requirements
  - Elaborate technology needs

Advantages

- Can be accurate with respect to the data generation process (e.g., WIC participation, Google searches conducted, etc.)
- Often very large samples
- Missingness and bias in administrative data can be quite different (less correlated) from survey data
- Access can be quick and inexpensive

A few examples of administrative data can illustrate their variability

- Unemployment insurance data on workers’ earnings
  - State-by-state rules on data access, extensive legal negotiations
  - Highly accurate, fairly extensive coverage of labor force
- Commercial lists of schools or child care providers
  - Inexpensive and accessible
  - Coverage rarely matches populations of research interest
  - Data quality and documentation are variable
- Census data (almost not administrative data)
  - Well curated and documented
  - Extremely high quality
  - Many public use resources available for easy download, linkable/identifiable data can be accessed at very significant cost and effort if appropriate

Introduction to the NSECE

NORC at the University of Chicago

The NSECE is made up of four coordinated nationally-representative surveys that provide information on:

- early care and education (ECE) services available to families in settings that are
  - **home-based** (for children under age 13) and
  - **center-based** (for children ages birth through age 5, not yet in kindergarten);
- characteristics of the **workforce** providing these early care and education services; and
- **households** with children under age 13.

More about the NSECE

- NSECE data have been collected in 2012 and 2019 and are in preparation for 2024.
- All four surveys that are part of the NSECE use a variety of modes for outreach to respondents and for interview completion, always including mail, web, phone and in-person.
- OPRE funds the NSECE, with co-Federal Project Officers Ivelisse Martinez-Beck and Ann Rivera.
- A primary focus of the NSECE is to measure how the availability of care interacts with how families use that care, and to make these measurements at local and national levels

Selected mitigations in the NSECE: Uses of administrative data

NORC at the University of Chicago

Selected examples of how administrative data are used to mitigate missingness in the NSECE

- Identifying and correcting for undercoverage in the sample frame
- Implementing data collection methodologies to avert or minimize potential non-response
- Monitoring data collection for non-response
- Developing data collection interventions in response to emerging non-response
- Partially correcting for final survey non-response

Using administrative data: Identifying and correcting for undercoverage in the sample frame

- **The household survey uses a sampling frame of housing units in the U.S. based on information from the U.S. Postal Service**
  - For each sampled census tract, we compare counts of housing units in our frame with counts of housing units in the Decennial Census; where the frame count is low relative to census data, we supplement the frame
- **The center-based provider survey uses a frame we build from state and national lists of providers**
  - We supplement these state and national lists of early care and education (ECE) providers with a commercially-available list of all schools providing any grades kindergarten through fourth, which may also provide ECE but not appear on ECE provider lists

Using administrative data: Implementing data collection methodologies to avert or minimize potential non-response

- Household advance letters are either English-dominant with alternative Spanish text, or fully bilingual with equal English and Spanish content.
  - Use census data to identify communities with high rates of Spanish spoken at home to receive fully bilingual materials.
- Special permissions are sometimes required to interview within protected areas, such as tribal lands or public school districts.
  - Use boundaries of tribal areas and public school district records to identify sampled units that may require special permissions and seek those permissions in advance of data collection.

Using administrative data: Monitoring data collection for non-response

- **We sample households based only on an address, with no other information.**
  - Monitor screening and interview completion rates by commercially-available flags for possible presence of children, low-income household, and other characteristics. Investigate unequal rates.
- **Some states' lists of home-based providers include limited address information.**
  - We compared response rates of providers that had complete vs incomplete address information in state lists to ensure that we could equalize response rates as much as possible. In addition, we used manual review and data science techniques to correct for duplication in these state lists.

Using administrative data: Informing data collection interventions in response to emerging non-response

- When qualitative information suggested that some Head Start programs were indicating barriers to participation
  - We used Head Start and licensing list data to identify sampled addresses that might host Head Start programs and delivered customized information about benefits of participation to only those addresses.
- Some housing units that are challenging to screen are in fact uninhabited
  - Match unscreened addresses against up to date postal service information on possible vacant units. For possibly vacant units, take steps to determine whether the unit is occupied before additional outreach to complete the screener.

Using administrative data: Partially correcting for achieved survey non-response

- **At the close of 2019 data collection, permission to collect data had not been received from many large public school districts. Centers associated with these districts were not missing at random.**
  - We were able to recover high priority data elements such as enrollment, receipt of public funding, and ages of children served from publicly-available administrative data from these school districts. These data elements were incorporated into the survey data files to allow researchers to include these centers in analyses and also to estimate the bias for other data elements that could not be retrieved.

Concluding comments

Mitigation of missingness when conducting surveys

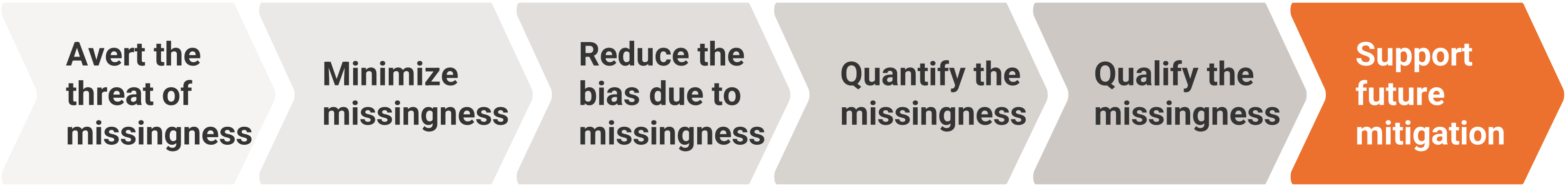
- **We always have missingness, but we want to prioritize mitigation of missingness that harms our estimates and inferences.**

- **Mitigation of missingness can happen at every stage of the survey process**

- **Even flawed data can be valuable when mitigating missingness. Administrative and commercial data are good sources to consider.**

- **Remember the continuum (next slide)**

Mitigation strategy: How can we feasibly minimize the harm that non-response will do to inferences and estimates?


Avert the threat of missingness
Minimize missingness
Reduce the bias due to missingness
Quantify the missingness
Qualify the missingness
Support future mitigation

Thank you.

A Rupa Datta
Distinguished Senior Fellow
Datta-rupa@norc.uchicago.edu

Research You Can Trust™

NORC at the University of Chicago