Missing and Incomplete Data Reporting Guidelines: Overview

HPA Health Analytics

Version 1.0 – Updated August 2024

If you have questions about this document or if you would like this document in other languages, large print, braille, or another format you prefer, please email HPA.IDEA.Team@oha.oregon.gov

OREGON HEALTH AUTHORITY

Contents

Overview

This document describes the guidelines for reporting missing and incomplete data that Health Analytics staff use for external facing reporting and when filling data requests.

Reporting vs. Operationalizing: This document provides guidance on how to treat data for public-facing reporting purposes; it is not intended to guide program operations. The treatment of missing data for business operations (e.g., committee member selection or incentive programs) should be guided by appropriate business rules.

What is missing data and why is it a problem?

Missing data occur when you don't have data for either certain variables ("item level") or for certain respondents in your sample ("unit level"). In addition to the loss of information and therefore reduced statistical power, missing data may also indicate selection bias in the data set. Nearly all of the data that we work with will have missing or incomplete data and it is essential to understand what type of missing data you are working with to treat it appropriately when reporting. **Failure to treat missing or incomplete data appropriately and to report it transparently may result in reporting that is misleading or that may not be generalizable to the population of interest.**

When a respondent provides a response, it is an "active" response. This includes responses like "Decline to answer" and "Unknown." Even though these responses are valid, they may affect analysis of the data in the same way as other missing data. Consequently, the guidelines in this document are applicable for both active responses like "Unknown" and "Decline to answer" and data that are truly missing.

Minimum reporting standards

We cannot point to a single number as the gold standard for determining when to report or suppress data. However, at a minimum the following standards should be followed when working with all Health Analytics data:

1) Release of any data element with missingness rate[1] > 20% must be approved by the team manager or contract manager prior to release.
2) Release of survey data with overall survey response rate < 20% must be approved by the team manager or contract manager prior to release.
3) All reporting should be accompanied by missingness rates at the data element level. When possible, the percentage of blank responses, as well as the percentage of each active response (e.g., "decline to answer" and "unknown") should be reported separately. Survey data should also be accompanied by overall survey response rates.
4) When imputation methods are used, imputation rates and methodology should be thoroughly described (see page 11 for more on imputation).
5) When there is likely bias in the data, this should be noted. Work with the team manager or contract manager to determine if the data should be suppressed with an annotation (e.g., "not available due to selection bias") or if the data may be released. If the data are released, the data user must be informed of any concerns around bias in the data.

Health Analytics staff work with full count data (e.g., claims, enrollment, and hospital discharge data), as well as survey data (e.g., OHIS and CAHPS data). These standards apply to all data regardless of data type (i.e., full count or survey data).

Note: Suppression also may be needed to ensure confidentiality and reliability. All reporting should be done in accordance with the HA Small Numbers Reporting Guidelines.

[1]Missingness includes blank responses and active responses such as "decline to answer and "unknown" (see page 5).

Important concepts

Active responses

When a respondent provides a response, it is considered an active response. This holds true for all responses. E.g., “Decline to answer”, “Don’t understand”, and “Unknown.”

Level of missing data: Unit vs item level

Unit level: Data are missing at the unit of observation level. In most cases, this is at the person level and occurs when no information is available for a respondent. There are many reasons that data may be missing at the unit level (e.g., an individual chooses not to respond to a survey invitation, an individual does not receive a survey invitation because they have moved, or even due to data loss by the data collector).

Item (or question) level: Data are missing at the item or question level. This most commonly occurs when respondents skip or decline to answer *specific questions*, however it can also occur in a data set due to programmed branching or question display logic.

Non-response bias (or self-selection bias)

Bias differs from the issue of precision of estimates in that **bias is non-random error**. Non-response bias occurs when respondents are unwilling or unable to respond to a question (item or question level) or an entire survey (unit level). Non-response bias may occur in survey data as well as administrative data. **Data missing due to non-response bias is unlikely to be missing completely at random.**

Survey response rate

In survey research, the response rate refers to the number of people who answered the survey divided by the number of people in the sample.

Missingness rate

Missingness rate refers to the number of missing responses to a given question divided by the number of survey respondents (for survey data) or number of individuals in an administrative data set (for administrative data) who were presented the question (i.e., structurally missing respondents are excluded). Missing responses include truly missing (blank) responses, as well as active missing responses such as “Decline to answer” and “Unknown”.

Example: You send a survey to 400 people and receive 100 responses. For question #1, 11 respondents left the question blank, and 14 selected ‘decline to answer.’ The missingness rate for question #1 is (11+14)/100, or 25%.

Types of missing data

There are three types of *unintentional* missing data that you may encounter. A fourth type of missing data (‘structurally missing’) should be noted, but it is missing by design and does not warrant the same level of attention:

1) **Missing completely at random (MCAR)**: Data are missing completely at random when there is **no way to attribute the missingness to other data elements**. That is to say, the missingness cannot be explained or predicted by any of the other information. This is the best-case scenario. However, it is generally unsafe to assume that data are missing completely at random.

    Example: Data missing due to equipment malfunction, lost data, or questions that were accidentally skipped by a respondent.

2) **Missing at random (MAR)**: Data are missing at random when **we can attribute the missingness to other data elements observed in the data set, but not the missing values themselves**. That is to say the missingness can be explained or predicted *by other information in the dataset*. Note: This name is a misnomer, in that MAR data are NOT actually missing at random.

   Example: A survey collects information on age and income. Compared to other age groups, many respondents in the youngest age group (18-25 years) did not answer the income question. Provided that the income responses for those in the 18-25 year age range that were collected have a wide distribution, it is possible that those in this age range were less likely to answer the income question. In that case, because the missing income data are related to other values in the data set (i.e., age), these data would be missing at random.

3) **Missing not at random (MNAR) (non-ignorable):** Data are missing not at random when the **missingness can be attributed to the missing values themselves**. That is to say the missingness is explained by the values that are missing, an unknown factor. In this case, data from key sub-groups may be missing from your sample (e.g., due to non-response bias), and your sample will not be representative of your population.

   Example 1: In the same income survey from above, you find few low incomes have been collected. It seems that respondents with low incomes were less likely to respond to this question. Because the missing data are related to the values themselves, these data would be missing not at random.

   Example 2: Looking at the racial demographics for an annual survey, you notice that response rates for some groups are much lower this year than they have been historically. If the response rates for other groups are consistent with past surveys, it is likely that for some reason some groups were less likely to respond to the survey this year.

4) **Structurally missing:** Data are structurally missing when missingness is due to how the data were collected (i.e., the design of the data collection tool or survey).

   Example: In the same income survey as above, income data are missing for respondents under 18 years of age. Because the missing data are a product of the study design, these data would be structurally missing.