Guidance Document:
Missing Data in SEER-CAHPS

Background

Because SEER-CAHPS links data from multiple sources, there are different types of missing data in each data source. The guidance below details intended and unintended missing data, with recommendations for handling each type.

Please note: Consistent with Medicare CAHPS analyses, we recommend *never* imputing CAHPS items or composites.* The guidance below applies to other variables included within the CAHPS survey.

*CAHPS items and composites include: Global ratings of health plan, health care, personal doctor, specialist, and prescription drug plans; Composite measures of Getting Needed Care, Getting Care Quickly, Provider Communication, Customer Service, Getting Needed Prescription Drugs, Care Coordination

Types of missing data

Missing data are often categorized as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). **Table 1** below defines each type and explains how to test for type of missing data. These tests are recommended for each new analytic sample.

Table 1. Categories of Missing Data

Type	Definition	Example	How to Determine
Missing Completely at Random (MCAR)	The propensity for a data point to be missing is completely random.	A survey respondent flips a coin to decide whether to complete a course evaluation.	Little's MCAR test (not totally accurate)
Missing at Random (MAR)	The propensity for a data point to be missing is not related to the missing data, but it is conditional on another variable.	Male respondents are more likely to decline to complete surveys, but declining *does not* depend on their level of satisfaction.	Test for interactions between observed variables: No significant interactions = MAR; Significant interactions = MNAR
Missing Not at Random (MNAR)	The propensity for a data point to be missing is not random.	Respondents with disabilities are less likely to complete surveys.	

Missing data in SEER-CAHPS

In the SEER enrollment-type files (PEDSF and SUMDENOM), missing data are generally designated with a separate category (for example, unknown stage) or a period (".") with no information on why a value might be missing. However, there are low fractions of missing information (FMI <1%) overall, since most of the information comes from administrative records that are largely complete.

In CAHPS, missing data on survey items are designated with a dot that is *sometimes* followed by a letter that provides additional information on why data are missing. It is possible to separate these types of missing data into intended and unintended types:

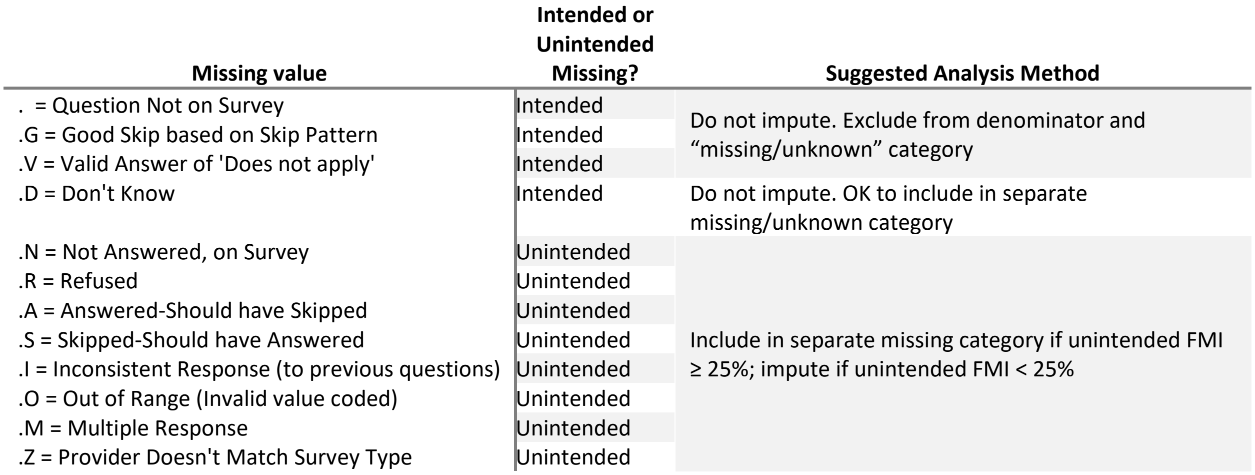
- Intended missing data occurs when the question was not on the survey, or the respondent had a valid skip or a valid answer of "don't know".
  - We recommend that analysts *not* impute these values.
- Unintended missing data arises when a respondent *should* have some data but does not, whether because they skipped it, refused, or gave an invalid response.
  - We recommend that analysts include such response values in a separate missing/unknown analytic category if the unintended FMI ≥ 25%.
  - If the unintended FMI < 25%, we recommend that analysts apply multiple imputation to that variable, if the predictive imputation model appears to have validity (see section III below).

Note that the MCAR, MAR, and MNAR categories are separate from intended/unintended. However, intended missing data are often MAR – for example, missingness is conditional on a variable such as survey year or type, but missingness is unrelated to care experiences.

Conversely, unintended missing data are often MNAR. For example, proxy respondents may skip or answer "don't know" to certain items AND proxy respondents generally perceive care quality as lower than do patient respondents.[1] It is important to note that if MNAR data are handled as if they are MAR or MCAR, analysts are likely to arrive at inaccurate parameter estimates.[2]

Table 2 below lists each type of missing data in CAHPS along with recommendations on how to handle missing values for each.

Table 2. Types of missing data in CAHPS and suggested methods for analysis



Missing value	Intended or Unintended Missing?	Suggested Analysis Method
. = Question Not on Survey	Intended	Do not impute. Exclude from denominator and “missing/unknown” category
.G = Good Skip based on Skip Pattern	Intended	
.V = Valid Answer of 'Does not apply'	Intended	
.D = Don't Know	Intended	Do not impute. OK to include in separate missing/unknown category
.N = Not Answered, on Survey	Unintended	Include in separate missing category if unintended FMI ≥ 25%; impute if unintended FMI < 25%
.R = Refused	Unintended	
.A = Answered-Should have Skipped	Unintended	
.S = Skipped-Should have Answered	Unintended	
.I = Inconsistent Response (to previous questions)	Unintended	
.O = Out of Range (Invalid value coded)	Unintended	
.M = Multiple Response	Unintended	
.Z = Provider Doesn't Match Survey Type	Unintended	

The following text and examples are taken from the SEER-CAHPS documentation:

. Question Not on Survey

Since multiple survey types and survey years are found within a single file and since the surveys were not consistent over time or type, some questions are not found on every survey. If a question was not asked at all, then for those years and types the missing value will be a simple . value.
For example, the question about getting a flu shot this year was only asked in 1998 and 1999 Medicare Advantage (MA) surveys. It would have a . value for 1997 and 2000-2005 MA surveys, which are stored in the same file structure.

If you are calculating percentages missing for a question or percent complete for a respondent, this value should not be included in the Numerator or Denominator.

.G Good Skip based on Skip Pattern

Some questions have leading Skip Pattern questions and the respondent is instructed to skip the question if they answered No.
For example, the question “Were you seen for an illness or injury?” is a Yes/No question. If the respondent answered NO, he should have skipped the question “How often did you get care for an illness or injury as soon as you wanted?” We will use this example again. If the respondent answered NO to being seen for an illness or injury and skipped “How often did you get care as soon as you wanted”, then it is a good skip based on the skip pattern question.

If you are calculating percent complete for a respondent, this value should not be counted against them.

.V Valid Answer of 'Does not apply'

Some questions provided a valid response which effectively means ‘This doesn’t apply to me.’
For example, in the 2009 Medicare Advantage + Prescription Drug Plan (MA PDP) survey, question 37 is “In the last 6 months, how often did the PDP’s customer service give you the information/help you needed about prescription drugs?” The last choice is “I did not try to get information or help from my health plan’s customer service in the last 6 months”. This is a valid response, but was recoded to .V as it does not affect calculations of how satisfied people were in this area.

If you are calculating percentages missing for a question or percent complete for a respondent, this value should not be included in the Numerator or Denominator.

.D Don't Know

Some questions provide a valid response which effectively means “I don’t know.” However, these surveys were also given by phone and sometimes the respondent said they didn’t know. Both these types of responses are classified as .D values.
An example of the survey based response can be found in the 2009 MA PDP survey, question 34, “Have you ever asked anyone at your health plan to reconsider a decision not to provide or pay for health care or services?” The last choice is “Don’t know”. This was coded to .D Don’t know, but is a valid response.

If you are calculating percentages missing for a question or percent complete for a respondent, you should handle this value with care. This link includes CMS CAHPS surveys: https://www.cms.gov/Research-Statistics-Data-and-Systems/Research/CAHPS/

.N Not Answered, on Survey

When a question is on the survey, but the respondent just didn’t answer it, the field will get a .N value. This is used for questions outside of the skip pattern set. For example, if gender was not answered, it would be set to .N value.

.R Refused

This was used for phone-based surveys when the respondent refused to answer the question.

.A Answered-Should have Skipped

As with .G, this value is used for questions associated with a skip pattern question when the respondent answered a question that they should have skipped.
For example, if the respondent answered NO to being seen for an illness or injury and then said he always got care as soon as wanted, then he answered a question that should have been skipped.

.S Skipped-Should have Answered

As with .G, this value is used for questions associated with a skip pattern question when the respondent answered a question that they should have skipped.
For example, if the respondent answered YES to being seen for an illness or injury and then skipped "How often did you get care as soon as wanted", then he skipped a question that should have been answered.

.I Inconsistent Response (to previous questions)
If the respondent provided inconsistent responses outside of a skip pattern, then the 2nd question was set to .I inconsistent response.
For example, in the 2010 MA PDP survey, question 15 is "Do you have a personal Doctor?" If the respondent responded NO to that question, but in question 28, "How often did your personal doctor seem informed about the care you got from specialists?" he answered Never, Sometimes, Usually or Always, that is inconsistent with the original information that he didn't have a personal doctor, and was coded as .I value.

.O Out of Range (Invalid value coded)
If a question had 3 valid responses, but the coded value was 4, the value is out of range. This would be the result of bad coding, but the value is not useable and was reset to be .O out of range.

.M Multiple Response
If a question had multiple responses, for example the respondent answered that he 'Sometimes' and 'Always' got care as soon as wanted it, then it is coded as a .M, multiple responses given.

.Z Provider Doesn't Match Survey Type
For the 2012 MA PPO data, some people were sent this survey type even though the Part D contract was not of the correct type. These respondents had Part D that were known to be from PDP or had Part D where it was unclear whether the contract was from MAPD or PDP. There were 4660 such respondents. All their responses to Prescription Drug related questions were masked with .Z as they were not providing information on the same type of plan as the other respondents of this survey type.

.B Blanked out responses
4660 respondents were sent the MA-PPO survey in 2012. However, based on their contracts, they should have been sent a FFS-PDP survey. Comparison of this data is questionable, so the responses were blanked out.

Table 3 provides FMIs (unintended, intended, and total) for predictor variables among respondents to the **2007-2013** Medicare CAHPS surveys whose data have been linked to SEER data (n=524,929). The file includes people who responded before or after their cancer diagnosis (i.e., those in the SEER PEDSF file) as well as those without cancer histories residing in SEER areas (i.e., those in the SEER SUMDENOM file). It includes both fee-for-service (FFS) and Medicare Advantage (MA) beneficiaries of all ages. Please note that missingness tabulations will vary for other analytic samples.

We have highlighted the 14 variables (out of 46) with 10% unintended missing or greater. The variable with the highest percentage of unintended missing data is self-reported cancer history (21%).

Table 3. Fractions of missing information (FMIs) in the 2007-2013 SEER-CAHPS Analytic File

	Variable name	Availability		FMI: Intended	FMI: Unintended	FMI: Total
		Years	Survey types			
PEDSF/SUMDENOM variables						
Age	age_dx	All	All	0%	0%	0%
Sex	m_sex	All	All	0%	0%	0%
Race	race	All	All	0%	0%	0%
Dual enrollee	duals (sbi flags; sc_dual_status)	All	All	0%	0%	0%
Census region	code_sys; tract	All	All	0%	1%	1%
Urbanicity	urbrur	All	All	0%	0%	0%
Neighborhood poverty	census_pov_ind	All in PEDSF	All	80%	0%	80%
Marital status	mar_stat	All in PEDSF	All	80%	0%	80%
Cancer site	site02v	All in PEDSF	All	85%	0%	85%
Stage at diagnosis	dajcc7t	All in PEDSF	All	78%	0%	78%
Primary/first cancer is malignant	firstprm	All in PEDSF	All	78%	0%	78%
CAHPS variables						
Lives alone	living_alone	All	All	26%	7%	33%
Education	education	All	All	6%	2%	8%
Proxy assistance	proxy	All	All	5%	20%	25%
Self-reported poor/fair general health status	ghs	All	All	0%	5%	5%
Self-reported poor/fair mental health status	mhs	All	All	0%	5%	5%
Has had a condition/problem lasting 3+ months	cnd_last3mo10	All	All	45%	20%	65%
Smoke now	smokenow	All	All	0%	7%	7%
Physical condition interferes with independence	lim_physcond	2007-2010	All	42%	8%	50%
Need help with personal care	helpperscare	2007-2010	All	42%	8%	50%
Need assistance with routine activities	helproutine	2007-2010	All	42%	8%	50%
Limitations in activities of daily living (ADLs)						
Bathing	lim_bathing	MA: 2007-2010; FFS: all	MA, FFS	26%	11%	37%

	Variable name	Availability		FMI: Intended	FMI: Unintended	FMI: Total
		Years	Survey types			
Dressing	lim_dressing	MA: 2007-2010; FFS: all	MA, FFS	26%	12%	38%
Eating	lim_eating	MA: 2007-2010; FFS: all	MA, FFS	26%	12%	38%
Using chairs	lim_chairs	MA: 2007-2010; FFS: all	MA, FFS	26%	11%	37%
Walking	lim_walking	MA: 2007-2010; FFS: all	MA, FFS	26%	10%	36%
Using the toilet	lim_toilet	MA: 2007-2010; FFS: all	MA, FFS	26%	12%	38%
Climbing stairs	lim_climb	2007	All	88%	6%	94%
Moderate activities	lim_modact	2007	All	88%	5%	93%
Regular activities	lim_regact	2007	All	88%	6%	94%
Social activities	lim_socacts	2007	All	88%	8%	96%
Pain interferes	lim_painint	2007	All	88%	7%	95%
Little/no energy most of the time	sf_energy	2007	All	88%	7%	95%
Self-reported cancer	cnd_cancer	2008-13	All	14%	21%	35%
Self-reported heart attack or angina	cnd_heartattack; cnd_angina	2008-13	All	14%	19%	33%
Self-reported stroke	cnd_stroke	2008-13	All	14%	21%	35%
Self-reported COPD	cnd_cpod	2008-13	All	14%	20%	34%
Self-reported diabetes	cnd_diabetes	2008-13	All	14%	16%	30%
Depression (PHQ-2 ≥ 3)	ds_phq2	2009	All	82%	3%	85%
SF-12 Physical Component Score (PCS)	sc_pcs_vr12	2007-2010	All	42%	20%	62%
SF-12 Mental Component Score (MCS)	sc_mcs_vr12	2007	All	88%	4%	92%
3+ doctor visits for same condition, past 12 mos.	cnd_md3time	All	All	0%	7%	7%
Taking Rx for any condition	cnd_rxmeds	All	All	0%	6%	6%
Delayed/didn't get Rx because of cost, past 6 mos.	rx_delay	All	All	4%	7%	11%
Overnight hospital stay	pl_hospovn	2013	All	87%	6%	93%

Considerations for Imputation

The main goals of the strategies for handling missing data are to minimize bias, maximize use of available information, and generate appropriate estimates of uncertainty (such as standard errors or confidence intervals). Many books and articles have been written about imputation. Common approaches to dealing with missing data include:

- **Complete case analysis** (also known as listwise deletion)

  - **Approach**: Drop cases with missing data on any variable of interest (done automatically in most software packages)
  - **Drawbacks:** loss of data/observations; biased estimates unless data are MCAR

- **Unconditional mean imputation**
  - **Approach**: Replace missing values for a variable with its overall estimated mean
  - **Drawbacks:** Artificially reduces variability; changes correlations between variables
- **Singular regression-based imputation**
  - **Approach**: Replace missing values with predicted scores from a regression equation
  - **Drawbacks:** Decreases variability; underestimates uncertainty; may have dubious face validity if regression model does not fit data well (e.g., if the R^2 is low); inflates correlation between variables and biases R^2 statistics from analysis of imputed data
- **Stochastic imputation**
  - **Approach**: Add randomly drawn residual to imputed value from regression imputation. Distribution of residuals based on residual variance from regression model.
  - **Drawbacks:** Standard errors are still attenuated (biased downward)
- **Multiple imputation**
  - **Approach**: Multiple values are imputed rather than a single value to reflect the uncertainty around the “true” value. Each imputed value includes a random component whose magnitude reflects the extent to which other variables in the model cannot predict its “true” value. Variants include multiple imputation with chained equations (MICE) and Fully Conditional specifications that do not assume normal distributions for all variables and allow for different types of regression (linear, logistic, etc) for imputation.
  - **Drawbacks:** Auxiliary variables need to be correlated with missing variable (rule of thumb: $r \geq 40\%$. Biased estimates may result when N is relatively small and the FMI is high. Requires substantial computing power for larger Ns. Assumes data are MAR.

Newer methods of imputation are gaining proponents. Among them is multiple imputation using various machine learning methods, such as random forests (RF).[3] Some researchers have found that RF imputation produces less biased results with narrower confidence intervals than regression-based imputation.[4] Evidence suggests that RF-based imputation methods may be theoretically sound even for large percentages of missing values (up to 50%).[3,5]

Conclusions

In the SEER-CAHPS 2007-2013 sample, a little less than a third of major predictor variables had more than 10% *unintended* missing data, and none had more than 21%. However, when combining both *intended* and *unintended* missingness types, up to 96% of respondents have missing data; some variables, such as limitations in social activities, have particularly high total FMIs because they were asked in only one year.

One question that is often raised by reviewers is how much data are missing from particular covariates. We would advise that analysts using the SEER-CAHPS data distinguish between intended and unintended missing when tabulating missingness in their articles for publication. This may pre-emptively address concerns about missing data that are endemic to survey research.

Distinguishing between intended and unintended missing data is challenging but important in any analysis. It is particularly important when using methods that impute missing data by default. Analysts using the SEER-CAHPS data resource would be advised to decide in advance whether to use imputation and how to account for missing data on key predictors.

References

1. Elliott MN, Beckett MK, Chong K, Hambarsoomians K, Hays RD. How Do Proxy Responses and Proxy-Assisted Responses Differ from What Medicare Beneficiaries Might Have Reported about Their Health Care? *Health Services Research.* 2008;43(3):833-848.
2. Stockdale M, Royal K. Missing data as a validity threat for medical and healthcare education research: problems and solutions. *Int J Health Care.* 2016.
3. Tang F, Ishwaran H. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal.* 2017;10(6):363-377.
4. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American journal of epidemiology.* 2014;179(6):764-774.
5. Bodner TE. What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal.* 2008;15(4):651-675.