Family Practice, 2021, 1–4
doi:10.1093/fampra/cmaa134

OXFORD

CASFM Methods Briefs

Missing data in primary care research: importance, implications and approaches

Miguel Marino[a,b,*], Jennifer Lucas[a], Emile Latour[c] and John D. Heintzman[a,d]

[a]Department of Family Medicine, Oregon Health & Science University, Portland, OR, USA [b]School of Public Health, Oregon Health & Science University – Portland State University, Portland, OR, USA [c]Knight Cancer Institute, Oregon Health & Science University, Portland, OR, USA [d]OCHIN, Portland, OR, USA

*Correspondence to Miguel Marino, Department of Family Medicine, Division of Biostatistics, Oregon Health & Science University, 3181 SW Sam Jackson Park Rd, Portland, OR 97239, USA; E-mail: marinom@ohsu.edu

Introduction

Missing data are a common occurrence in all study designs used to conduct primary care research (pragmatic randomized trials, observational studies, quality improvement efforts, etc.). We define missing data as values not available to the investigator that would have contributed to the final analysis had they been observed. Examples of missing data include patients lost to follow-up, partially filled-out surveys or incomplete medical records. Missing data can compromise the validity of study findings (e.g. risk for bias increases, subgroups of the population may be underrepresented, loss of information, reduced statistical power, etc.) (1–5). Multiple studies have shown that high rates of missing data may negatively impact conclusions in primary care studies (6–8).

A holistic approach for addressing missing data should include approaches in the design phase (prior to the study), conduct phase (during the study) and in the analytic phase (after data collection is complete). This article presents key concepts regarding types of missing data, and discusses good practices to observe when conducting primary care research with missing data.

Dealing with missing data in the design phase (prior to the study)

Whatever approach is used to address missing data, all require making assumptions about the mechanism that resulted in missing data. Determining these mechanisms can be imprecise and impossible to verify; thus, taking steps to prevent missing data in the first place is just as important as missing data analytic approaches.

There are numerous and unique randomized trial designs in primary care; these designs use administrative, electronic health record (EHR) and patient-reported data. In the design of randomized trials, there is extensive literature that provides guidance on how to avoid missing data (3,9). Generally, these include: (1) identifying target populations underserved by interventions that may have incentives to remain in the study; (2) under appropriate research questions, utilizing patient-centered (10) and adaptive interventions (11) to meet the specific needs of different clinics, their patients and avoid patient/clinic dropout by motivating patients/physicians to stay engaged throughout the study; (3) utilizing designs that add interventions to existing accepted interventions; (4) selecting primary outcomes that are less susceptible to extensive missing data (e.g. easily defined, standardized quality indicators or clinical outcomes, biomarkers that have widely available standardized protocols for collection, and brief, validated, language and literacy-appropriate survey instruments).

For observational studies that utilize administrative and clinical data such as EHRs, a common data source in primary care research, several considerations can be utilized to reduce missingness including: (1) focusing on variables that are consistently collected as part of the routine clinical practice, (2) designing clear data collection procedures focused on only meaningful data to minimize burden on those extracting data, (3) identifying and linking several data sources for better capture of important variables (12,13) (e.g. race/ethnicity fields in administrative and EHR data can be linked and used to fill in missing values (14)). The STROBE guidelines provide great guidance on what to consider as one designs observational studies (15) and provide suggestions for how to document missing data.

Avoiding missing data in survey research has been extensively detailed (16,17) and we refer the reader to those citations. In short, there is an emphasis on reducing survey burden (e.g. limiting survey length) and pilot testing surveys before large-scale implementation to ensure capture of elements that address the main research question, as well as information that might be useful in approaches to address missing data.

Dealing with missing data in the conduct phase (during the study)

For study designs with prospective data collection, missing data can also occur during the study (prior to analysis). Table 1 provides ideas

Table 1. Suggestions for reducing missing data in the conduct phase of a study

Consider multiple methods of assessment. If a patient can't come to the clinic to get assessed for study measures, consider virtual, at-home visits, self-administered surveys, etc.
Consider auxiliary data sources to retrieve data. If patient consent is given, consider pulling some measures from their electronic health records, chart review, etc.
If dropout occurs, try to collect reason for dropout. If a patient or clinic decides to drop out, consider engaging them one last time to capture reason for withdrawing and utilize that information to include in analytic phase or to provide context for study findings.
Monitor data collection. Performing data checks regularly during the conduct of the study can identify missing data issues and prompt action to address missing data. Action can include modifications to the process of data collection and additional training.
Collaborate with experts. Recruit and engage study team members who have a strong track record of enrolling and successfully following patients/clinics in previous studies.
Remind study team and participants about the importance of complete data. Periodically, remind everyone that information collected is important regardless of whether the patient or clinic continue on the assigned study arm.
Update contact information. Check in with clinics and/or patients to keep their information up to date in order to limit clinics and patients lost to follow-up.
Employ engagement strategies. For clinics and participants, utilize newsletters, blogs and incentives to get survey completion (e.g. random draw prize for survey completion, vouchers, etc.).
Work with clinics to understand basic clinical research. Within clinic constraints, work with clinicians and clinical staff to review the basic principles of research methods and data collection including the importance of complete data.

based on published research (9,18,19) on how to minimize missing data during this phase.

Dealing with missing data in the analysis phase (after the study)

Even with much care during the study, missing data will be unavoidable. It is good practice to plan one's analytic methods to address missing data in the design phase. There is no single best method to address missing data that is appropriate for all circumstances as study designs and reasons for missing data may vary. Prior to selecting a method, it is important to identify the potential mechanism leading to missing data.

Types of missing data

Missing data can be classified into three types: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). When data are MCAR, the probability of being missing is equal for all observations. If data are MAR, this suggests that the probability of being missing is equal only within groups defined by observed data. When data are MNAR, the probability of being missing varies for unobservable reasons (20). To illustrate these concepts, consider a scenario where a researcher is interested in studying if there are disparities in colorectal cancer screening by race/ethnicity in their health system. If a sample of patient faecal occult blood tests (FOBT) mail-in kits were lost in the mail due to a random glitch, we would say that the data are MCAR. Missing data scenarios that have nothing to do with the patient being studied (e.g. weight scale breaks down, lost mail, etc.) are typically considered MCAR. Most simple solutions for missing data (e.g. complete case analysis, missing-indicator method, single-value imputation, sensitivity analyses with best/worst case scenarios) only provide unbiased results under the MCAR assumption, though estimates may be less precise (21–23).

Data are MAR if the missingness does depend on the patient but the missing component can be predicted from other information about the patient. For example, if a patient does not mail in the FOBT kit because the instructions are in English and they speak Spanish and you know the patient's preferred language, and the reason that the patient did not send in the kit was not related to the outcome of screening, this missing data are MAR. In MAR settings, one can leverage other available data to address missing data; most statistical methods that address missing data assume MAR.

Scenarios that do not meet MCAR or MAR fall into MNAR. When data are MNAR, it is because the missing data are specifically related to what is missing. For example, if a patient does not mail in the FOBT kit because they were uncomfortable with specimen collection, we would say that these missing data are MNAR; anytime the missing data are related to what is missing is an indication of MNAR. This setting is the most problematic and is the most complex situation to handle. Within a study, data may be both MAR and/or MNAR, and rarely is it MCAR.

General analytic approaches to address missing data

There are four common types of methods that address missing data (9): (1) complete case analyses, (2) single imputation, (3) inverse probability weighting and (4) multiple imputation.

Complete case analysis (i.e. listwise deletion) is a method that removes all patients with any missing data from the analysis. This is typically the default approach to handle missing data in many statistical software programs and while it is convenient to implement, it relies on the often-unrealistic assumption that the data are MCAR which can result in biased estimates and reduced statistical power (3).

Single imputation methods are methods that fill in missing data with a single value. Examples include last observation carried forward (e.g. in a longitudinal study, fill in missing values with the last value you observed), mean imputation (e.g. if missing body mass index [BMI], estimate the mean from the patients with observed BMI and use that mean to fill in the missing values), among others.

Inverse probability weighting is a method that utilizes only complete cases (i.e. patients who had fully observed data). These complete cases are weighted by the inverse of their probability of being a complete case (24–26). Intuitively, complete cases may be different than those with missing data and thus may not be representative of the broader population. Thus, one can assign larger weights to patients who are underrepresented in the sample and lower weights to those that are overrepresented with the goal to make the sample look more like the population.

Multiple imputation (27), applicable to cross-sectional and longitudinal designs (28), is an extension of single imputation that fills in missing data, typically using standard regression methods, with multiple plausible values that account for the uncertainty of the imputed values (29). Figure 1 shows the main three stages involved in multiple imputation: (1) generate *m* imputed data sets according to the a priori analytic study plan, (2) analyse the *m* imputed data sets, (3) pool the results from the *m* analyses using multiple imputation combining rules to yield a single estimate (30). Starting with observed, incomplete data, multiple imputation creates several versions of the data by replacing missing values with plausible ones. The parameters of interest are estimated for each of the imputed data sets. Then the multiple parameter estimates are pooled into one estimate. While there are no universal rules for how many imputations to perform, it is recommended that the number of imputations should be similar to the percentages of incomplete cases (31,32).

In most cases, single imputation methods are considered an improvement to complete case analyses but they do not reflect the uncertainty in the imputations and thus can lead to standard errors that are too small and have the potential for incorrect conclusions (33). Inverse probability weighting and multiple imputation approaches are generally recommended as they often assume a MAR missing data mechanism and thus can use supplementary information about the missing data in the final analysis. In practice, a MAR assumption is more likely to be valid as compared to a MCAR assumption, especially when all relevant important variables are collected.

How much missing data are too much?

There are no universal guidelines for the amount of missing data that make statistical inference is valid. Several characteristics play a role including the amount of missingness (e.g. percentage of data missing), the correlation between cause of missingness and variable containing missingness and the correlation between cause of missingness and missingness itself (3). In general, the lower percentage of missing data, the better. Missing percentages of ≤5% are thought to be trivial (27). One study found that an analysis is likely to be biased if 10% or more of the data are missing (34). Another suggests that if more than 40% of data on important variables are missing, the results should be considered hypothesis-generating rather than confirmatory (35). Lastly, one simulation study showed that the proportion of missing data should not be used as a guide to whether to use a method like multiple imputation or not. Instead, they encouraged the use of methods such as multiple imputation to reduce bias and improve efficiency at any proportion of missing data (36).

In survey research, acceptable response rates vary. One 'rule of thumb' is 60% (37). Some agencies may ask for the use of survey methods that result in a 75% response rate yet response rates lower than 30% are seen in published studies (38). Response bias—the idea that respondents are different than non-respondents and not representative of the whole survey population—is more likely in surveys with lower response rates (38).

Even in settings with minimal missingness, it is generally recommended that an analytic method that addresses missing data should be employed. Though outside the scope of this primer, sensitivity analyses that assume MNAR should be performed to check the robustness of study findings, such as pattern-mixture models (39) and selection models (40,41).

Summary

Primary care studies are multifaceted and will encounter missing data at the patient, clinician or clinic level that could negatively impact conclusions. Investigators, in collaboration with statisticians, should include missing data approaches or their mitigation in the design phase (prior to the study), conduct phase (during the study)

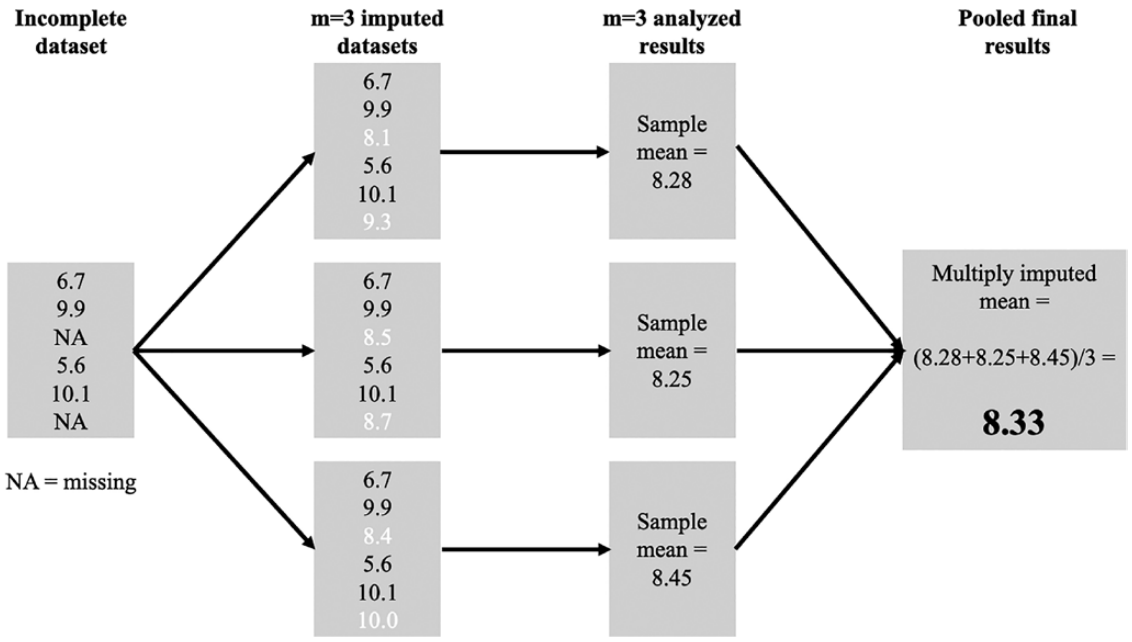

Figure 1. Stages of multiple imputation with an illustrative example where the mean haemoglobin A1c (HbA1c) among patients is the parameter of interest and m = 3 imputations were performed. Note: Multiple imputation begins with an incomplete data set. Multiple imputed data sets are created by replacing missing values with plausible values based on a random draw from a distribution that is specifically modelled for the missing values, typically using regression models. Analysis is done on each imputed data set to estimate a parameter of interest. These estimates are pooled into one single estimate of the parameter and its variance.

and in the analytic phase (after the study has finished data collection). Understanding the missing data mechanism (MCAR/MAR/MNAR) is important to guide the researcher in understanding study limitations, and identifying the appropriate analytic approach to employ. Several methods are available to address missing data and one or more should always be used. By ensuring that missing data are properly addressed, primary care research studies will continue to increase rigour and inform evidence-based practice.

Declaration

Funding: This work was supported by grant numbers R01MD011404 and R01MD014120 from the US National Institute for Minority Health and Health Disparities and by grant number R01AG056337 from the US National Institute for Aging.
Ethical approval: none.
Conflict of interest: none.

References

1. Papageorgiou G, Grant SW, Takkenberg JJM, Mokhles MM. Statistical primer: how to deal with missing data in scientific research? *Interact Cardiovasc Thorac Surg* 2018; **27**(2): 153–8.
2. Little RJ, D'Agostino R, Cohen M, *et al*. The prevention and treatment of missing data in clinical trials. *N Engl J Med* 2012; **36**(14): 1355–60.
3. Graham JW. Missing data analysis: making it work in the real world. *Annu Rev Psychol* 2009; **60**: 549–76.
4. Joseph R, Sim J, Ogollah R, Lewis M. A systematic review finds variable use of the intention-to-treat principle in musculoskeletal randomized controlled trials with missing data. *J Clin Epidemiol* 2015; **68**(1): 15–24.
5. Howe CJ, Cole SR, Lau B, Napravnik S, Eron JJ Jr. Selection bias due to loss to follow up in cohort studies. *Epidemiology* 2016; **27**(1): 91–7.
6. Petersen I, Welch CA, Nazareth I *et al*. Health indicator recording in UK primary care electronic health records: key implications for handling missing data. *Clin Epidemiol* 2019; **11**: 157–67.
7. Stiglic G, Kocbek P, Fijacko N, Sheikh A, Pajnkihar M. Challenges associated with missing data in electronic health records: a case study of a risk prediction model for diabetes using data from Slovenian primary care. *Health Informatics J* 2019; **25**(3): 951–9.
8. Power MJ, Freeman C. A randomized controlled trial of IPT versus CBT in primary care: with some cautionary notes about handling missing values in clinical trials. *Clin Psychol Psychother* 2012; **19**(2): 159–69.
9. National Research Council. The prevention and treatment of missing data in clinical trials. *Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press, 2010.
10. Mullins CD, Vandigo J, Zheng Z, Wicks P. Patient-centeredness in the design of clinical trials. *Value Health* 2014; **17**(4): 471–5.
11. Almirall D, Nahum-Shani I, Sherwood NE, Murphy SA. Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. *Transl Behav Med* 2014; **4**(3): 260–74.
12. Marino M, Angier H, Valenzuela S *et al*. Medicaid coverage accuracy in electronic health records. *Prev Med Rep* 2018; **11**: 297–304.
13. Angier H, Gold R, Gallia C *et al*. Variation in outcomes of quality measurement by data source. *Pediatrics* 2014; **133**(6): e1676–82.
14. Grundmeier RW, Song L, Ramos MJ *et al*. Imputing missing race/ethnicity in pediatric electronic health records: reducing bias with use of U.S. census location and surname data. *Health Serv Res* 2015; **50**(4): 946–60.
15. Vandenbroucke JP, von Elm E, Altman DG *et al*.; STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Med* 2007; **4**(10): e297.
16. De Leeuw ED. Reducing missing data in surveys: an overview of methods. *Qual Quant*. 2001; **35**(2): 147–60.
17. De Leeuw ED, Hox JJ, Huisman M. Prevention and treatment of item nonresponse. *J Off Stat*. 2003; **19**:153–76.
18. Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol* 2013; **64**(5): 402–6.
19. Mack C, Su Z, Westreich D. *Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide*. 3rd edn. [Internet]. Rockville, MD: Agency for Healthcare Research and Quality (US), 2018. Approaches to Prevent Missing Data. https://www.ncbi.nlm.nih.gov/books/NBK493616/
20. van Buuren, S. *Flexible Imputation of Missing Data*. 2nd edn. Boca Raton, FL: CRC Press, 2018.
21. Pedersen AB, Mikkelsen EM, Cronin-Fenton D *et al*. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol* 2017; **9**: 157–66.
22. Groenwold RH, White IR, Donders AR, Carpenter JR, Altman DG, Moons KG. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ* 2012; **184**(11): 1265–9.
23. Knol MJ, Janssen KJ, Donders AR *et al*. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *J Clin Epidemiol* 2010; **63**(7): 728–36.
24. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res* 2013; **22**(3): 278–95.
25. Sun B, Perkins NJ, Cole SR *et al*. Inverse-probability-weighted estimation for monotone and nonmonotone missing data. *Am J Epidemiol* 2018; **187**(3): 585–91.
26. Perkins NJ, Cole SR, Harel O *et al*. Principled approaches to missing data in epidemiologic studies. *Am J Epidemiol* 2018; **187**(3): 568–75.
27. Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res* 1999; **8**(1): 3–15.
28. Kalaycioglu O, Copas A, King M, Omar RZ. A comparison of multiple-imputation methods for handling missing data in repeated measurements observational studies. *J R Stat Soc A* 2016; **179**: 683–706.
29. Kontopantelis E, White IR, Sperrin M, Buchan I. Outcome-sensitive multiple imputation: a simulation study. *BMC Med Res Methodol* 2017; **17**(1): 2.
30. Little RJ, Rubin DB. *Statistical Inference with Missing Data*. 2nd edn. New York: Wiley, 2002.
31. Bodner TE. What improves with increased missing data imputations? *Struct Equ Modelling* 2008; **15**(4): 651–75.
32. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011; **30**(4): 377–99.
33. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 2011; **20**(1): 40–9.
34. Bennett DA. How can I deal with missing data in my study? *Aust N Z J Public Health* 2001; **25**(5): 464–9.
35. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. *BMC Med Res Methodol* 2017; **17**(1): 162.
36. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. *J Clin Epidemiol* 2019; **110**: 63–73.
37. Johnson TP, Wislar JS. Response rates and nonresponse errors in surveys. *JAMA* 2012; **307**(17): 1805–6.
38. Draugalis JR, Coons SJ, Plaza CM. Best practices for survey research reports: a synopsis for authors and reviewers. *Am J Pharm Educ* 2008; **72**(1): 11.
39. Thijs H, Molenberghs G, Michiels B, Verbeke G, Curran D. Strategies to fit pattern-mixture models. *Biostatistics* 2002; **3**(2): 245–65.
40. Kenward MG. Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Stat Med* 1998; **17**(23): 2723–32.
41. Leurent B, Gomes M, Faria R, Morris S, Grieve R, Carpenter JR. Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: a tutorial. *Pharmacoeconomics* 2018; **36**(8): 889–901.