Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.JournalofSurgicalResearch.com

Research review

Missing data in surgical data sets: a review of pertinent issues and solutions

Sherene E. Sharath, PhD, MPH,[a] *Nader Zamani, MD,*[a] *Panos Kougias, MD, MSc,*[a] *and Soeun Kim, PhD*[b,*]

[a] *Division of Vascular Surgery and Endovascular Therapy, Michael E. DeBakey Department of Surgery, Baylor College of Medicine/Michael E. DeBakey Veterans Affairs Medical Center, Houston, Texas*

[b] *Department of Biostatistics and Data Science, University of Texas Health Science Center – School of Public Health, Houston, Texas*

ARTICLE INFO

Article history:
Received 26 January 2018
Received in revised form
30 March 2018
Accepted 11 June 2018
Available online xxx

Keywords:
Statistical methodology
Multiple imputation
Missing data
Complete case analysis
Single imputation

ABSTRACT

Incomplete data is a common problem in research studies. Methods to address missing observations in a data set have been extensively researched and described. Disseminating these methods to the greater research community is an ongoing effort. In this article, we describe some of the basic principles of missing data and identify practical, commonly used methods of adjustment relevant to surgical data sets. Through an example data set, we compare models generated through complete case analysis, single imputation (SI), and multiple imputation (MI). We also provide information on the steps to conduct MI using Stata IC. In our comparisons, we found that differences in odds ratios were greatest between the results from complete case analysis compared to the SI and MI models indicating that in this case the reduction in statistical power has a non-negligible effect on the parameter estimates. Odds ratio estimates from the SI and MI methods were largely similar. In some instances, when compared to the MI method, the SI method tended to overestimate effect sizes. While in this example the differences in odds ratios do not vary greatly between the SI and MI methods, there are clear indications supporting the use of MI over SI. By describing the issues surrounding missing data and the available options for adjustment, we hope to encourage the use of robust imputation methods for missing observations.

Introduction

Study databases are sought after sources from which exploratory analyses describe associations, generate thoughtful prospective trials, and ultimately help guide clinical practice. As research questions increase in complexity, so does the methodology that propels studies to generate accurate results that minimize bias and maximize generalizability. Given that the merit of a study depends heavily on *a priori* design and appropriate statistical analyses, researchers are trained in a system of checks and balances that magnify the impact of results by reviewing common threats to validity and precision throughout the conduct of the study.

One such threat, that is almost universal in its occurrence, is missing data. Missing data can insidiously bias results by decreasing the precision of parameter estimates and reducing

* *Corresponding author*. Department of Biostatistics and Data Science, University of Texas Health Science Center – School of Public Health, 1200 Pressler Street, Houston, TX 77030. Tel.: +713 500.9567.
E-mail address: Soeun.S.Kim@gmail.com (S. Kim).

0022-4804/$ – see front matter
https://doi.org/10.1016/j.jss.2018.06.034

the statistical power of a study.[1,2] The presence of missing data can significantly impact generalizability and the degree to which study conclusions reflect the truth. The consequences of missing data are magnified if the issue is ignored or handled incorrectly. Few studies can claim the achievement of a complete data set without any missing observations. In surgical research, commonly used data sets provided by the Veterans Affairs Surgical Quality Improvement Program and the National Inpatient Sample are subject to the limitations of missing data. For example, patient medical history and lab values such as preoperative serum albumin, blood urea nitrogen, hematocrit, and body mass index are particularly susceptible to unavailability in these data sets.[3,4] While the National Surgical Quality Improvement Program uses predictive model-based multiple imputation (MI) methods to account for missing data,[4] other existing data sets require the end-user to account for missing observations. The absence of covariate observations contributes to an incomplete understanding of the role of these factors in associative and predictive relationships. Despite the other limitations inherent to retrospective data analyses, adjusting for the missing data-related biases within these data sets can maximize overall study strength and therefore, the derived conclusions. Although the occurrence of missing data is inescapable, it can be accounted for throughout the design, implementation, and analysis of a study. Appropriate handling of missing data is a crucial component of efforts to contribute accurate findings to the surgical literature. The goal of this article is to succinctly describe the types of missing data and provide reliable solutions to easily improve the validity of a study's results.

Patterns and mechanisms of missing data

When the reason for missing data is known, or if there is a reasonable assumption that the data are missing at random, appropriate corrective measures may be identified to minimize the negative impact of unavailable data. Before steps may be taken to account for missing observations, it is critical to examine the underlying patterns and mechanisms of the missing data.[5,6] The patterns and mechanisms are two separate constructs that work together and define the nature of the missing data. Patterns describe how missing data appear in a data set (i.e., how the observed and missing values appear in a data matrix[7]). Missing data patterns may be described as follows: (1) the absence of observations in only one variable in the entire data set (univariate nonresponse); (2) missing observations in the same multiple variables throughout the data set (multivariate two patterns); (3) the absence of repeated measures after participants drop out in longitudinal studies (monotone); or (4) no pattern in the missing observations (general pattern).[7] On the other hand, mechanisms deal with how the missing data are related to observed variables. Data may be (1) missing completely at random (MCAR); (2) missing at random (MAR)[8]; or (3) missing not at random (MNAR).[2,5,6,9] Each mechanism and pattern, in turn, has underlying assumptions that impact the strategy that is used to mitigate the effect of missing observations.

MCAR data describe a mechanism in which complete and incomplete data for cases of a given variable are both random samples of the total. In this situation, there is no association between a given observation and that observation's propensity for being missing.[5,10] In other words, missing observations are assumed to be a random sample of the population. Unlike MCAR data, however, MAR not only refers to the instances where missingness does not depend on unobserved data but also accounts for whether a researcher has access to data that can account for the missingness.[5] If a variable related to the missingness is available and included in the model, then any remaining missing observations for that variable are assumed to be MCAR.[5] Therefore, MAR is defined as missingness that is dependent on observed data.

Data that are not missing at random, however, are characterized by either their dependence on a variable that has not been measured or the possibility that the reason data are unavailable is inherently related to the variable itself. In other words, the researcher may know the possible cause of the missing data, but since it is unobserved, it cannot be included in the model. On the other hand, as previously described, including such a variable in the model would make the mechanism MAR. Thus, when the missingness is dependent on unknown or unobserved variables, the data are said to be MNAR.[5,6] Though MCAR and MAR mechanisms are often referred to as "ignorable," MNAR mechanisms are "non-ignorable" and should be handled appropriately using methods developed specifically for MNAR. These methods include selection models[11] or pattern mixture models.[12]

Handling missing data

Accounting for missing data in the design of a study

Because missing observations in any data set are expected, countermeasures can be implemented from the design phase of a study. As MNAR data are, arguably, the hardest and most important type of missing data to work with, many strategies in the design phase address reducing MNAR. The first strategy is something as simple as measuring any potential cause of absent observations. This allows MAR mechanism to be assumed instead of MNAR.[5] For example, if it is expected that a large proportion of observations for the body mass index variable will be missing because either the data are unavailable or participants decline to share weight or height information, collecting information on why the values are absent (whether due to participant refusal or not) could inform the choice of missing data model. Collecting data on auxiliary variables is a second strategy that can reduce estimation bias. Auxiliary variables are factors that are highly correlated ($r > 0.40$[13-15]) with one or more independent variables that can later be used to predict observations for a given variable of interest. Therefore, it is helpful to inlcude such variables in the missing data model. A third method is to collect data from a random sample of individuals whose data are missing.[5] However, this method requires that the individuals are tracked, available, and willing to provide data—a problematic trifecta.

Finally, in some instances, missing data can be designed into the study as a means of collecting more data—this design is known as a planned missing data design.[5] An example is the 3-form design, most useful when collecting data using