JAMA Guide to Statistics and Methods

Missing Data
How to Best Account for What Is Not Known

Craig D. Newgard, MD, MPH; Roger J. Lewis, MD, PhD

Related article page 884

Missing data are common in clinical research, particularly for variables requiring complex, time-sensitive, resource-intensive, or longitudinal data collection methods. However, even seemingly readily available information can be missing. There are many reasons for "missingness," including missed study visits, patients lost to follow-up, missing information in source documents, lack of availability (eg, laboratory tests that were not performed), and clinical scenarios preventing collection of certain variables (eg, missing coma scale data in sedated patients). It is particularly challenging to interpret studies when primary outcome data are missing. However, many methods commonly used for handling missing values during data analysis can yield biased results, decrease study power, or lead to underestimates of uncertainty, all reducing the chance of drawing valid conclusions.

In this issue of *JAMA*, Bakris et al evaluated the effect of finerenone on urinary albumin-creatinine ratio (UACR) in patients with diabetic nephropathy in a randomized, phase 2B, dose-finding clinical trial conducted in 148 sites in 23 countries.[1] Because of the logistical complexity of the study, it is not surprising that some of the intended data collection could not be completed, resulting in missing outcome data. Bakris et al used several analysis and imputation techniques (ie, methods for replacing missing data with specific values) to assess the effects of different approaches for handling missing data. These methods included complete case analysis (restricting the analysis to include only patients with observed 90-day UACR values); last observation carried forward (LOCF; typically this involves using the last recorded data point as the final outcome; Bakris et al used the higher of 2 UACR values and, separately, the most recent UACR obtained prior to study discontinuation); baseline observation carried forward (using the baseline UACR value as the outcome UACR value, therefore assuming no treatment effect for that patient); mean value imputation (replacing missing values with the mean of observed UACR values); and random imputation (using randomly selected UACR values to replace missing UACR values).[1] Multiple imputation[2] to handle missing values was also performed. With the exception of multiple imputation, each of the imputation approaches replaces a missing value with a single number (termed "single" or "simple" imputation) and can threaten the validity of study results.[3,4] The authors concluded that finerenone improved the UACR, a result that was consistent regardless of the method for handling missing data.

Use of the Method

Why Are These Methods Used?

It is rare for a research investigation not to have any missing data. If patients with missing variables are omitted from an analysis, the effective sample size is reduced and the treatment effect estimate may be incorrect.[3] This is known as complete (observed) case analysis and is the default methods used by most statistical software.

Strategies for handling missing values are each based on different assumptions and have different limitations. Key questions to consider when selecting a method for handling missing values include (1) Why are data missing? (2) How do patients with missing and complete data differ? and (3) Do the observed data help predict the missing values? To better understand this last concept, suppose a physician was asked to make a best guess about a characteristic of one of her patients that was missing from their chart; eg, weight, systolic blood pressure, fasting serum cholesterol, or serum creatinine. The chance of guessing a value close to the true value would likely be substantially improved if the physician was given related data about the patient, such as his or her age, comorbidities, and prior laboratory values.

The cause for missing data, called censoring, is "noninformative" when the reason a value could not be measured provides no information for what it should be. Censoring is "informative" when the absence of a value indicates something about what it should be. For example, a patient lost to follow-up may have quit the study because declining health made traveling to follow-up visits more difficult, implying that patients with complete follow-up data may have better health status than those with missing data.

There are 3 ways by which data may be missing.[3,4] The first is that data may be missing completely at random (MCAR), meaning the probability of being missing is completely unrelated to all observed and unobserved patient characteristics. This is the least plausible mechanism but is the only one for which complete case analysis will yield unbiased results.

The next mechanism, missing at random (MAR) or "ignorable," does not assume patients with missing values are similar to those with complete data but instead assumes that observed values can be used to "explain" which values are missing and help predict what the missing values would be.[3] This mechanism of missingness is a more realistic assumption than MCAR, and MAR is assumed by most of the currently used valid techniques for handling missing data. However, most simple imputation methods still yield biased or falsely precise results when MAR is assumed.

Missing not at random (MNAR) is the most problematic censoring mechanism and occurs when missing values are dependent on unobserved or unknown factors. When MNAR is present, statistical adjustment for missing information is virtually impossible.

Because an investigator usually cannot determine the actual mechanism for missingness, statistical analyses usually proceed assuming the data conform to a MAR mechanism. Collecting information to explain why data are missing (eg, participants' mode of transportation and distance to the clinic) can help predict certain values and make the MAR assumption more plausible.[3,4]

What Are the Limitations of These Methods?

Simple imputation methods (eg, LOCF, complete case analysis, mean value imputation, and random imputation) are considered "naive" because they fail to account for the uncertainty in imputing missing values, do not use information available in observed values, can introduce bias, and artificially increase precision (ie, inappropriately narrow confidence intervals and result in smaller *P* values).[3,4] Each of these limitations can cause spurious results. Better estimates and measures of uncertainty (eg, confidence intervals) can be obtained by using maximum likelihood–based methods, hot deck imputation, and multiple imputation.[3]

The primary limitation of complete case analysis is bias and reduced sample size, resulting in reduced study power.[4] Unless the data are MCAR (an unlikely event), estimates using observed case analysis will be biased and the direction of the bias unpredictable. Last observation carried forward is a commonly performed simple imputation technique. This strategy requires the tenuous assumption that the final outcome (eg, 90-day UACR) does not change from the last observed value. In mean value imputation, all missing values are replaced with the mean of observed values (eg, 90-day UACR). With an increasing proportion of missing data, mean value imputation results in larger numbers of patients with an identical imputed value, creating smaller measures of variance and greater bias, artificially increasing the apparent precision of inaccurate estimates.[4,5] Random number imputation avoids the repetitive use of the same imputed estimate but fails to use observed values to inform the selected estimate.

Why Did the Authors Use This Method in This Particular Study?

In the study by Bakris et al, the primary outcome had missing values requiring the use of missing data methods. Several imputation methods were used so that results obtained by the various approaches could be compared.

How Should This Method's Findings Be Interpreted in This Particular Study?

Because of the inherent limitations of simple imputation methods, the multiply imputed results provide the most valid results in the study by Bakris et al. Provided the underlying assumptions are met and rigorous imputation methods (eg, multiple imputation) are used, study results can be interpreted as if all values had been observed.

Caveats to Consider When Looking at the Results in This Study Based on This Method

The LOCF method for handling missing values (as used in the primary analysis by Bakris et al[1]) has the same fundamental limitations as other simple imputation methods, generating potentially biased results with inappropriately narrow confidence intervals. Because results from the post hoc multiple imputation analysis were reported to be no different from those of the LOCF analysis,[1] the primary results can be considered valid despite the risks of using simple imputation methods. Nonetheless, results from the multiple imputation analysis are more rigorous (despite the post hoc selection of this strategy) because of the advantages of this method over simple imputation methods.[5] Caution is required when using traditionally defined "conservative" methods for handling missing outcomes (eg, LOCF) over more sophisticated missing data methods. While they may be conservative in assigning the outcome of a participant with missing data, they can lead to both false-positive and false-negative results in measured treatment effects. In general, multiple imputation is the best approach for modeling the effects of missing data in studies.

ARTICLE INFORMATION

Author Affiliations: Center for Policy and Research in Emergency Medicine, Department of Emergency Medicine, Oregon Health and Science University, Portland (Newgard); Department of Emergency Medicine, Harbor–UCLA Medical Center, Torrance (Lewis); Los Angeles Biomedical Research Institute, Torrance, California (Lewis); David Geffen School of Medicine, University of California, Los Angeles (Lewis).

Corresponding Author: Craig D. Newgard, MD, MPH, Center for Policy and Research in Emergency Medicine, Department of Emergency Medicine, Oregon Health and Science University, 3181 SW Sam Jackson Park Rd, Mail Code CR-114, Portland, OR 97239-3098 (newgardc@ohsu.edu).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: The authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Bakris GL, Agarwal R, Chan JCN, et al; Mineralocorticoid Receptor Antagonist Tolerability Study–Diabetic Nephropathy (ARTS-DN) Study Group. Effect of finerenone on albuminuria in patients with diabetic nephropathy: a randomized clinical trial. *JAMA*. doi:10.1001/jama.2015.10081.

2. Rubin DB. *Multiple Imputation for Nonresponse in Surveys.* New York, NY: Wiley; 1987.

3. Little RJA, Rubin DB. *Statistical Analysis With Missing Data.* 2nd ed. Princeton, NJ: Wiley; 2002.

4. Haukoos JS, Newgard CD. Advanced statistics: missing data in clinical research, I: an introduction and conceptual framework. *Acad Emerg Med*. 2007;14(7):662-668.

5. Newgard CD, Haukoos JS. Advanced statistics: missing data in clinical research, II: multiple imputation. *Acad Emerg Med*. 2007;14(7):669-678.