Review Article

Korean J Anesthesiol 2013 May 64(5): 402-406
http://dx.doi.org/10.4097/kjae.2013.64.5.402

The prevention and handling of the missing data

Hyun Kang

Department of Anesthesiology and Pain Medicine, Chung-Ang Universtiy College of Medicine, Seoul, Korea

Even in a well-designed and controlled study, missing data occurs in almost all research. Missing data can reduce the statistical power of a study and can produce biased estimates, leading to invalid conclusions. This manuscript reviews the problems and types of missing data, along with the techniques for handling missing data. The mechanisms by which missing data occurs are illustrated, and the methods for handling the missing data are discussed. The paper concludes with recommendations for the handling of missing data. (Korean J Anesthesiol 2013; 64: 402-406)

Key Words: Expectation-Maximization, Imputation, Missing data, Sensitivity analysis.

Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data [1]. Accordingly, some studies have focused on handling the missing data, problems caused by missing data, and the methods to avoid or minimize such in medical research [2,3].

However, until recently, most researchers have drawn conclusions based on the assumption of a complete data set. The general topic of missing data has attracted little attention in the field of anesthesiology.

Missing data present various problems. First, the absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false. Second, the lost data can cause bias in the estimation of parameters. Third, it can reduce the representativeness of the samples. Fourth, it may complicate the analysis of the study. Each of these distortions may threaten the validity of the trials and can lead to invalid conclusions.

Types of Missing Data

Rubin first described and divided the types of missing data according to the assumptions based on the reasons for the missing data [4]. In general, there are three types of missing data according to the mechanisms of missingness.

Missing completely at random

Missing completely at random (MCAR) is defined as when the probability that the data are missing is not related to either the specific value which is supposed to be obtained or the set of observed responses. MCAR is an ideal but unreasonable assumption for many studies performed in the field of anesthesiology. However, if data are missing by design, because of an

Received: February 13, 2013. Accepted: February 20, 2013.
Corresponding author: Hyun Kang, M.D., Ph.D., Department of Anesthesiology and Pain Medicine, Chung-Ang Universtiy College of Medicine, 224-1, Heuksuk-dong, Dongjak-gu, Seoul 156-756, Korea. Tel: 82-2-6299-2571, Fax: 82-2-6299-2585, E-mail: roman00@naver.com

equipment failure or because the samples are lost in transit or technically unsatisfactory, such data are regarded as being MCAR.

The statistical advantage of data that are MCAR is that the analysis remains unbiased. Power may be lost in the design, but the estimated parameters are not biased by the absence of the data.

Missing at random

Missing at random (MAR) is a more realistic assumption for the studies performed in the anesthetic field. Data are regarded to be MAR when the probability that the responses are missing depends on the set of observed responses, but is not related to the specific missing values which is expected to be obtained.

As we tend to consider randomness as not producing bias, we may think that MAR does not present a problem. However, MAR does not mean that the missing data can be ignored. If a dropout variable is MAR, we may expect that the probability of a dropout of the variable in each case is conditionally independent of the variable, which is obtained currently and expected to be obtained in the future, given the history of the obtained variable prior to that case.

Missing not at random

If the characters of the data do not meet those of MCAR or MAR, then they fall into the category of missing not at random (MNAR).

The cases of MNAR data are problematic. The only way to obtain an unbiased estimate of the parameters in such a case is to model the missing data. The model may then be incorporated into a more complex one for estimating the missing values.

Techniques for Handling the Missing Data

The best possible method of handling the missing data is to prevent the problem by well-planning the study and collecting the data carefully [5,6]. The following are suggested to minimize the amount of missing data in the clinical research [7].

First, the study design should limit the collection of data to those who are participating in the study. This can be achieved by minimizing the number of follow-up visits, collecting only the essential information at each visit, and developing the user-friendly case-report forms.

Second, before the beginning of the clinical research, a detailed documentation of the study should be developed in the form of the manual of operations, which includes the methods to screen the participants, protocol to train the investigators and participants, methods to communicate between the investigators or between the investigators and participants, implementation of the treatment, and procedure to collect, enter, and edit data.

Third, before the start of the participant enrollment, a training should be conducted to instruct all personnel related to the study on all aspects of the study, such as the participant enrollment, collection and entry of data, and implementation of the treatment or intervention [8].

Fourth, if a small pilot study is performed before the start of the main trial, it may help to identify the unexpected problems which are likely to occur during the study, thus reducing the amount of missing data.

Fifth, the study management team should set a priori targets for the unacceptable level of missing data. With these targets in mind, the data collection at each site should be monitored and reported in as close to real-time as possible during the course of the study.

Sixth, study investigators should identify and aggressively, though not coercively, engage the participants who are at the greatest risk of being lost during follow-up.

Finally, if a patient decides to withdraw from the follow-up, the reasons for the withdrawal should be recorded for the subsequent analysis in the interpretation of the results.

It is not uncommon to have a considerable amount of missing data in a study. One technique of handling the missing data is to use the data analysis methods which are robust to the problems caused by the missing data. An analysis method is considered robust to the missing data when there is confidence that mild to moderate violations of the assumptions will produce little to no bias or distortion in the conclusions drawn on the population. However, it is not always possible to use such techniques. Therefore, a number of alternative ways of handling the missing data has been developed.

Listwise or case deletion

By far the most common approach to the missing data is to simply omit those cases with the missing data and analyze the remaining data. This approach is known as the complete case (or available case) analysis or listwise deletion.

Listwise deletion is the most frequently used method in handling missing data, and thus has become the default option for analysis in most statistical software packages. Some researchers insist that it may introduce bias in the estimation of the parameters. However, if the assumption of MCAR is satisfied, a listwise deletion is known to produce unbiased estimates and conservative results. When the data do not fulfill the assumption of MCAR, listwise deletion may cause bias in the estimates of the parameters [9].

If there is a large enough sample, where power is not an

issue, and the assumption of MCAR is satisfied, the listwise deletion may be a reasonable strategy. However, when there is not a large sample, or the assumption of MCAR is not satisfied, the listwise deletion is not the optimal strategy.

Pairwise deletion

Pairwise deletion eliminates information only when the particular data-point needed to test a particular assumption is missing. If there is missing data elsewhere in the data set, the existing values are used in the statistical testing. Since a pairwise deletion uses all information observed, it preserves more information than the listwise deletion, which may delete the case with any missing data. This approach presents the following problems: 1) the parameters of the model will stand on different sets of data with different statistics, such as the sample size and standard errors; and 2) it can produce an intercorrelation matrix that is not positive definite, which is likely to prevent further analysis [10].

Pairwise deletion is known to be less biased for the MCAR or MAR data, and the appropriate mechanisms are included as covariates. However, if there are many missing observations, the analysis will be deficient.

Mean substitution

In a mean substitution, the mean value of a variable is used in place of the missing data value for that same variable. This allows the researchers to utilize the collected data in an incomplete dataset. The theoretical background of the mean substitution is that the mean is a reasonable estimate for a randomly selected observation from a normal distribution. However, with missing values that are not strictly random, especially in the presence of a great inequality in the number of missing values for the different variables, the mean substitution method may lead to inconsistent bias. Furthermore, this approach adds no new information but only increases the sample size and leads to an underestimate of the errors [11]. Thus, mean substitution is not generally accepted.

Regression imputation

Imputation is the process of replacing the missing data with estimated values. Instead of deleting any case that has any missing value, this approach preserves all cases by replacing the missing data with a probable value estimated by other available information. After all missing values have been replaced by this approach, the data set is analyzed using the standard techniques for a complete data.

In regression imputation, the existing variables are used to make a prediction, and then the predicted value is substituted as if an actual obtained value. This approach has a number of advantages, because the imputation retains a great deal of data over the listwise or pairwise deletion and avoids significantly altering the standard deviation or the shape of the distribution. However, as in a mean substitution, while a regression imputation substitutes a value that is predicted from other variables, no novel information is added, while the sample size has been increased and the standard error is reduced.

Last observation carried forward

In the field of anesthesiology research, many studies are performed with the longitudinal or time-series approach, in which the subjects are repeatedly measured over a series of time-points. One of the most widely used imputation methods in such a case is the last observation carried forward (LOCF). This method replaces every missing value with the last observed value from the same subject. Whenever a value is missing, it is replaced with the last observed value [12].

This method is advantageous as it is easy to understand and communicate between the statisticians and clinicians or between a sponsor and the researcher.

Although simple, this method strongly assumes that the value of the outcome remains unchanged by the missing data, which seems unlikely in many settings (especially in the anesthetic trials). It produces a biased estimate of the treatment effect and underestimates the variability of the estimated result. Accordingly, the National Academy of Sciences has recommended against the uncritical use of the simple imputation, including LOCF and the baseline observation carried forward, stating that:

Single imputation methods like last observation carried forward and baseline observation carried forward should not be used as the primary approach to the treatment of missing data unless the assumptions that underlie them are scientifically justified [13].

Maximum likelihood

There are a number of strategies using the maximum likelihood method to handle the missing data. In these, the assumption that the observed data are a sample drawn from a multivariate normal distribution is relatively easy to understand. After the parameters are estimated using the available data, the missing data are estimated based on the parameters which have just been estimated.

When there are missing but relatively complete data, the statistics explaining the relationships among the variables may be computed using the maximum likelihood method. That is,

the missing data may be estimated by using the conditional distribution of the other variables.

Expectation-Maximization

Expectation-Maximization (EM) is a type of the maximum likelihood method that can be used to create a new data set, in which all missing values are imputed with values estimated by the maximum likelihood methods [14]. This approach begins with the expectation step, during which the parameters (e.g., variances, covariances, and means) are estimated, perhaps using the listwise deletion. Those estimates are then used to create a regression equation to predict the missing data. The maximization step uses those equations to fill in the missing data. The expectation step is then repeated with the new parameters, where the new regression equations are determined to "fill in" the missing data. The expectation and maximization steps are repeated until the system stabilizes, when the covariance matrix for the subsequent iteration is virtually the same as that for the preceding iteration.

An important characteristic of the expectation-maximization imputation is that when the new data set with no missing values is generated, a random disturbance term for each imputed value is incorporated in order to reflect the uncertainty associated with the imputation. However, the expectation-maximization imputation has some disadvantages. This approach can take a long time to converge, especially when there is a large fraction of missing data, and it is too complex to be acceptable by some exceptional statisticians. This approach can lead to the biased parameter estimates and can underestimate the standard error.

For the expectation-maximization imputation method, a predicted value based on the variables that are available for each case is substituted for the missing data. Because a single imputation omits the possible differences among the multiple imputations, a single imputation will tend to underestimate the standard errors and thus overestimate the level of precision. Thus, a single imputation gives the researcher more apparent power than the data in reality.

Multiple imputation

Multiple imputation is another useful strategy for handling the missing data. In a multiple imputation, instead of substituting a single value for each missing data, the missing values are replaced with a set of plausible values which contain the natural variability and uncertainty of the right values.

This approach begin with a prediction of the missing data using the existing data from other variables [15]. The missing values are then replaced with the predicted values, and a full data set called the imputed data set is created. This process iterates the repeatability and makes multiple imputed data sets (hence the term "multiple imputation"). Each multiple imputed data set produced is then analyzed using the standard statistical analysis procedures for complete data, and gives multiple analysis results. Subsequently, by combining these analysis results, a single overall analysis result is produced.

The benefit of the multiple imputation is that in addition to restoring the natural variability of the missing values, it incorporates the uncertainty due to the missing data, which results in a valid statistical inference. Restoring the natural variability of the missing data can be achieved by replacing the missing data with the imputed values which are predicted using the variables correlated with the missing data. Incorporating uncertainty is made by producing different versions of the missing data and observing the variability between the imputed data sets.

Multiple imputation has been shown to produce valid statistical inference that reflects the uncertainty associated with the estimation of the missing data. Furthermore, multiple imputation turns out to be robust to the violation of the normality assumptions and produces appropriate results even in the presence of a small sample size or a high number of missing data.

With the development of novel statistical software, although the statistical principles of multiple imputation may be difficult to understand, the approach may be utilized easily.

Sensitivity analysis

Sensitivity analysis is defined as the study which defines how the uncertainty in the output of a model can be allocated to the different sources of uncertainty in its inputs.

When analyzing the missing data, additional assumptions on the reasons for the missing data are made, and these assumptions are often applicable to the primary analysis. However, the assumptions cannot be definitively validated for the correctness. Therefore, the National Research Council has proposed that the sensitivity analysis be conducted to evaluate the robustness of the results to the deviations from the MAR assumption [13].

Recommendations

Missing data reduces the power of a trial. Some amount of missing data is expected, and the target sample size is increased to allow for it. However, such cannot eliminate the potential bias. More attention should be paid to the missing data in the design and performance of the studies and in the analysis of the resulting data.

The best solution to the missing data is to maximize the data collection when the study protocol is designed and the data collected. Application of the sophisticated statistical analysis

techniques should only be performed after the maximal efforts have been employed to reduce missing data in the design and prevention techniques.

A statistically valid analysis which has appropriate mechanisms and assumptions for the missing data should be conducted. Single imputation and LOCF are not optimal approaches for the final analysis, as they can cause bias and lead to invalid conclusions. All variables which present the potential mechanisms to explain the missing data must be included, even when these variables are not included in the analysis [16]. Researchers should seek to understand the reasons for the missing data. Distinguishing what should and should not be imputed is usually not possible using a single code for every type of the missing value [17]. It is difficult to know whether the multiple imputation or full maximum likelihood estimation is best, but both are superior to the traditional approaches. Both techniques are best used with large samples. In general, multiple imputation is a good approach when analyzing data sets with missing data.

References

1. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol 2009; 60: 549-76.
2. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med 2012; 367: 1355-60.
3. O'Neill RT, Temple R. The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it. Clin Pharmacol Ther 2012; 91: 550-4.
4. Rubin DB. Inference and missind data. Biometrika 1976; 63: 581-92.
5. DeSarbo S, Green PE, Carroll JD. An alternating least-squares procedure for estimating missing preference data in product-concept testing. Decision Sciences 1986; 17 : 163-85.
6. Wisniewski SR, Leon AC, Otto MW, Trivedi MH. Prevention of missing data in clinical research studies. Biol Psychiatry 2006; 59: 997-1000.
7. Scharfstein DO, Hogan J, Herman A. On the prevention and analysis of missing data in randomized clinical trials: the state of the art. J Bone Joint Surg Am 2012; 94 Suppl 1: 80-4.
8. Wilcox S, Shumaker SA, Bowen DJ, Naughton MJ, Rosal MC, Ludlam SE, et al. Promoting adherence and retention to clinical trials in special populations: a women's health initiative workshop. Control Clin Trials 2001; 22: 279-89.
9. Donner A. The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. Am Stat 1982; 36: 378-81.
10. Kim JO, Curry J. The treatment of missing data in multivariate analysis. Sociol Methods Res 1977; 6: 215-41.
11. Malhotra N. Analyzing marketing research data with incomplete information on the dependent variable. J Mark Res 1987; 24: 74-84.
12. Hamer RM, Simpson PM. Last observation carried forward versus mixed models in the analysis of psychiatric clinical trials. Am J Psychiatry 2009; 166: 639-41.
13. Panel on Missing Data in Clinical Trials. The prevention and treatment of missing data in clinical trials. 2nd ed. Washington DC, National Academies Press. 2010, pp 107-14.
14. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. JRSSB 1997; 39: 1-38.
15. Sinharay S, Stern HS, Russell D. The use of multiple imputation for the analysis of missing data. Psychol Methods 2001; 6: 317-29.
16. Rubin DB. Multiple imputation after 18+ years (with discussion). J Am Stat Assoc 1996; 91: 473-89.
17. Acock AC. Working with missing values. J Marriage Fam 2005; 67: 1012-28.