Review | Vol. 4, No. 1 / Oct-Dec 2018 /p. 01-05/ PPCR Journal

Principles and Practice of Clinical Research
A Global Journal in Clinical Research

PPCR
ISSN: 2378-1890

Missing data in clinical trials: a different point of view

W. Fandino[*1]

[1]*The Walton center NHS Foundation Trust, Lower Lane, Liverpool, United Kingdom.*
**Corresponding author: Wilson Fandino, The Walton Center NHS Foundation Trust, Lower Lane, L9 7LJ, Liverpool, United Kingdom. E-mail: wilson.fandino@hotmail.com.*

Received July 1, 2017; accepted October 16, 2017; published July 25, 2018.

Abstract

Despite the high frequency of missing data observed in most clinical trials, the problem continues to be overlooked. While single imputation techniques consistently underestimate the variance, multiple imputation approaches yield more accurate estimators. Even so, the unverifiable assumption of missing at random renders these strategies to be unreliable in many instances. In this short review, a clinical perspective is proposed to revise the main concepts related to missing data in the context of clinical trials.

Keywords: *Clinical trials, missing data, intention to treat analysis*

DOI: http://dx.doi.org/10.21801/ppcrj.2018.41.1

INTRODUCTION

Missing data are a major problem of significant concern in clinical research. According to the intent-to-treat principle, all randomized patients need to be included in the analysis of data, and outcomes should be assessed in the group they were originally allocated to (Dziura, 2013). However, when this principle is violated, the benefits of randomization are lost, rendering the sample to have important imbalances of variables, thus diminishing the statistical power, increasing the type II error and finally biasing the results (O'Neill, 2012). Despite significant advances in the development of statistical software packages, the problem continues to be overlooked (Jaukoos 2007, Bell 2014). Therefore, in this mini-review, the frequency of missing data, main mechanisms, prevention strategies, alternatives to handling the problem and most commonly used techniques to minimize potential biases are presented in a clinical fashion.

Epidemiology

Most clinical trials are expected to have missing data, regardless of how carefully they have been designed. In a descriptive study that evaluated handling of missing data in clinical trials published in major medical journals, as much as 95% had missing outcomes, and only 33% had reportedly used strategies to avoid high attrition rates (Bell, 2014). The problem becomes meaningful when the proportion is unexpectedly high (a subjective threshold has been set at 20%), and it was not planned in the study protocol how to deal with dropouts (Dziura, 2013).

Pathophysiology

There are three different mechanisms based on the relationship of missing data to the outcome and the independent variables, and also with the pattern of censoring: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The former two mechanisms are also known as ignorable, while the latter mechanism is also called non-ignorable (Dziura, 2013). Although the actual mechanism is unverifiable, documenting the reasons of dropping out is usually helpful to formulate hypotheses (Little, 2012).

1. Missing completely at random

This mechanism is unrelated to the study variables (i.e., outcome or independent variables), and therefore, it is deemed to be a random sample of the study population. The most illustrative example is the measurement of blood pressure in a given population sampling. One possible explanation for missing blood pressure values is the breakdown of the sphygmomanometer. If that is the case, the observed and unobserved (missing) values are

theoretically comparable, and consequently, their distributions are similar. This assumption is, however, unrealistic, because it is unlikely that missing data are not related to any of the variables. In most cases, such an optimistic approach should be avoided, because it may lead to biased estimates (Little, 2012).

2. Missing at random

Data are considered MAR when the missing pattern is related to the observed values of the independent variables, but unrelated to both the unobserved values and the outcome (Dziura, 2013). Accordingly, missing data can be fully explained by the observed values of one of several independent variables. Following the example cited above, if a researcher aims to record mean blood pressure readings from clinical notes, it is possible that young people have missing measurements in their notes, because they are presumably healthy and subsequently, they are not expected to have abnormal registrations. In this example, missing data of blood pressure are probably related to age (independent variable), which has been recorded in the dataset.

3. Missing not at random

When missing data are related to the outcome, and possibly to unobserved values of independent variables, the underlying mechanism is deemed MNAR. For the same hypothetical study, if the researcher is not interested in registering the age of the subjects, the mechanism of missing data is MNAR, provided that unobserved blood pressure values are more likely to happen in young people, and therefore, the distribution of age between observed and unobserved blood pressure values would be different. It is impossible to estimate how different those distributions would be –age has not been recorded-, and consequently, there will remain systematic differences between missing and obtained data, even after controlling for the observed values of the independent variables. In other words, the differences cannot be entirely explained by the observed data. This is the worst-case scenario, because an ideal technique to deal with these missing data does not exist, and even sophisticated methodologies (e.g., multiple imputation and maximum likelihood) are unable to completely eliminate the bias (Haukoos, 2007).

In clinical practice, it is challenging to distinguish among these mechanisms. Although typically the MAR mechanism is assumed for most analyses, the validity of the results strongly depends on the actual pattern of censoring (Haukoos, 2007).

Diagnosis

It is useful to visually examine the dataset, looking for specific patterns, time to dropout (e.g., using Kaplan-Meier estimators), factors likely to be associated (for instance, using logistic regression models) and proportions of missing data, thus suggesting an association with observed or unobserved values (Dziura, 2013). In clinical research, dataset obtained from biological variables quite often look randomly distributed. This pattern is called non-monotone, as it cannot be organized in any sequential order. Alternatively, a monotone pattern can also be observed, particularly in longitudinal studies, in which case missing data have a regular pattern of distribution along the dataset, mainly due to the progressive dropouts

- Consider shorter studies with earlier study endpoints.
- Run a pilot study to test the study protocol and identify potential problems with missing data.
- Increase sample size according to the rate of dropouts in similar studies.
- Conduct a withdrawal design (only responders to the treatment are randomized after a run-in period).
- Limit participants burden (e.g., minimizing face to face visits in the follow-up) and provide monetary and non-monetary incentives to continue in the trial.
- Avoid outcome measures with higher probability of missing data.
- Enroll investigators adequately qualified to collect data, and train them to maximize adherence to the treatment.
- Allow rescue medications, when possible.
- Motivate investigators to enhance the quality and completeness of data.
- Keep contact information of dropouts and follow them up, when possible (participants may agree to stop the treatment but remain in the study).
- Determine the reason for withdrawal, when possible.

Table 1. Proposed key strategies to prevent missing data during the design and conduct of clinical trials. Adapted from Little et al. (2012), and Dizura et al. (2013).

(Haukoos, 2007). The distinction between these patterns is important, because certain methodologies to deal with missing data may be more suitable than others.

Prevention

Table 1 provides key strategies to prevent high rates of missing data. Although increasing the sample size according to the expected dropouts is recommended, it does not account for potential imbalances between groups, and the probability of bias will depend on the underlying mechanisms (Little, 2012). On the other hand, increased awareness of the problem by researchers, editors of medical journals and regulatory authorities, including FDA (Food and Drug Administration) and CONSORT (Consolidated Standards of Reporting Trials) guidelines, may have an impact in the manner this topic is addressed (O'Neill, 2012). While researchers need to be familiar with appropriate statistical methods used to deal with missing data, regulatory authorities have the responsibility of making a cultural shift from the methodologies to deal with missing data towards strategies to prevent them in a well-conducted clinical trial.

Treatment

Conservative treatment: complete-case analysis

With this approach, only the observed values are included in the analysis. Consequently, the sample size and the precision of the estimators are invariably reduced (Dziura, 2013). This practice has been strongly discouraged in modern clinical research -unless the dropout rate is very low-, since the intrinsic assumption of MCAR is unrealistic -most missing data will be related to the outcome, the independent variables, or both of them- (Little 2012, Dziura 2013). However, this methodology continues to be overused in the report of primary analyses (Bell, 2014).

There are two subsets of methodologies derived from complete-case analysis, namely, available-case analysis and weighted complete-case analysis. The former accounts for data available for statistical analysis, thus varying the number of cases analyzed with each statistical test. The latter is based on a weight assigned to each case depending on whether or not they are complete, thus increasing the variance and diminishing the precision of the estimators. Although these methods are more recommended than complete-case analysis, the magnitude of bias is still unpredictable and highly dependent on the underlying mechanism (Haukoos, 2007).

Interventional treatment

Missing data can be replaced according to the observed values and analyzed following the intent-to-treat principle. Options include single imputation, multiple imputation and maximum likelihood estimation.

1. Single imputation

Although this technique has long been used, the results are variable, since they rely on the MCAR assumption and, importantly, variance is artificially decreased (Haukoos, 2007). Some of the preferred techniques are mean or median imputation (depending on the data distribution), hot and cold deck imputation, regression imputation, last observation carried forward (LOCF), baseline observation carried forward (BOCF) and worst-case analysis.

Mean and median imputation techniques replace the censored data with central tendency measures obtained for each variable within each group. However, the information provided from other variables is ignored. In addition, missing data are replaced based on the observed values, thus increasing the risk of biased parameters. The inherent variance decrease leads to false improvement of the statistical precision, thereby increasing the risk of type I error. On the other hand, hot and cold deck imputations use a "matched" case from the same or an external dataset, respectively, containing similar values to replace censored data (Haukoos, 2007).

Regression imputation provides the predicted data for each patient obtained from a linear regression model, in which the variable containing censored values is included as the outcome and the remainder variables are independent. Thus, for each subject the predicted value will be different and dependent on the individual information from other variables. The model requires parametric assumptions. Also, since no additional variance is added, it tends to be underestimated. The problem can be minimized adding residual errors to the predicted values with stochastic techniques (Haukoos, 2007). Inverse probability weighting is a modality of regression imputation, in which variables more likely to be unobserved have more weight in the analysis (Diziura, 2013).

With LOCF technique, missing data are replaced with the last obtained value for each subject. Thus, the values are assumed to be unchanged. Although it is widely used in longitudinal studies, easy to perform and accepted by most peer-reviewers, the risk of type I and type II error is increased (depending on whether the outcome improves or gets worse in the unobserved subjects and the imbalances of dropouts between treatment and control groups), particularly in samples with earlier or high rate of dropouts (O'Neill, 2012). Similarly, BOCF uses baseline values to replace data,

thereby artificially diminishing the variance and increasing the probability of type I error. Lastly, the worst-case analysis is commonly used in logistic regression models, in which the binary outcome for the missing data is assumed to have the less favourable values. However, this assumption may lead to type II error (Haukoos, 2007).

2. Multiple imputation

The uncertainty of missing values has led to the development of multiple imputation techniques. They are based on the generation of several datasets, which are analyzed to provide estimation of parameters with standard deviation and confidence intervals, and eventually merged to provide a more plausible estimation of the censored data (Newgard 2007, Sterne 2009).

Technically, the process uses Bayesian methods (e.g., Monte Carlo simulations), taking the observed data as the prior distribution and the complete dataset -which includes the estimation of missing values generated with a likelihood function- as the posterior distribution (Newgard, 2007). After an iterative process, multiple datasets -usually 5 to 10, depending on the rate of missing data- are generated from the posterior distribution, thus creating several estimations for a given missing value that are followed by statistical analyses of variance (ANOVA) of between and within datasets, thus fitting the model to each dataset. However, since datasets are randomly created, the estimations are not exactly the same each time the computational algorithm is repeated (Sterne 2009, Dziura 2013). The reliability of the technique depends on the MAR assumption -which cannot be tested-, the normal distribution of the data and the variables included in the model (Haukoos 2007, Sterne 2009).

It is possible to estimate the relative efficiency of the process by dividing the number of datasets over the sum of datasets plus the rate of missing data. For instance, with 10 datasets and 40% of missing data, multiple imputation yields 96% of relative efficiency [10/(10 + 0.4) x 100]. However, even with 10 imputations, the between-groups (datasets) variance is expected to be large, and consequently the estimated standard error may have a low precision (Newgard 2007, Sterne 2009).

3. Maximum likelihood estimation

This model includes a likelihood function that rather than replacing data, yields unique estimators with more accurate standard errors. Provided that the number of datasets used is appropriate, the relative efficiency of the technique is comparable with that from multiple imputation (Newgard, 2007). Nevertheless, the method is reliant on parametric assumptions and is only suitable for replacing outcome values (Dziura, 2013).

Follow-up

After replacing the missing values, the robustness and reliability of the technique needs to be examined with descriptive analyses of the imputed values. If inconsistencies are suspected, the whole process should be revised to find potential explanations of unstable values. A sensitivity analysis is also useful to evaluate how plausible the imputation was. This technique compares different simulations (e.g., intention-to-treat versus per-protocol analyses) to evaluate the stability of the estimators (i.e., p values and confidence intervals), before considering it has been successful (Newgard, 2007). However, it is noteworthy that sensitivity analyses assume a MAR mechanism, and since this assumption is unverifiable, the reliability of the results is not guaranteed.

Prognosis

It depends on the mechanism of missing data.

1. MCAR: this condition has a benign prognosis, since the results are not considered biased. Unfortunately, this scenario is not common and most missing data will be classified into MAR or MNAR.
2. MAR: Data can still be handled with single imputation, although multiple imputation and maximum likelihood techniques are recommended.
3. MNAR: There is no definitive treatment for this dataset, and the results will be biased despite any treatment. A "palliative" approach with multiple imputation techniques and maximum likelihood estimation may be helpful, but the information lost will be important, particularly in high rate of dropouts (Haukoos, 2007).

Sequels

Frustration and disappointment of the researcher, rejection of the manuscript and the waste of time and resources represent the aftermath when missing data are not suitable to be replaced by imputation techniques. However, the most important consequence of inappropriate handling of missing data is obtaining biased estimates of the results, which can mislead the conclusions and influence the clinical practice.

CONCLUSION

Considering that in modern clinical research advanced statistical software packages are readily available, sophisticated methods including regression imputation, multiple imputation and maximum likelihood are

becoming the preferred techniques to deal with missing data. However, a definitive solution does not exist, and prevention is by far the best treatment, since the uncertainty of missing outcomes is always difficult to address. Regardless of which strategy was planned to deal with missing data, it is important to be determined a priori for the transparency of data and avoid the temptation of modifying the clinical trial protocol for the benefit of the results.

Conflict of interest and financial disclosure

The authors followed the International Committee or Journal of Medical Journals Editors (ICMJE) form for disclosure of potential conflicts of interest. All listed authors concur with the submission of the manuscript, the final version has been approved by all authors. The authors have no financial or personal conflicts of interest.

REFERENCES

Bell, M. L., Fiero, M., Horton, N. J., & Hsu, C. H. (2014). Handling missing data in RCTs; a review of the top medical journals. BMC medical research methodology, 14(1), 118.

Dziura, J. D., Post, L. A., Zhao, Q., Fu, Z., & Peduzzi, P. (2013). Strategies for dealing with missing data in clinical trials: from design to analysis. Yale J Biol Med, 86(3), 343-358.

Haukoos, J. S., & Newgard, C. D. (2007). Advanced statistics: missing data in clinical research-part 1: an introduction and conceptual framework. Academic Emergency Medicine, 14(7), 662-668.

Little, R. J., D'agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T. et al. (2012). The prevention and treatment of missing data in clinical trials. New England Journal of Medicine, 367(14), 1355-1360.

Newgard, C. D., & Haukoos, J. S. (2007). Advanced statistics: missing data in clinical research-part 2: multiple imputation. Academic Emergency Medicine, 14(7), 669-678.

O'Neill, R. T., & Temple, R. (2012). The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it. Clinical Pharmacology & Therapeutics, 91(3), 550-554.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G. et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ, 338, b2393.