HHS Public Access

Author manuscript

Stroke. Author manuscript; available in PMC 2016 June 01.

Published in final edited form as:

Stroke. 2015 June ; 46(6): e130–e132. doi:10.1161/STROKEAHA.115.007984.

What is Missing from my Missing Data Plan?

Sharon D. Yeatts, PhD and **Renee' H. Martin, PhD**

Department of Public Health Sciences, Medical University of South Carolina, Charleston SC

Keywords

missing data; imputation

Under the Intention-to-Treat (ITT) principle, all randomized subjects should be analyzed according to their randomly assigned treatment, regardless of treatment actually received or protocol compliance. Adherence to this principle requires that even subjects with missing outcome data be included in the analysis; in fact, the exclusion of such subjects can have important implications regarding power and bias. Statistical methods for dealing with missing data exist, but many questions remain unclear. Much statistical research has been devoted to the development and assessment of various methods for handling missing data[1]. The choice of appropriate methodology requires assumptions regarding the mechanism underlying the missing data. All of these decisions should be made *a priori*, preferably before the trial starts but certainly before unblinding the trial. Related conversations between clinical investigators and the study statistician during the design phase often focus on more practical questions. Is there some threshold for the missing data rate below which the trial's conclusions are unlikely to be affected? Under what circumstances can the missing data be excluded from the analysis without biasing estimation, or is imputation always the preferred approach?

In this manuscript, we discuss implications of missing outcome data from a practical standpoint. We describe potential reasons for missing data and suggest strategies to minimize its occurrence. We also present common imputation approaches and emphasize that, since none of these approaches are universally preferred, the best analytic plan includes a series of sensitivity analyses.

Why does missing data occur?

In any longitudinal trial where subjects are followed over some "extensive" period of time, lengthy follow-up makes missing data somewhat unavoidable. In stroke clinical trials, the primary outcome assessment often occurs at 90 days, although there is evidence to suggest that additional follow-up may be beneficial. Subjects may expire, or withdraw informed consent, prior to primary outcome ascertainment. Subjects may become "lost" to the study team because of incomplete contact information, or because they move out of the relevant

Corresponding Author: Sharon D. Yeatts, Department of Public Health Sciences, Medical University of South Carolina; 135 Cannon St, Ste 305, Charleston SC 29425; Phone 843-513-9085; yeatts@musc.edu.

Disclosures: NONE

catchment area. When developing an approach for handling missing data, the best defense is a good offense; that is, the best approach is to proactively prevent the occurrence of missing data.

Various protocol strategies can be considered, based on careful consideration as to why missing data might occur in a population. The first such strategy is to recognize the distinction between discontinuation from study treatment and discontinuation from the study; subjects may discontinue study treatment for a variety of reasons, but such subjects remain part of the study, and follow-up attempts should be made until or unless consent has been withdrawn.[2] This distinction is unlikely to be an issue in acute trials where treatment is completed relatively early compared to the total duration of follow-up, but is likely to be extremely important in prevention studies involving adherence to a treatment regimen for the duration of the follow-up period.

Another strategy involves detailed review of the protocol's requirements, with careful consideration of those elements which might impact a subject's ability to complete the protocol. If travel to/from the clinic is likely to be difficult because of age or underlying disability, primary outcome ascertainment could be dramatically impacted. In such cases, the investigator might outline specific efforts to overcome this obstacle, including assistance in the scheduling of transportation or reimbursement for associated expenses, the option to conduct visits via telemedicine or to send an investigator to the subject's residence (home, nursing home, rehab facility, etc). If missing data is instead likely because of the transient nature of the population, frequent contact with the subject, such as periodic telephone calls between clinic visits, may help to avoid such loss to follow-up; the use of private investigators to find such patients has been employed in other disease areas.

The occurrence of missing data may vary with timing and complexity of the outcome determination. Two popular outcome assessments in stroke trials, the National Institutes of Health Stroke Scale (NIHSS) and modified Rankin Scale (mRS), illustrate this point. The NIHSS requires in-person assessment, whereas the mRS can be reliably administered via telephone, and mortality can be established via public record. Therefore, one might expect minimal missing data for a mortality endpoint, with the missing data rate higher for the mRS and higher still for the NIHSS. The relevance of the endpoint to subjects who have died might also be a consideration when selecting an endpoint. Since death is a category of the mRS, the mRS would not be considered missing for subjects who expire prior to the primary outcome assessment. Even for the same endpoint, one could expect the missing data rate to increase with length of follow-up; completion of protocol-specified visits is likely easier to maintain over the course of a 3-month trial than a one-year trial.

How much missing data is acceptable?

The question is often asked: how much missing data can a trial tolerate without jeopardizing the validity of its conclusion? Though such thresholds appear in the literature, there is no consensus as to their utility. Schulz and Grimes[3] reference a rule of thumb indicating that trial validity is threatened when the missing data rate reaches 20% or more, whereas the bias resulting from less than 5% is likely to be trivial. Bennett[4] suggests that traditional methods

may generate biased results when the missing data rate is larger than 10%. Researchers should be most concerned with the impact of missing data in the 5% to 20% range, where missing data is sufficiently common to cause statistical concern but not common enough for the clinical community to reject trial results on this basis alone. Rather than trying to achieve a somewhat arbitrarily tolerable rate, investigators should focus on minimizing the occurrence of missing data.

Recent trials have been able to achieve minimal missing data pertaining to the primary outcome, despite facing some of the obstacles described above. The Interventional Management of Stroke (IMS) III trial was designed to assess the efficacy of a combined approach of endovascular therapy following intravenous (IV) tissue plasminogen activator (tPA) versus IV t-PA alone. Approximately 4% missing data was reported for the primary endpoint of favorable outcome, defined by mRS of 0–2 at 90 days; this percentage included both subjects for whom the primary outcome was not collected and subjects for whom the outcome was assessed outside of the specified window.[5]

Do I need to impute (and what does that mean)?

In order to conduct the analysis according to the ITT principle, the missing outcome data must be accounted for so that all randomized subjects can be analyzed. The default approach of most statistical packages can be thought of as complete case analysis, meaning that only subjects with complete data (non-missing values for all variables incorporated in the analysis) are analyzed. This approach essentially ignores the missing observations and, as such, is not in concordance with the ITT philosophy. Statistically, this approach reduces the available sample size, and hence the anticipated power, and is likely to result in biased estimation.

Imputation has been described as “the practice of ‘filling in’ missing data with plausible values”[6]. Imputation assumes that the observed data can be used to generate a response, or a distribution of potential responses, based on pertinent subject characteristics. Single imputation methods replace the missing outcome with a single value; in multiple imputation (MI), the process of imputing a response is repeated in order to incorporate uncertainty in the outcome assignment. Both approaches (single or multiple) result in a likeness of the original data set which contains no missing data. A sampling of such approaches are described below; however, since none of these is universally recommended for all missing data scenarios, a sensitivity analysis of various approaches is recommended to support trial findings.

Last Observation Carried Forward (LOCF) is a commonly encountered form of single imputation, potentially useful in longitudinal trials, wherein the missing outcome is replaced with the last observed outcome assessment. For example, in IMS III, the mRS was measured at 1, 3, 6, 9, and 12 months. For a subject who withdrew consent after the 6 month mRS assessment, LOCF would replace both the missing 9 and 12 month outcomes with the available 6 month outcome. This approach assumes that the subject’s outcome would have remained constant, neither improving nor declining, beyond the last available assessment date. For degenerative diseases/disorders such as Parkinson’s disease or Amyotrophic

Lateral Sclerosis, such an assumption may be immediately declared invalid based on knowledge of the disease course. But even for non-degenerative conditions, the validity of the assumption may depend on both timing of the last available assessment in relation to the missing outcome, as well as the response observed at the last available assessment. A subject who is deceased at 1 month will remain deceased at 3 months, and a subject who has no remaining symptoms at 1 month may be unlikely to decline to disability by 3 months. But is it reasonable to assume that a subject with moderate disability, represented by a mRS score of 3 at 1 month, will remain a 3 at 3 months, or might he continue to improve? How about at 12 months?

Alternative single imputations, such as the worst- and best- case imputations, assign the worst and best possible outcome, respectively. A combination of these can also be used to create the most extreme result possible, by replacing missing data with the best outcome possible in the control arm and the worst outcome possible in the experimental arm. The resulting treatment effect estimate would represent a conservative estimate for efficacy. Single imputation methods, though easy to implement and interpret, ignore the uncertainty associated with the imputation; the resulting standard error and p-value are biased downward, thereby making it easier than it should be to declare a difference.[2]

MI[6] is generally preferred over single imputation, because the inference correctly reflects the uncertainty associated with the procedure. Multiply imputed data sets are generated and analyzed, and the inference from each statistically combined to estimate the treatment effect and its corresponding standard error and p-value. For instance, in an acute treatment trial such as IMS III, favorable outcome at 90 days could be imputed via logistic regression, where favorable outcome is predicted according to baseline characteristics (e.g. stroke severity, age, sex, time to treatment) and treatment assignment, as well as post-treatment characteristics (24-hour NIHSS, discharge location, and 1-month mRS).

Missing Data Mechanisms

The properties of the described approaches may depend on the mechanism underlying the missing data: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Data are considered MCAR if the missing data do not depend on either observed or unobserved subject characteristics. If instead the missing data are related only to observed data, it is considered MAR. In a post-stroke rehabilitation trial, a subject might withdraw consent after early visits suggesting decline, rather than improvement, in physical function; then the missing outcome is associated with the outcome data obtained previously. Conversely, if missingness depends on the value of the unobserved outcome, then it is considered MNAR. A subject in a post-stroke rehabilitation trial might experience the same decline in physical function but withdraw prior to the first outcome visit; since missingness depends on the unobserved outcome, which cannot be predicted by observed trial data, this would be MNAR. Though mechanism is important in justifying the analysis approach, the assumption cannot be formally tested. Under the situation of MCAR and MAR, MI is the most straightforward approach, and most software that implement MI make this assumption. While MI can be applied in the MNAR setting, this requires proper specification of the MNAR mechanism which can be challenging.[7]

Preparations during Trial Design

From a statistical perspective, the primary concern related to missing data is impact on the inference associated with the treatment effect. Unless missing data are MCAR, complete case analysis results in biased estimates of the treatment effect, as well as power loss owing to the reduction in available sample size. Sample size inflation is required to compensate. An inflation factor equal to the proportion of missing data anticipated is insufficient, unless subjects with missing data are excluded from the analysis; and even in this case, the likely result is a biased estimate of the treatment effect. When an imputation procedure (multiple or otherwise) is specified, sample size inflation should consider both proportion of missing data anticipated and the resulting dilution of the treatment effect estimate in order to maintain target power.[2,8]

Conclusions

The best approach for handling missing data in a clinical trial is to minimize the likelihood of its occurrence by selecting appropriate trial endpoints and making extensive efforts to achieve complete follow-up. Guidance literature does not offer a preferred handling method for all situations, and the previously described imputation approaches are just a sampling of those available. It is important to contemplate the rationale and implications for proposed approaches, to prespecify during the design phase an approach based on relevant clinical and statistical considerations, and avoid data-driven changes to prespecified analysis plans. Since no universally preferred approach exists, sensitivity analysis under varying missing data approaches should be undertaken; consistent conclusions under multiple paradigms will lead to increased confidence in trial conclusions.

Acknowledgments

Funding: Partially supported by National Institutes of Health U01 NS077304.

References

1. Little, RJA.; Rubin, DB. Statistical analysis with missing data. 2. Hoboken, New Jersey: Wiley & Sons; 2002. Introduction; p. 3-23.
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. The New England Journal of Medicine. 2012; 367:1355–1360. [PubMed: 23034025]
3. Schulz KF, Grimes DA. Sample size slippages in randomized trials: exclusions and the lost and wayward. Lancet. 2002; 359:781–785. [PubMed: 11888606]
4. Bennett DA. How can I deal with missing data in my study? Australian and New Zealand Journal of Public Health. 2001; 25:464–469. [PubMed: 11688629]
5. Broderick JP, Palesch YY, Demchuk AM, Yeatts SD, Khatri P, et al. for the Interventional Management of Stroke (IMS) III Investigators. Endovascular Therapy after Intravenous t-PA versus t-PA Alone for Stroke. The New England Journal of Medicine. 2013; 368:893–903. [PubMed: 23390923]
6. Schafer JL. Multiple Imputation: A Primer. Statistical Methods in Medical Research. 1999; 8:3–15. [PubMed: 10347857]
7. Molenberghs, G.; Kenward, MG. Multiple Imputation. In: Senn, S.; Barnett, V., editors. Missing Data in Clinical Trials. West Sussex, England: John Wiley & Sons; 2007. p. 105-117.

8. Friedman, LM.; Furberg, CD.; DeMets, DL. Fundamentals of Clinical Trials. 3. New York, NY: Springer-Verlag; 1998. Sample Size; p. 107-109.