• Biostatistics in psychiatry (9) •

How to avoid missing data and the problems they pose: design considerations

Julia Y. LIN[1*], Ying LU[1,2], Xin TU[3]

One of the most common challenges in biomedical and psychosocial research is missing data, which occurs when respondents refuse to provide answers to sensitive questions and when study subjects are lost to follow-up during the repeated assessments of longitudinal trials. This paper is the first in a 3-part series focusing on this important topic; it describes different types of missing data and their differential effects on model estimates, focusing on study design strategies that can be used to prevent or minimize missing data and, thus, maintain the scientific integrity of the research. The second paper in the series will discuss implementation strategies to manage and reduce missing data while conducting the study, and the third paper will discuss analytic strategies for dealing with missing data after completion of data collection.

It is always worth devoting careful attention to issues like missing data that may have significant effects on model estimates. As the saying goes, 'an ounce of prevention is worth a pound of cure,' so it is much better to focus on these issues during the planning stage of a study rather than having to deal with them later in the study. In this paper, we focus squarely on such preventive strategies as the first line of defense against the ubiquitous problem of missing data in clinical research studies.

1. Types of missing data and their effects on model estimates

The reasons for missing data vary, and the degree to which missing data decreases the validity of the estimates depends on how the missing data arises. Thus it is important to make plausible assumptions about how missing data occurs in a study and, based on these assumptions, select appropriate models for addressing the effect of the missing data on inference from the observed results of the study. There are three statistical models with increasing levels of generalizability that are commonly used to classify different types of missing data.

If missing data occur in a random fashion—that is, with no particular pattern that determines which data are observed and which are missing—it is typically referred to as 'missing completely at random' (MCAR). Data that are MCAR have no influence on any of the study participants' outcomes and, thus, may be ignored because they do not result in an inferential bias.

In many follow-up studies participants may be lost to follow up because of deteriorated or improved health conditions. In this situation the missed data are not MCAR because the probability of the missed visit depends on the outcome. For example, if an investigational medication worsens depression, subjects may drop out from the study over time, creating a so-called 'monotone missing data' pattern. In cases like this where whether or not data is missing is influenced by treatment-related effects, ignoring the missing data and focusing only on the subjects with complete data will usually give rise to biased estimates of treatment effects. If depression measures taken at baseline (or prior to dropping out) can be used to model this dependence structure in the missing data pattern (i.e., the relationship between baseline severity of depression and dropping out) this information can then be incorporated into the model for treatment effects to address the bias. If the dependence structure between the outcome of interest and the missing data can be modeled based on observed data, the missing data are classified as 'missing at random' (MAR).

If, however, it is not possible to model the pattern of missing data (i.e.,the pattern of dropping out is not related to any observed baseline measures) the biases resulting from missing data cannot be estimated and the missing data are classified as 'not missing at random' (NMAR). Fortunately, in most carefully designed studies, missing data follows the MAR mechanism. For more in-

doi: 10.3969/j.issn.1002-0829.2012.03.010

[1]US Department of Veterans Affairs Cooperative Studies Program Coordinating Center, Palo Alto VA Health Care System, Palo Alto, CA, USA
[2]Department of Health Research and Policy, Stanford University, Stanford, CA, USA
[3]Department of Biostatistics and Computational Biology, University of Rochester, NY, USA
*Correspondence:Julia.Lin@va.gov

depth discussions of the types of missing data, readers are referred to Tang & Tu[1] and to Little & Rubin.[2]

To illustrate how the three types of missing data affect model estimates, consider a hypothetical example of an intervention study where we compare depressed patients receiving intervention A versus patients receiving intervention B and assess the outcome based on improvement in their Hamilton Depression Rating Scale (HAMD) scores from baseline to the 3-month follow-up. We would like to know whether intervention A results in a greater reduction in HAMD scores than intervention B.

Study participants who relocate to other cities end up missing their post-intervention follow-up visits, resulting in missing HAMD scores. These missing outcome measures follow the MCAR mechanism because it is assumed that relocating to other cities is not related to the severity of the participant's depression or to any other characteristics related to depression, either at baseline or during the course of the study. In this case, the missing data do not result in biased estimates of change in HAMD scores between the two intervention groups.

If we find that some study participants who are more depressed at baseline are less likely to return for follow-up visits, missing HAMD scores for these subjects are likely to follow the MAR mechanism, since their failure to participate in follow-up is related to the severity of depression at baseline. In this case, if we only compared the change in HAMD scores between the two interventions for participants who returned for the follow-up visit, thus excluding those with missing data, our treatment difference estimate may be biased, especially if the severity of depression at baseline is not balanced between the two groups. However, if we adjust for baseline depression severity (e.g., by using regression adjustments) we can reduce or even eliminate bias in the estimate of the treatment effect.

If study participants' failure to participate in follow-up assessments cannot be explained by the existing data in the study then the missing data are considered NMAR. For example, if subjects with suicide ideation at baseline are more likely to drop out and suicide ideation was not recorded in the study database, then we would not be able to model the relationship of missing data to observed study outcomes. In this situation it is more difficult to adjust for potential biases that arise when limiting the comparison of the two interventions to those subjects who complete the trial.

Even when missing data do not result in significant bias in model estimates, as occurs for data that are MCAR, the missing data may, nevertheless, lead to a reduction in the usable sample size, potentially causing the study to become underpowered to detect the true effect sizes. It is also important to clarify that when there are few missing data (e.g., <5% of data) the bias in model estimates and the decreased power to detect differences introduced by the missing data are much smaller than when there are many missing data.

2. Simpler designs

> *'Good design may not eliminate the problem of missing data, … it can reduce it, so that the modern analytic machinery can be used to extract statistical meaning from study data. Conversely, we note that when insufficient attention is paid to missing data at the design state, it may lead to inferential problems that are impossible to resolve in the statistical analysis phase.'*(Lavori et al. 2008)[3]

Before planning a study, the researcher should always ask, 'What is the objective of the study?' The study should be designed in a way that will serve to answer the research question. For example, if the research objective is to test the hypothesis that decreased social support mediates the association between traumatic life events (e.g., death in the family, divorce, unemployment, etc.) and increased psychological distress, then the study should plan to have at least three assessment periods: to ensure proper temporal order for all variable in the hypothesis, the presence of the traumatic event is assessed at baseline and social support and psychological distress are assessed at the first and second follow-up assessments, respectively.

One way to avoid having missing data is to simplify the study design while collecting sufficient data to address the research questions posed. Here are some of our suggestions.

a) Focus on the research objectives of the study and only collect data that are absolutely necessary to fulfill the objectives. This prevents imposing an unnecessary burden on research staff and study participants, and allows the study to dedicate limited resources to improving the quality of the data collected. If having three assessments is sufficient for fulfilling the study objectives, it would be frivolous to require additional assessments.

b) Target the appropriate patient population. For example, if the study objective is to assess the three-month treatment outcome, it behooves the study to exclude patients who cannot participate for three months.[4] However, be mindful that such exclusion may limit generalizability of study findings.

c) Reduce the complexity of data collection procedures so they are straightforward to carry out. For example, in a study that uses a self-administered sleep diary, a question about sleep onset latency will elicit more usable and valid responses if it clearly defines sleep onset latency as the amount of time it takes to fall asleep or,

even more specifically, the lapse between the time when the respondent went to bed and the time when the respondent fell asleep.

d) Allow and encourage multiple methods of assessment. For example, if study participants are not available to come to the clinic for study visits, allow alternative methods of assessment such as telephone interviews, self-administered surveys or, if appropriate, interviews of surrogates or reviews of medical charts.

e) If it is possible to obtain participant consent, incorporate methods of continuing to follow up study participants even after they withdraw from the study, such as through patient chart review or the use of an electronic medical database.

f) Thoroughly consider all possible factors that could result in missing data and include assessment of these factors so appropriate adjustments can subsequently be made for missing data.

g) Estimate the anticipated amount of missing data and account for it in the sample size calculation to minimize the possibility of having an underpowered study.

h) Specify the analytical strategies for dealing with missing data a priori. This will limit the use of post hoc approaches that could negatively affect the validity of study conclusions.

3. Resource allocation

Ensuring proper allocations of resources to facilitate data collection can effectively reduce missing data and their impact on the scientific integrity of the research. Even the most well thought-out study would be difficult to execute without adequate resources. The first thing that comes to mind is, of course, budgeting sufficient funds to carry out the study, including the salary support for required time and effort. However, there are other types of resource issues that researchers should consider when planning and designing a study.

a) If the study requires participants to travel for study visits, assess the feasibility and appropriateness of travel reimbursements or other incentives to reduce the burden on participants and compensate them for their time and effort. Be mindful of ethical concerns that such compensation does not become a source of coercion for study participation that could compromise the integrity of the study.

b) If the study protocol requires laboratory tests, such as blood work or brain imaging, make sure study participants and/or study personnel have easy access to appropriate laboratories, along with plans for alternative solutions.

c) It is sometimes useful to have an external body of experts to assess whether missing data poses threats to the integrity of inference. If so, it is highly recommended that the study establishes an advisory board or a data monitoring committee (DMC) to carefully monitor missing data during the study so timely actions can be taken to minimize the amount of missing data and collect relevant information to address their impact on study findings during the analysis.

In conclusion, missing data is a common challenge in biomedical and psychosocial research and has the potential to negatively affect research integrity. While it is not always avoidable, especially in longitudinal studies, it can be minimized with careful design considerations.

References

1. Tang W, Tu XM. *Modern Clinical Trial Analysis*. New York: SpringerScience, 2012.
2. Little RJA, Rubin DB. *Statistical Analysis with Missing Data.* 2nd ed. New York: Wiley, 2002.
3. Lavori PW, Brown CH, Duan N, Gibbons RD, Greenhouse J. Missing data in longitudinal clinical trials. Part A: design and conceptual issues. *Psychiatric Annals* 2008; **38**(12): 784-792.
4. Myers WR. Handling missing data in clinical trials: an overview. *Drug Information Journal* 2000; **34**: 525-533.

Julia Lin is a biostatistician at the US Department of Veterans Affairs (VA) Palo Alto Cooperative Studies Program Coordinating Center (CSPCC). Her research interests include design and analysis of clinical trials and causal modeling methods, with particular interest in the area of psychiatry. She has been involved in clinical trials, observational studies and survey research advising on methodological and statistical issues, providing statistical analytical support and documentation of research findings, and managing complex datasets. Her recent research is in the areas of psychiatry, critical care and health services research. Dr. Lin has recently been appointed a Biostatistical Editor for the Shanghai Archives of Psychiatry.

Dr. Lu is Professor of Biostatistics at Stanford University and the Director of the US Department of Veterans Affairs (VA) Palo Alto Cooperative Studies Program Coordinating Center (CSPCC), which has more than 50 staff members who provide comprehensive research support to the VA's nationwide large-scale multicenter clinical trials and DNA bank studies. Originally from Shanghai, Dr. Lu received his BS in Mathematics from Fudan University and his MS in Applied Mathematics from Shanghai Jiao Tong University followed by a Ph.D. in Biostatistics from the University of California at Berkeley. Dr. Lu's work, which has been published in more than 200 peer-reviewed publications, covers a wide range of clinical domains including several trials in mental health that he is currently overseeing at the CSPCC. Dr. Lu is an elected fellow of the American Statistical Association and a recipient of the Evelyn Fix Memorial Award and the Healthstar Osteoporosis Medical Research Award. As an alumnus of Shanghai Jiao Tong University, Dr. Lu is honored to serve as a Biostatistical Editor for the Shanghai Archives of Psychiatry.

Dr. Tu is Professor and Co-Director of the Division of Psychiatric Statistics (DPS) and Director of the Biostatistics Consulting Service Center of the Department of Biostatistics and Computational Biology at the University of Rochester Medical Center. His biostatistical research over the last 20 years has focused on the application of longitudinal data analysis, SEM, counterfactual outcome based causal models, distribution-free models, latent growth mixture models and functional response models to observational and randomized controlled trials across a range of disciplines, particularly in the behavioral and social sciences. Dr. Tu has authored over 150 peer-review papers on these topics and authored or edited several volumes on applied biostatistics that focus on topics such as the theory and application of U-statistics, instrumentation and agreement analysis, causal inferences and longitudinal data, and quality of life and cost effectiveness analysis. Dr. Tu has recently been appointed Biostatistical Editor for the Shanghai Archives of Psychiatry.