On the Prevention and Analysis of Missing Data in Randomized Clinical Trials: The State of the Art

Daniel O. Scharfstein, ScD, Joseph Hogan, ScD, and Amir Herman, MD, PhD

Abstract: We summarize and elaborate on the recently published National Research Council report entitled “The Prevention and Treatment of Missing Data in Clinical Trials.” We tailor our discussion to orthopaedic trials. In particular, we discuss the intent-to-treat principle, review study design and prevention ideas to minimize missing data, and present state-of-the-art sensitivity analysis methods for analyzing and reporting the results of studies with missing data.

Introduction

Randomized clinical trials are considered the gold standard for comparing medical interventions. Randomization ensures that the treatments being compared are balanced, on average, with respect to measured and, more importantly, unmeasured risk factors that are associated with outcomes. In the prototypical two-arm study, patients are randomized to an active or a control treatment and are followed for a specified amount of time, at which point an outcome will be measured. The intention-to-treat (ITT) effect is the effect of being assigned to the active treatment versus being assigned to the control group. In the absence of missing outcome data, the ITT effect is estimated by contrasting the mean outcomes of those assigned to the two treatment arms, regardless of compliance with treatment assignment.

Missing data complicate the estimation of the ITT effect because they require the imposition of nonverifiable assumptions about the distribution of missing outcomes. To address this issue, the U.S. Food and Drug Administration recently commissioned a report from the National Research Council entitled “The Prevention and Treatment of Missing Data in Clinical Trials.”¹ We summarize some of the findings of this report and discuss its relevance to the conduct and analysis of orthopaedic surgical trials. Throughout this paper, we assume that the ITT effect is the target parameter of interest. The reader is encouraged to read the National Research Council report, wherein key references can be found.

A recent literature review by Herman et al. serves to support the relevance of our work about missing data to orthopaedic trials.² Specifically, they reviewed 274 randomized trials reported in eight leading orthopaedic journals between January 2005 and August 2008 and demonstrated that (1) strict adherence to the ITT principle was rare, (2) missing data were

common in studies that sought to estimate an ITT effect, and (3) the methods used for handling missing data were problematic.

In this paper, we first explain the ITT principle. We then review study design and prevention ideas that can help to minimize missing data and also discuss state-of-the-art methods for analyzing and reporting the results of studies with missing data, including the use of sensitivity analysis.

Intention to Treat

In the absence of missing outcome data, and without appealing to additional assumptions, the ITT effect is the only truly causal quantity that can be estimated from a randomized study. The ITT effect measures treatment effectiveness because it ignores treatment compliance. That is, it averages over the distribution of compliance patterns. This can be relevant to understanding the causal effect of a policy that assigns treatment, particularly when compliance patterns in the population where the policy will be implemented are similar to those in the trial.

In the presence of noncompliance, the ITT effect does not measure treatment efficacy, which is the causal effect of taking the assigned treatment. A common misconception is that treatment efficacy can be estimated by restricting the analysis to those who comply with their assigned treatment, the so-called “per-protocol analysis.” However, those who comply with their assigned treatment are typically a nonrandom subset of those randomized to that treatment. Compliers may exhibit characteristics or behaviors that influence their outcome independently from the effect of the treatment. For example, in a trial of a new medication to prevent heart disease, those who comply with their assigned treatment may also be more likely to exercise and diet (or less likely). There are specialized methods that require additional assumptions (even in the absence of missing outcome data) for

Disclosure: None of the authors received payments or services, either directly or indirectly (i.e., via his or her institution), from a third party in support of any aspect of this work. None of the authors, or their institution(s), have had any financial relationship, in the thirty-six months prior to submission of this work, with any entity in the biomedical arena that could be perceived to influence or have the potential to influence what is written in this work. Also, no author has had any other relationships, or has engaged in any other activities, that could be perceived to influence or have the potential to influence what is written in this work. The complete **Disclosures of Potential Conflicts of Interest** submitted by authors are always provided with the online version of the article.

estimating treatment efficacy, but a discussion of these approaches is outside the scope of this paper.

When outcome data are missing, additional assumptions and models are needed to compute the ITT effect; importantly, these assumptions cannot be validated empirically. Missing observations are caused by a number of factors, but in many cases this situation is avoidable. For example, in many studies, investigators fail to follow patients after treatment discontinuation, thereby creating a missing data problem and making it impossible to estimate the ITT effect without imposing additional assumptions. In many cases, this situation can be avoided by continuing to record patient outcomes even if treatment is discontinued. Although it is difficult to avoid missing data altogether, the ITT principle dictates that outcomes should be recorded on all individuals randomized to treatment, regardless of compliance.

Design and Prevention

Given a plan to follow all patients, regardless of treatment compliance, what types of design and prevention strategies can be put in place to minimize missing data? The National Research Council report identifies many techniques that study designers and managers can implement to reduce the frequency of missing data. The overarching idea of the points discussed below is to create procedures and a climate and culture that seek to maximize the collection of complete data.

First, the study design should limit the burden of data collection on study participants. This can be achieved by (1) minimizing the number of follow-up visits; (2) collecting only essential information at each visit; (3) having user-friendly case-report forms; (4) using, when possible, direct data capture that does not require clinic visits; and (5) providing relatively large time frames for follow-up visits. Data collection also should include auxiliary health status measures that are associated with a patient's decision to withdraw from follow-up and the outcome(s) under investigation. As will be discussed in the next section, these auxiliary factors will be useful in correcting for the potential bias introduced by missing outcome data.

Second, incentives for follow-up and study completion should be provided to both participants and study sites. Incentives to participants can include monetary compensation, which can be slightly backloaded so that there are greater payments as time in the study increases. Paying for voluntary participation in a clinical trial is considered ethical. The institutional review board must ensure that compensation is not coercive or does not reach a level of undue influence. Additional incentives to participants might include access to effective treatments after study termination. Incentives to study sites might include payment for the number of study visits completed rather than payment for each enrolled patient.

Third, study sites should be selected carefully to ensure that they have a good track record for enrolling, following, and collecting complete data on patients. Training of site personnel should emphasize the importance of complete data collection, regardless of whether the patient adheres to the assigned therapy.

Fourth, the consent process should ensure that patients understand the commitment they are making with regard to

providing complete data regardless of adherence to treatment. Patients should understand the difference between their rights to discontinue therapy and discontinue follow-up; they may discontinue therapy but can opt to continue to be followed and have outcomes recorded. Payment for follow-up visits should not be contingent on adherence to therapy.

Fifth, the study management team should set a priori targets for "unacceptable" missing data. With these targets in mind, data collection by each site should be monitored and reported in as close to real time as possible during the course of the study. These reports should be made widely available to study personnel; they should be presented to study investigators and site personnel at regular meetings/teleconferences and posted on the study web site. The identification of poorly performing sites can expedite remediation efforts, including additional training, site visits, or possibly site closure. In addition, the Data Safety Monitoring Board (DSMB) should closely monitor data completion rates. These rates should factor into their evaluation regarding the quality of the information being produced by the study. The overarching idea is to create a climate and culture that maximize the collection of complete data.

Sixth, study investigators should identify and aggressively, but not coercively, engage those patients at greatest risk for dropping out of follow-up. Identification of patients may involve asking all patients about their intention to attend the next clinic visit and understanding the barriers to participation³. Addressing these barriers on either the study or individual level should serve to maximize retention.

Seventh, it is essential to proactively engage study participants so that they feel invested in the research enterprise and appreciated for their efforts. Ideas include study newsletters, frequent updates of web site material, access to study reports and findings, study-branded gifts, regular expressions of thanks, social networking, solicitation of input on study conduct, and creation of an enjoyable experience at study visits.

Eighth, a reminder/contact system (e.g., e-mail, phone, or text) for study visits that is tailored to the individual patient should be used. This system should be reevaluated for the individual patient throughout the trial.

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

Analysis

In studies with missing data, the validity of any analysis will rely on untestable assumptions about the distribution of missing outcomes. Imagine a two-arm randomized trial in which a continuous outcome (higher values representing better health status) is scheduled to be recorded at baseline and K visits after baseline. Additionally, assume that some patients prematurely withdraw from the trial but that all outcomes prior to withdrawal are observed. Suppose that, in each treatment group, interest is focused on estimating the mean outcome at the final clinic visit, assuming that all patients will have their outcome measured at the final visit. Consider approaches for estimation of this mean for one of the treatment arms. To understand commonly

used assumptions, it is useful to stratify patients on the basis of the last visit at which they are observed. Patients whose last visit is the final visit are said to be completers; patients who are not completers are stratified into dropout cohorts on the basis of the number of completed visits.

Two commonly used analytic approaches are last observation carried forward (LOCF) and complete case (CC) analysis. In LOCF analysis, the outcome for a patient who drops out before the final visit is replaced by the last recorded observation. This analysis is based on the assumption that, within each dropout cohort, the mean outcome at the final visit would have been equal to the mean outcome at the cohort's last visit. This assumption is often considered tenuous for two reasons: (1) if patients withdraw because of deterioration of their health status after their last visit, their imputed outcomes will be biased high; and (2) outcomes further removed from the final assessment are less likely to be truly reflective of the health status at that time.

In CC analysis, the mean is estimated by the sample mean among the completers. This analysis will be valid if the patients who finish the study are a completely random sample from those randomized to the treatment arm. This assumption is often referred to as missing completely at random (MCAR). If patients who drop out are sicker or healthier than those who complete the study, then the CC estimator will be biased high or low, respectively.

To illustrate, consider the randomized clinical trial conducted by the Canadian Orthopaedic Trauma Society in which patients were randomized to receive surgical or nonsurgical treatment for displaced midshaft clavicular fractures⁴. In this trial, sixty-seven and sixty-five patients were randomized to the surgical and nonsurgical groups, respectively. Within these groups, five patients (7%) and sixteen patients (25%), respectively, were lost to follow-up at one year. Table II (complications) and Table III (appearance of shoulder) in their manuscript present the results for patients who were followed to one year (i.e., a CC analysis). Herman et al. analyzed the complication data with the LOCF method; patients who did not have follow-up data at one year were all assumed to be complication-free (see Table III of their manuscript)². The significance of the results for nonunion and overall complications differs between the two analysis procedures. In a letter to the editor, McKee criticizes the use of the LOCF analysis and points to information external to the trial indicating that at least four patients in the nonsurgical arm who were not followed to one year had complications⁵. We would suggest that the CC analysis is also problematic since the patients who are followed to one year are likely different (perhaps healthier) with respect to risk factors than those who were lost to follow-up.

An alternative (untestable) assumption that may be more tenable is the missing at random (MAR) assumption. This assumption states that within levels of outcomes recorded through visit k , the distribution of outcomes after visit k is the same for those who are last seen at visit k and for those who are last seen after visit k . Another way of expressing this assumption is to say that, among those who show up for their k th post-baseline assessment, the decision to drop out before visit $k + 1$ is like a “flip

of a coin,” with a probability depending on the outcomes recorded through visit k .

Under the MAR assumption, a common analytic approach is to specify a fully parametric model for the full data. A random effects (or mixed) model is often specified to account for the correlation of outcomes within individuals. When using these models, especially for non-normally distributed outcomes, the mean outcome at the final visit may not correspond to a unique model parameter, but instead may be a complicated function of all of the model parameters. One can also use the model to impute the missing outcomes to create a “filled-in” dataset. An inference can then be drawn on this dataset, with proper variance correction. To improve efficiency, multiple complete datasets can also be created, and the inference can be “averaged” over these datasets. This is referred to as “multiple imputation.”

To illustrate, consider the trial in which patients with acute vertebral fractures were randomized in equal numbers to receive kyphoplasty or nonsurgical care⁶. Outcomes of back pain and opioid use were scheduled to be measured at baseline; at five to ten days; and at months one, three, six, and twelve. Approximately 80% of patients randomized to the kyphoplasty arm and 70% of patients randomized to the nonsurgical arm had observed outcomes at twelve months. The data for each outcome type were analyzed with use of the (normally distributed) random effects model, with treatment group, visit, treatment by visit interactions, and baseline outcome as covariates. Each analysis is valid under the MAR assumption specific to the outcome type and correct specification of the random effects model. In our opinion, it is necessary to question the validity of the normality assumption, especially for the binary variable of opioid use.

The MAR assumption discussed above is limited to outcome variables. Usually, there are many “auxiliary” variables collected at each visit that can be useful to incorporate into the analysis. Specifically, these variables are useful because they help explain the reasons for dropout as well as help predict the missing outcomes. They can also serve to make the MAR assumption more tenable. A more general MAR assumption states that within levels of outcomes and *auxiliary variables* recorded through visit k , the distribution of outcomes after visit k is the same for those that are last seen at visit k and those who are seen after visit k . Another way of expressing this assumption is to say that, among those who show up for their k th post-baseline assessment, the decision to drop out before visit $k + 1$ is like a “flip of a coin” with probability depending on the outcomes and *auxiliary variables* recorded through visit k . In the kyphoplasty study, opioid use could serve as an auxiliary variable when analyzing the back pain outcome.

Under the more general MAR assumption, there are two main ways of estimating the mean of interest. These methods are not widely used in the scientific literature, primarily because of lack of statistical software. However, we believe that they represent the state of the art, and software is currently being developed. The two approaches differ in model specification. Both posit models for the distribution of the observable data and can therefore be subjected to goodness-of-fit procedures.

In the first approach, called the “G-computation algorithm,” fully parametric models are specified for the distribution

of the observable outcomes and prognostic factors scheduled to be measured at visit k ($k = 1, \dots, K - 1$), among those on-study at that visit, conditional on the history of these variables. A model for the conditional mean (conditional on the history of outcomes and auxiliary factors through visit $K - 1$) of the outcome at visit K , among those who complete the study, is also specified. The parameters from these models can be estimated with the use of standard software. The mean of interest can be estimated by essentially predicting the conditional mean of the outcome at visit K as a function of the longitudinal history of outcomes and auxiliary factors that would be observed through visit $K - 1$, all in the absence of dropout, and then averaging over these longitudinal histories. This can be accomplished by using the following sequential simulation procedure:

S0: Simulate from the distribution of baseline covariates (other than treatment assignment). This can be done by randomly selecting an individual from the pool of all patients (regardless of treatment assignment) and recording the covariates. Set $k = 1$.

S1: Given the simulated data at the previous step(s), simulate from the estimated conditional distribution of the outcomes and auxiliary prognostic factors at visit k . Set $k = k + 1$.

S2: If $k < K$, go to S1; otherwise go to S3.

S3: Given the simulated data at the previous steps, predict the conditional mean of the outcome at visit K using the conditional mean model.

Repeat this procedure (e.g., 10,000 times) and take an average of the predicted means.

An alternative, albeit less efficient, procedure is to impute the outcomes for just those who dropped out. For the imputation procedure, one needs to specify not just a model for the conditional mean of the outcome at visit K , but actually its full conditional distribution; the prediction in S3 in the above algorithm is replaced by simulating from this conditional distribution given the data from the previous steps. To impute the outcome for a patient who was last seen at visit k^* , one starts the simulation at S1 with $k = k^* + 1$ and uses the patient's observed data through visit k^* . A "filled-in" dataset will then be created, and the sample mean of the outcome at the final visit can be computed. To improve the precision of the estimator, multiple imputation is recommended.

The second approach under the more general MAR assumption involves modeling the conditional (discrete-time) hazard of last being seen at visit k , given the history of observable data through visit $k - 1$. The parameters of this model can be estimated with standard software. For each individual who is a completer, the probability of being a completer is estimated from this model; this probability is estimated as the product (from visits 1 to K) of one minus the estimated hazard. The mean at the final visit is then estimated as the weighted mean of the outcomes for completers, where the weights are equal to the inverse of the estimated probability of completion. This procedure is called "inverse-weighted estimation." The idea behind this technique is that each completer is given increased influence to reflect himself plus people like him (in terms of prognostic risk

factors) who dropped out prematurely. For example, if a completer's probability of completion is estimated to be 0.25, then he counts for himself plus three other individuals like him who did not complete the study. This inverse-weighting procedure is relatively easy to implement, but the resulting estimator can have high standard errors, especially when some of the estimated probabilities are close to zero. There is active research into modifications of the inverse-weighted estimation procedure to improve efficiency and robustness.

For all of these methods, the easiest approach to assess uncertainty due to sampling variability (i.e., to compute standard errors and confidence intervals) is to use resampling-based procedures, such as the bootstrap method.

Sensitivity Analysis

When MAR fails (a possibility that is untestable), the missing data are said to be missing not at random (MNAR). The National Research Council report suggests that a sensitivity analysis be conducted to evaluate the robustness of the results to deviations from the MAR assumption. Specifically, the report recommends positing a class of assumptions for modeling the missing data that are indexed by sensitivity analysis parameters that govern the degree of departure from MAR. There are many ways that this can be done.

For example, for patients who are participating in the study through visit k , it might be assumed that the decision to withdraw before the next visit is like a "flip of a coin," with a probability depending on past observable factors (including baseline covariates, outcomes, and auxiliary prognostic covariates) through visit k and on the outcome scheduled to be measured at visit $k + 1$, but with no additional outcomes or auxiliary factors. The residual influence of the outcome at visit $k + 1$ on the probability cannot be estimated from the observed data, so its influence is assumed to be governed by a sensitivity analysis parameter (e.g., α) that is first fixed (for estimation purposes) and then varied. The parameterization is designed so that when $\alpha = 0$, there is no residual influence and MAR is obtained; when $\alpha > 0$ (< 0), patients who are healthier (sicker) with regard to the outcome at visit $k + 1$ are more likely to withdraw than those who are sicker (healthier). Another way of articulating this approach is to link the conditional distributions (conditional on past observable factors) of outcomes at visit $k + 1$ for those who were last seen at visit k and those who remain in the study through that visit. The parameter α governs differences between these distributions, where $\alpha > 0$ (< 0) implies that the distribution of outcomes at visit $k + 1$ for those last seen at visit k is more heavily weighted to higher (lower) values than for those who remain in the study through visit $k + 1$. For a fixed α , natural extensions of the G-computation, multiple imputation, and inverse weighting methods exist.

If only extreme values of α yield a conclusion different from that under MAR, then the study results are said to be robust; otherwise, the results are said to be nonrobust. Since we view the general MAR assumption (with appropriate auxiliary factors) as a likely closer approximation to the true mechanism by which data are missing, this form of sensitivity analysis is

preferred to an approach in which robustness is determined by analyzing the data with multiple disparate approaches, such as LOCF, CC, and random effects modeling.

Conclusions

In this paper, we have argued that randomized trials should be designed to estimate, at a minimum, the ITT effect, as this is the only causal effect that can be estimated in the absence of missing data without additional assumptions. With this viewpoint, follow-up of patients after treatment has been discontinued must take priority.

There is no question that the best advice regarding missing data is to avoid it entirely. Resorting to the sophisticated statistical analysis techniques should only be done after serious efforts have been employed to minimize missing data with the design and prevention techniques. Underinvesting in design and prevention techniques is penny wise and pound foolish because the resulting uncertainty about treatment effects, which must reflect both sampling variability and lack of knowledge about the mechanism by which the data were missing, can be too large to reach definite inferences.

We recognize that, even with reasonable investment in design and attempts at prevention, missing data will still occur. When this happens, naive methods such as LOCF and CC analysis

should not be employed unless there is strong underlying scientific justification. For studies in which missing data are a result of patient dropout, the general MAR assumption is often considered a reasonable benchmark assumption. Under this assumption, a rigorous statistical analysis using G-computation, multiple imputation, or inverse weighting is essential. Since MAR is untestable, the sensitivity of the inferences to deviations from the assumption must be evaluated and reported. ■

Daniel O. Scharfstein, ScD
Department of Biostatistics,
Johns Hopkins Bloomberg School of Public Health,
615 North Wolfe Street, Baltimore, MD 21205.
E-mail address: dscharf@jhsph.edu

Joseph Hogan, ScD
Department of Biostatistics, Box G-S121-7,
Brown University, Providence, RI 02912

Amir Herman, MD, PhD
Department of Orthopedic Surgery,
Chaim Sheba Medical Center,
Tel-Hashomer, 52621, Israel

References

1. Panel on Missing Data in Clinical Trials. The prevention and treatment of missing data in clinical trials. Washington, DC: The National Academies Press. 2010.
2. Herman A, Botser IB, Tenenbaum S, Chechick A. Intention-to-treat analysis and accounting for missing data in orthopaedic randomized clinical trials. *J Bone Joint Surg Am.* 2009 Sep;91(9):2137-43.
3. Leon AC, Demirtas H, Hedeker D. Bias reduction with an adjustment for participants' intent to dropout of a randomized controlled clinical trial. *Clin Trials.* 2007;4(5):540-7.
4. Canadian Orthopaedic Trauma Society. Nonoperative treatment compared with plate fixation of displaced midshaft clavicular fractures. A multicenter, randomized clinical trial. *J Bone Joint Surg Am.* 2007 Jan;89(1):1-10.
5. McKee MD. Response to "Intention-to-treat analysis and accounting for missing data in orthopaedic randomized clinical trials." 2010 Jan 11. <http://www.jbjs.org/article.aspx?articleid=29336>. Accessed 2012 Mar 27.
6. Wardlaw D, Cummings SR, Van Meirhaeghe J, Bastian L, Tillman JB, Ranstam J, Eastell R, Shabe P, Talmadge K, Boonen S. Efficacy and safety of balloon kyphoplasty compared with non-surgical care for vertebral compression fracture (FREE): a randomised controlled trial. *Lancet.* 2009 Mar 21;373(9668):1016-24. Epub 2009 Feb 24.