Missing Data in Clinical Studies

Amit K. Chowdhry, MD, PhD,* Vinai Gondi, MD,† and Stephanie L. Pugh, PhD‡

*Department of Radiation Oncology, University of Rochester Medical Center, Rochester, New York; †Northwestern Medicine Cancer Center Warrenville and Northwestern Medicine Proton Center, Warrenville, Illinois; and ‡NRG Oncology Statistical and Data Management Center, American College of Radiology, Philadelphia, Pennsylvania

Received Feb 17, 2021. Accepted for publication Feb 19, 2021.

Case Vignette

The NRG Oncology Cooperative Group conducted a randomized phase 3 trial comparing whole brain radiation therapy (WBRT) plus memantine to hippocampal avo"
dant WBRT (HA-WBRT) plus memantine in patients with brain metastases. HA-WBRT was shown to significantly delay the time to neurocognitive failure, which was the study’s primary endpoint, without a significant decrease in overall survival or increase in intracranial progression.1 Patient-reported outcomes (PROs) were collected on the trial as a secondary endpoint. The UT—MD Anderson Cancer Center Symptom Index — Brain Tumor module is a validated disease site-specific module that consists of the 19 multisymptom items related to symptom severity and symptom interference with daily life, as well as 9 items specific to patients with brain tumors.2,3 In NRG-CC001, the change from baseline to 6 months in symptom burden was the primary PRO endpoint, with symptom interference, cognitive symptoms, and neurologic symptoms domains as the secondary PRO endpoints. There is significant interest in comparing PRO results between the arms, but some patients were lost to follow-up in the study, thus leading to the question of how the missing data should be handled to perform the most meaningful analysis.

The Problem of Missing Data

Virtually all clinical studies involve some proportion of unobserved or missing data. These data may be missing for a variety of reasons, including subject dropout, survey nonresponse, and data entry errors from institutions. There are a variety of approaches to deal with missing data. Although no approach can guarantee accurate estimations and conclusions, principled approaches (which take advantage of all available data and are valid under relatively loose assumptions) are preferred because they are more likely to yield reliable estimates and inferences.4,5 In general, one should always prospectively try to limit the amount of data that are missing in a clinical study, and a 2012 report in the New England Journal of Medicine highlighted approaches to limit the amount of missing data in a study.6 However, loss to follow-up and other causes of missing data remain a problem, and other causes are unavoidable, such as death. In addition, nonsurvival outcomes (eg, PROs), which require extra effort on the part of the patient, their caregivers, and/or their health care team, create more barriers to data collection.

The goal of this article is to educate readers about missing data and to discuss recommended ways to handle them using appropriate statistical methods.

Data Sharing Statement: All data in this article have been presented in previously presented and published works, and thus there are no new data to share.

Acknowledgments—The authors thank David Sher for his helpful comments on this manuscript.
Missing Data Assumptions

Statisticians think about missing data by making certain types of assumptions about them. First, one may assume that there are no systematic differences between recorded and missing data, a condition termed missing completely at random (MCAR). Second, one may instead assume that, although there are systematic differences between observed and missing data in a way that might change the conclusions of the study, those differences can be explained by some other observed information for which the results can be adjusted (e.g., other demographic or tumor characteristics). This case is known as missing at random (MAR). In the case of NRG-CC001, if hypothetically patients with lung cancer brain metastases were consistently more likely to have missing PRO results, they would be considered MAR (assuming the differences in outcomes between lung cancer patients and other patients could adequately explained by the statistical model being used).

Finally, one may think that the missing data are systematically different than the observed data but that these differences cannot be explained by the observed data alone using the statistical model. For instance, the reason why the data are missing may be due to the missing data themselves. Consider the scenario in which WBRT patients with worse cognitive function would be more likely to have missing data. This situation is known as missing not at random (MNAR). It is never possible to determine whether the data are MNAR because this would require knowing the actual values of the missing data. Moreover, MNAR is not often considered a single assumption but as a family of assumptions that may be made about the missing data. Analyses under MNAR are typically performed in addition to the primary statistical approach (known as sensitivity analyses) and are conducted to determine whether the results change under different assumptions.

The biggest threat from missing data is that they can cause a form of selection bias. The goal of all principled methods for handling missing data is to avoid such bias in the effect estimate and thereby avoid making incorrect inferences. An additional benefit of methods that do not throw out information is that they generally increase statistical power. If there is a large proportion of missing data, the benefits of randomization can be lost, especially if the probability of missingness is differentially associated with the treatment arms. If there are very few missing datapoints, it is unlikely that different methods of handling missing data would yield different conclusions. However, it is uncertain at what proportion of missing data principled methods provide an advantage. Some authors have proposed rules of thumb that say if the proportion of missing data is <5% to 10%, then the choice of missing data method likely does not matter (barring large deviations from the MCAR assumption). If the proportion is >40%, then the results should be considered hypothesis generating. Even when the proportion of missing data is large, principled methods perform better than simple methods. In NRG-CC001, the percent of patients with completed PRO questionnaires ranged from 83% at 2 months to 75% at 6 months, highlighting the need for careful management of the missing data.

Methods to Analyze Data When Some Observations Are Missing

Methods that may yield poor estimates

Complete case analysis

Many implementations of statistical methods in commonly used software programs will often automatically delete whole subjects’ information if they are missing 1 assessment. This rule not only throws out a lot of information (which decreases power), but it can be a source of bias. This approach is known as complete case analysis. Consider a registry study where all subjects except 3 are missing only 1 of 20 covariates. For simplicity, assume roughly equal numbers of patients are missing each of the 20 covariates. If someone uses a package that defaults to complete case analysis, only 3 subjects would be included in the analysis, even though most covariates are present for most subjects. Most situations are not so extreme, but many observations can be deleted this way despite only having a small amount of missing information (and the reader may never know the difference). This dilemma would also apply to longitudinal clinical trials where a PRO is measured on many occasions. If a substantial number of patients happened to miss a single follow-up visit, then many patients may be excluded from an analysis despite having a lot of timepoints, which may have helped in drawing conclusions.

A natural conclusion from this discussion is that using a principled missing data method can sometimes do more than just reduce bias: It makes better use of the information we do have and thus allows for possibly more power. Of note, a complete case analysis cannot be an intent-to-treat analysis because some patients assigned to a particular treatment group are excluded from the analysis. Complete case analyses can be useful in the setting of performing multiple analyses under different missingness assumptions. If all analyses agree, then the results are likely robust to different assumptions of missing data. The only situations in which a complete case analysis should be performed (an analysis valid only under MCAR) are when sensitivity analyses are performed or in the setting of a very small proportion of missing data.

In NRG-CC001, approximately 10% of patients completed assessments outside the specified timeframe at each timepoint and thus were excluded from complete case analyses, such as the t-test comparing the change from baseline to 6 months between the treatment arms for each of the 4 UT—MD Anderson Cancer Center Symptom Index
Brain Tumor module domains assessed. A statistical modeling approach using all completed assessments at the time from randomization to completion was performed with the 6-month difference tested between the arms. The results from the 1-year follow-up analysis were presented by Armstrong et al. The secondary endpoint of patient-reported symptoms was powered using the symptom severity score at a type I error of 0.05. Symptom severity was found to be not statistically significant on complete case analyses (mean between arm difference = -0.26; \( P = 0.083 \)). The complete case analysis was used in addition to another method for handling missing data in 2 separate analyses. The second analysis performed will be discussed later in this section.

**Mean imputation and single regression-based imputation**

Mean imputation is a missing data approach that takes the mean of the observations and substitutes it for the missing observation. This technique has the potential to generate confidence intervals that are erroneously smaller than they should be and bias results. Consider a study where one wants to determine whether treatment is associated with a quality-of-life outcome, but some subjects withdraw from the study. If the quality-of-life information is missing and one substitutes the mean quality of life from all subjects that are available to analyze, then the statistical method may be biased unless the data are missing completely at random. For NRG-CC001, if mean imputation were used, we may have seen bias in either direction that may have led us to falsely conclude that neurocognition was either worse or better by some amount in either the HA-WBRT or standard WBRT groups. Moreover, the estimated standard error will be artificially smaller, which again may result in erroneous inferences.

A method that is similar to mean imputation is single regression-based imputation. Single-regression-based imputation uses regression to impute the missing value, but it treats the estimated point as if it were an observed datapoint. This method has less chance of being biased than mean imputation by using other variables to help predict the missing datapoint, but the standard error produced is also artificially low. For both mean and single-regression imputation, one is essentially acting as though one has data one does not actually have by filling the data in, with no adjustment to the standard error. If single-regression-based methods were used in NRG-CC001, falsely smaller \( P \)-values may have been seen, leading us to conclude statistically significant differences in neurocognition between the treatment groups when these differences do not, in fact, exist.

**Last observation carried forward**

Another method that has received criticism in recent years is the last observation carried forward approach. Consider a PRO measured every 3 months for 2 years. If a subject were to drop out of the study before the 12-month PRO survey, then his or her 9-month value would be substituted for the 12-month datapoint. A simple example of when this type of analysis might fail is when subjects with a poorer quality of life are more likely to drop out of a study. In that situation, the missing data would be either MAR or MNAR depending on whether there are sufficient data to explain the differences between the subjects with observed and unobserved 12-month timepoint data. This method is not valid under either assumption, except for the exact situation where there is no change in the outcome measure if the subject drops out of the study—a situation that is often not the case, including the setting of brain metastases. In the case of NRG-CC001, last observation carried forward would most likely yield unacceptably inaccurate estimates because neurocognitive effects of radiation unfold over time and would not necessarily be as apparent at early follow-up as during later visits.

**Methods that may yield better estimates**

**Likelihood-based methods**

Likelihood-based methods, including many commonly used mixed models, may be used to analyze the data. These methods are some of the most commonly used techniques in clinical trials because they are often valid under the MAR assumption and generally do not require additional complicated modeling decisions. These methods use optimization to find the statistical model with the best fit to the data in a process analogous to minimizing the cost function in intensity modulated radiation therapy, balancing the optimal dose to the tumor while sparing normal tissues. These models allow for covariates to be included that may possibly adjust for differences between patients with and without missing data and include all available data for each patient. For instance, if a patient is missing the third of 5 assessments, all available assessments will be included in the model. One limitation is, as standardly implemented in software packages, likelihood-based methods are valid under MAR but not MNAR assumptions because there is no single MNAR assumption. Rather, MNAR is a family of assumptions with many possibilities for how missing data may differ from observed data. It is not as straightforward to write software that works on a wide variety of MNAR assumptions for likelihood-based methods as it is for certain multiple imputation methods, as we will discuss.

**Multiple imputation**

One of the most flexible approaches to handling missing data is multiple imputation. Along with likelihood-based methods, this class of principled method is very commonly used to handle missing data. Multiple imputation may be performed under a number of assumptions, including MCAR, MAR, and MNAR assumptions, with MAR and MNAR being more commonly applied in practice. The technical details behind multiple imputation are beyond the scope of this article, but the idea behind it is simple. In
contrast to single imputation methods, such as mean imputation, with multiple imputation methods, the original data set is first replicated multiple times. In each data set, a statistical model is used to generate values to fill in the missing datapoints, with each model adding some randomness to the missing datapoints (either based on a sample of the data or even by incorporating random noise). Subsequently, the estimates of treatment effect are aggregated across the data sets, and a measure of the variation across data sets is used via a formula to increase the estimate of the standard error. Because the imputation and combination procedures are separate from the model-fitting procedure (ie, the process to generate the imputed data and the model to analyze the outcome are different and distinct), multiple imputation may be used with a wide variety of statistical models for analyzing data.

<table>
<thead>
<tr>
<th>Do’s and Don’ts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do report the proportion of missing data used for primary analyses (including covariates) in all studies.</td>
</tr>
<tr>
<td>Do know the reasons why the data are missing.</td>
</tr>
<tr>
<td>Do perform sensitivity analyses to determine whether changing the assumptions changes your interpretation of the study; if results are not conflicting, it suggests results are robust to different assumptions.</td>
</tr>
<tr>
<td>Do not perform a complete case analysis without additional sensitivity analyses when the number of missing participants is significant. Even under the tight assumption of missing-completely-at-random, complete case analysis generally has lower power than other methods such as likelihood-based methods and multiple imputation. Furthermore, when the missing-completely-at-random assumption is not met, complete case analyses can be biased.</td>
</tr>
<tr>
<td>Do not use single imputation methods for imputation of outcome variables.</td>
</tr>
<tr>
<td>Do play close attention to the imputation models used when using multiple imputation methods to make sure they are appropriate.</td>
</tr>
</tbody>
</table>

Glossary

Bias: The statistical definition of bias is somewhat different than its common English definition. In layman’s terms, a biased estimate is an estimate that is systematically either larger or smaller than the true value.

Missing-Completely-at-Random (MCAR): An assumption that there are no systematic differences between the data that are observed and the data that are missing.

Missing-at-Random (MAR): An assumption that, although there are systematic differences between observed and missing data in a way that might change the conclusions of the study, those differences can be explained by some observed information (eg, other demographic or tumor characteristics).

Missing-Not-at-Random (MNAR): The situation where the missing data are systematically different than the observed data, but these differences cannot be explained by the observed data used.

Complete Case Analysis: An analysis where all subjects with any missing data for any of the variables in the model are excluded from the analysis.

Impute: To substitute, or fill in, a missing data point, with a value.

Returning to NRG CC001, to account for missing data from an additional 60 patients using a principled method for the patient-reported symptoms analysis, an a priori specified multiple imputation was performed, rerunning the statistical model and testing the 6-month between-arm difference. In this imputed analysis, symptom severity was found to be statistically significantly less in the HA-WBRT arm compared with the WBRT arm (mean between arm difference = -1.37; \( P < .001 \)) in favor of the HA-WBRT arm. The results from the imputed analysis were in the same direction as the complete case analysis, but the additional 60 patients provided enough statistical power to find a significant difference for symptom severity.

Other methods

Other methods for handling missing data include inverse probability weighting (IPW) and Bayesian methods. IPW is an approach comparable to propensity score analyses. Similar to weighted propensity score analyses, IPW analyses weight the observed data in the analysis by the inverse of the probability that the data are not missing. There is a rich literature on Bayesian methods for handling missing data, including multiple imputation methods and data augmentation. A detailed discussion of these topics is beyond the scope of the article, but interested readers can see Molenberghs et al.\(^\text{12}\).

Missing not at random and sensitivity analysis

It is often beneficial to strengthen statistical inference by performing a variety of analyses to determine the sensitivity of the results to different assumptions about the missing data. The more consistent the findings, the more one can believe in the results. Such sensitivity analyses are performed under MNAR assumptions and thus involve making untestable hypotheses about the missing data (ie, why or how the data are not missing at random) and seeing how those assumptions influence the results. Selection models and pattern mixture models are 2 common approaches for conducting an MNAR analysis, the specifics of which are beyond the scope of this article.\(^\text{12,13}\)

Conceptually, in a selection model, one models the probability that an outcome is missing and incorporates this information into the final analysis. In a pattern mixture model sensitivity analysis, one does not make any assumption about why the data are missing but rather...
models how these missing results may differ from the known observations. Of course, they may differ in a number of ways, and thus one approach to this sensitivity analysis is to generate multiple models and examine how different the missing data must be from the observed data to change the conclusion of the analysis.

If there are systematic differences between patients with and without completed assessments, sensitivity analyses may actually show no statistically significant difference whereas the complete case analysis shows a significant difference. Hence, sensitivity analyses should not be restricted to nonsignificant complete case analyses. In NRG-CC001, estimates of the 6-month between-arm differences were similar in both the complete case and imputed analyses, suggesting that there were no systematic differences between patients with or without missing data and further strengthening the conclusions that can be drawn from these missing data analyses.

Conclusions

Knowing how much data are missing and how the missing data are handled can affect conclusions for clinical research studies. Researchers and readers of clinical studies should be cautious when a significant proportion of data are missing and the reasons for missingness are unknown, even when principled methods under the MAR assumption are used.

Particularly with an incorrect choice of missing data method, there is a high risk of bias with a large proportion of data missing, although the risk of bias always exists regardless of the method used. For this reason, analyses using at least 2 methods of handling missing data that are valid under different assumptions or sensitivity analyses can help with the robustness of conclusions.14

Nothing replaces the full complement of data, but methods valid under MAR generally perform better than methods that assume MCAR, and MNAR sensitivity analyses may allow for further exploration of whether deviations from the MAR assumption may change the conclusions of the study. Sensitivity analyses have a growing role in modern studies and are likely to have a much larger role in the statistical analysis of clinical studies going forward.

References