Missing Data in Longitudinal Trials - Part B, Analytic Issues

Juned Siddique, DrPH,
Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago

C. Hendricks Brown, PhD,
Prevention Science and Methodology Group, Department of Epidemiology and Biostatistics, College of Public Health, University of Illinois at Chicago

Donald Hedeker, PhD,
Division of Biostatistics, University of Illinois at Chicago

Naihua Duan, PhD,
Professor of Biostatistics in Psychiatry, Departments of Biostatistics and Psychiatry, Columbia University; and Director, Division of Biostatistics, N.Y. State Psychiatric Institute, New York, N.Y.

Robert D. Gibbons, PhD,
Professor of Biostatistics and Psychiatry, Director of the Center for Health Statistics, University of Illinois at Chicago

Jeanne Miranda, PhD, and
UCLA Health Services Research Center

Philip W. Lavori, PhD
Professor of Biostatistics, and Chair, Department of Health Research and Policy, Stanford University School of Medicine

Abstract

Longitudinal designs in psychiatric research have many benefits, including the ability to measure the course of a disease over time. However, measuring participants repeatedly over time also leads to repeated opportunities for missing data, either through failure to answer certain items, missed assessments, or permanent withdrawal from the study. To avoid bias and loss of information, one should take missing values into account in the analysis. Several popular ways that are now being used to handle missing data, such as the last observation carried forward (LOCF), often lead to incorrect analyses. We discuss a number of these popular but unprincipled methods and describe modern approaches to classifying and analyzing data with missing values. We illustrate these approaches using data from the WECare study, a longitudinal randomized treatment study of low income women with depression.

WHY DATA ARE MISSING AND HOW THIS AFFECTS INFERENCES

Missing Data Mechanisms

Missing data mechanism refers to the underlying process of generating missing data. For example, in a depression trial, subjects who remain depressed may be more likely to drop out of the study. The statistical properties of all missing data methods depend on the value itself and the values of the other variables. The most important question is how the chance of observing a particular value of a variable depends on what that value (and others) actually is. [1]

Rubin's[2] classification of missing data mechanisms into three types is now standard. The first, and least problematic type is "missing completely at random" (MCAR) where the probability that a value is missing does not depend on any values (observed or missing) in the dataset. Under MCAR, observed values can be thought of as a random sample from the full set of observed and unobserved values. For example, consider the problem of estimating the prevalence of a psychiatric disorder based on an in-person assessment with a psychiatric diagnostic instrument. If everyone in a representative sample of the population is assessed on this instrument, the prevalence estimate can be obtained readily. However, it is often cost-effective to conduct a study in two stages, beginning with a short interview using a screening instrument followed by the in-person interview on a subsample of subjects for diagnostic assessment of the disorder. To keep this example simple, we assume that the screen is given to everyone and the more expensive interview is missing on some subjects.

To illustrate MCAR, imagine that 1) there are no refusals to either the screening or in-person interviews, and 2) a random subsample of those given the screen is selected for an in-person interview. In

Bias and precision of the design depend on the realities of study conduct in the real world.

this situation, the missing data mechanism satisfies MCAR; the subsample interviewed in person is a representative subsample of the sample interviewed originally by phone.

It is unusual for conditions 1) and 2) to be met in practical field studies. For reasons to be discussed below, the selection of the in-person interview subsample might take the data obtained from the screen into consideration. In addition, refusals often arise in interviews, and it is common for them to be related to data values (eg, patients with the disorder might be more

likely to refuse to be interviewed). Therefore, MCAR is not a realistic mechanism for most practical applications.

A more realistic missing data mechanism is "missing at random: (MAR), where the probability that a value is missing may depend on observed values in the dataset but does not depend on any missing data. To illustrate the difference between MAR and MCAR, we continue with the example above but this time change the study design. More specifically, we now assume that a) the initial screen contains questions about the disorder, and b) the selection of the in-person interview subsample is stratified by the results of the screening assessment. For example, 100% of those who screened positive are selected for in-person interviewing, and a random 10% of those who screened negative are selected for in-person interviewing. (This design is discussed further in the article by Lavori et al in this issue, see page 784.[3])

Under this design, the missing data mechanism satisfies MAR but not MCAR—the missing data mechanism now depends on the screening results, violating the requirement in MCAR for the missing data mechanism not to depend on any data at all. MAR is satisfied because the missing data mechanism depends only on observed data (screening status) and does not depend on any missing data.

Note that the subsample interviewed in-person is not a representative subsample of those interviewed by phone. The subsample overrepresents those who screened positive in the phone interview. An appropriate analytic procedure needs to be used to address this bias. In particular, we can weigh the screened subsample by the sampling weights, defined as the reciprocal of the sampling probability (100% for screen positives, 10% for screening negatives). In other words, we weight each screen negative interviewee by 10, because each one of them represents 10 screen negatives out of which one is selected; the screen positives are weighted by 1, because each of them only represents himself/herself.

The third type of missing data mechanism is "not missing at random" (NMAR), corresponding to situations when missing data is neither MCAR nor MAR. Under this mechanism, the probability for a data element to be missing may depend on missing data. To illustrate the difference between MAR and NMAR, we continue with the previous example but assume this time that some subjects might refuse to be interviewed in person. If the refusal probability does not depend on any data value at all (unlikely in most practical applications), we will have MCAR. If the refusal probability depends on observed data for example, men are more likely to refuse than women, we will have MAR (here we assume that gender is observed for all subjects). If the refusal probability depends on missing data, for example, subjects with disorder are more likely to refuse, we will have NMAR (here we assume that the disorder status is not observed for subjects who refused to be interviewed).

With NMAR, missing values are systematically different from observed values, even after conditioning on observed values.[4] This is much harder to deal with than MCAR and MAR in statistical modeling and data analysis. Even our best statistical analyses can behave rather poorly if the missing data mechanism is NMAR, so it is important to minimize these effects either through design and/or analytic considerations.[5,6]

When the missing data mechanism satisfies MCAR or MAR and some other technical conditions hold, the missing data mechanism is sometimes referred to as ignorable.[2] The term ignorable means that it is not necessary to specify explicitly the missing data mechanism, (ie, the missing data mechanism can be ignored). But the analysis still needs to take the missing data into account to avoid bias (eg, to use weighted analysis in the two-phase design discussed earlier). To clarify, it is the missing data mechanism that is ignorable, not the missing data.

Summarizing and describing the pattern of missing data

In the context of repeated measures or other longitudinal data, missing data can potentially occur for any or all variables. The timing of when a subject's data first become missing during the course of a study is often relevant. For example, participants who provide every measurement up to a certain time point and then fail to do so for the remaining duration of the study are referred to as measurement dropouts (not to be confused with "treatment non-compliers," who may or may not also be missing measurements after they stop complying). Because participants may drop out of the measurement plan of a study for reasons related to the quantity being measured or related to the study treatment, it is often necessary to use missing data methods that take into account dropout status.

Another type of missing data pattern is intermittent missing data where a participant completes the study but does not respond to every survey. Finally, datasets may be complete for baseline covariates in the analysis model (for variables such as age, treatment status, gender) and partially missing for outcome data, or have both missing outcome and covariate data.

How can study design minimize the possibility and effect of missing data

Of course, the best way to handle missing data is to avoid it or limit the amount during data collection. Part A of this series on missing data in longitudinal trials[3] makes recommendations for minimizing the possibility and effect of missing data. Briefly, reconsideration of study goals and measured outcomes may avoid difficulties of design that masquerade as and compound missing data problems so that an investigator can minimize the rate of missing data, particularly of the NMAR variety. Information about the reasons for missing data and proxies for the missing data should be collected whenever possible because the more data that are available, the closer the mechanism approaches MAR, and high quality estimates will have limited bias or misleading confidence intervals. Later we will show how to use such ancillary information in analysis.

AD-HOC (AND GENERALLY FLAWED) APPROACHES FOR HANDLING MISSING DATA

We describe here the common ad-hoc approaches for handling missing values, which are often used when analyzing longitudinal data, because they are easy to implement and do not require special software. Despite their common use, they rely on implicit assumptions that are usually unreasonable and often lead to invalid inference.

Last observation carried forward

With last observation carried forward (LOCF), missing values are replaced with the most recent previously observed value in the same patient. The filled in dataset is then analyzed as if there had been no missing data. This substitution of previously observed values for missing data can be performed for both intermittent missing values and measurement dropouts in repeated measures designs. Very strong and often unrealistic assumptions have to be made to ensure the validity of this method. First, LOCF assumes that a subject's true but unmeasured status stays at the same level from the moment of truncation onward (or during the period they are unobserved in the case of intermittent missingness).[7] In other words, there is a perfect relationship between the last observation and those following it. The prior trajectory of the subject is not taken into account, and any change is assumed to level off immediately. For intermittent missing data, the subsequent trajectory of the subject after the "gap" is not taken into account either. Further, as will be discussed later in this article, LOCF (like all substitution and single imputation procedures) overestimates precision by treating imputed and actually observed values on equal footing. It is often believed, erroneously, that LOCF is conservative,

thus does not lead to an inflated type I error rate. For point estimates, LOCF might underestimate the improvement in the experimental arm, if there is a systematic improvement in the outcome over time. However, the same underestimation might also happen in the control/placebo arm. Therefore, it is not clear whether the treatment effect based on contrasting the trajectories in the two arms is under-estimated or not. Furthermore, the overestimation of the precision might lead to underestimation of the standard error and inflation of the type I error. There are several published examples where LOCF does poorly.[8-10]

Mean substitution

In the context of longitudinal studies, mean substitution is typically implemented by replacing a missing value with the average (over other patients') observed value for the same variable and then analyzing the dataset as if it were complete. Although this method does preserve the overall mean for the time period, it has two serious disadvantages. Mean substitution does not preserve relationships among other variables in the data. For example, if a subject's month 2 depression score is missing, substitution of the mean at month 2 ignores that person's depression score for months 1 and 3. Mean substitution, therefore, always attenuates correlations between the measures. Finally, as with all substitution and single imputation procedures, mean substitution does not take into account uncertainty in the true but unknown value.

Regression substitution

Regression substitution extends the mean substitution method by using a regression substitution estimate to replace a missing data point. For each subject's missing data, the predictor variables consist of all those that are non-missing, with regression substitution coefficients computed from the remaining data. Although this procedure is a substantial improvement over LOCF and mean substitution, it is still unsatisfactory because missing data are replaced with values having too little variability, resulting in bias in correlations and over-estimation of the precision.

Complete-case analysis

Complete-case analysis involves discarding all observed data elements for subjects who have any missing values and restricting the data analysis to the remaining complete cases. This is the simplest procedure for handling missing data. It is usually done automatically by most software packages when missing data are encountered so that the dataset can be analyzed using standard complete-data methods. Unless the observations with missing values are only randomly different from those without missing values (ie, unless the data are MCAR), complete-case analysis will produce biased estimates. Complete-case analysis can also result in substantial information loss, by discarding an entire subject's data because of a few missing items. Rather than discarding an entire observation because of a single missing value, methods that make better use of all available information will provide estimates that are more precise and less biased.

End-point analysis

End-point analysis, a form of LOCF (see Gibbons and colleagues[11] for a review of limitations) is a procedure that concentrates on baseline and the last observed measurement for each individual, ignoring all observations between these times. Although the baseline period is usually the same for each individual, the end point will be different for each individual depending on if and when they drop out of the study. Typically, some form of difference or adjusted score is calculated from the baseline and end-point scores, and these difference or adjusted scores are compared across treatment groups.

By using only the last observed measurement for each individual, missing values are no longer an issue (except for those who have no follow-up data). However, there are many drawbacks to this approach. First, data between the first and last time points are ignored. This is problematic because a large amount of information is being discarded leading to reduced efficiency of parameter estimates. In addition, the researcher is no longer able to study individual trends over time, one of the original goals of longitudinal research.

A further drawback to end-point analysis is that since the time of the last measurement can vary for each individual, time is effectively ignored in the analysis. As a result, between-group comparisons can be confounded with time, since subjects in one group may have been assessed under a different period than subjects from another group. Within each group the length of the period itself may be influenced by the treatment. For example, if placebo-treated participants are more likely to drop out earlier than participants receiving the active drug, estimates of the treatment effect will favor the active drug even if the improvement rate is identical.[12]

Single imputation

Single imputation is a general method of replacing missing values with plausible values. It differs from the previous methods in that the imputed value has the same distribution as the non-missing data. One way to do this is to correct the regression substitution method, which uses a prediction equation to adjust for a person's own non-missing variables by adding in a random component to mimic the additional variability that real data would be expected to have around this predicted value. For each variable that has any missing data, a regression substitution model for imputation is developed, which uses a person's non-missing data to form a best predictor of that person's missing data. To this predictor, a random component is added based on the residual variance of this regression substitution model. These singly imputed values are said to be drawn from a predictive distribution of the missing values, conditioning on that subject's observed data. Once each missing value is imputed, analyses are then conducted using the completed dataset. This procedure has the advantage of replacing missing data with values whose distributions are like the non-missing ones. However, when missing values are only imputed once, no distinction is made between those values that were observed and those that were made up. For an imputation procedure to be valid, it must take into account the fact that imputed values are only a guess and not the value that would have been observed had there been no missing values. Single imputation ignores this uncertainty and overstates precision. We will later discuss multiple imputation[13] a process of creating two or more imputations for each missing value, which can lead to valid inferences under certain circumstances.

WHY DOES IT TAKE MORE EFFORT TO GET VALID ESTIMATES AND TESTS WHEN THERE ARE MISSING DATA?

None of the ad-hoc methods described above provide correct inferences. In order for a missing data procedure to provide valid inferences, it must meet a number of objectives. Relationships among variables must be preserved, nonresponse bias must be corrected, and uncertainty must be incorporated into the standard errors of parameter estimates. Achieving these objectives often requires the use of special statistical methods and careful thought about why values are missing.

A missing data procedure should preserve relationships among the data

In the context of longitudinal data, preserving relationships among the data often means taking into account the participant's trajectory before dropout or nonresponse. For example, if a participant has a very steep slope prior to dropout, it may not be realistic to assume that the

participant leveled off after dropping out of the study as is assumed with LOCF. Missing data procedures should condition on the observed data so that these trajectories are preserved.

A missing data procedure should adjust for nonresponse bias

When observations with missing values are systematically different from observations with observed values, bias can be introduced into parameter estimates. In a simple example, if all of the participants in a study who have an adverse outcome to the treatment drop out of the study, then the treatment effect will appear more favorable unless a proper adjustment takes place. Careful thought about why observations are missing is necessary to develop and implement missing data methods that can correct for nonresponse bias.

A missing data procedure should take into account uncertainty

No matter how well a procedure performs in preserving relationships among variables and adjusting for nonresponse bias, the fact remains that not all the values in the dataset are known. This uncertainty needs to be incorporated into the standard errors of parameter estimates so that their confidence intervals are not overly precise.

THE WECARE STUDY, A LONGITUDINAL DEPRESSION TREATMENT TRIAL

Description

The WECare Study investigated outcomes during a 12-month period in which low-income, mostly minority women in the suburban Washington, D.C., area were treated for depression. Participants were screened for depression at Women, Infant, and Children (WIC) clinics and various pediatric clinics. The study screened 16,286 women and eventually enrolled 267 women into the treatment portion of the study. The participants were randomly assigned to three groups: Medication, Cognitive Behavioral Therapy (CBT), and Treatment-as-usual (TAU), which consisted of a referral to a community provider.

Participants were interviewed by phone at baseline, every months for 6 months, and then every other month for the duration of the study. Major clinical outcomes were depression status, measured by the Hamilton Depression Rating Scale (HRSD), and functioning, measured by the Short Form 36-Item Health Survey (SF-36) and the Social Adjustment Scale (SAS). Depression was measured at every interview, and functioning was measured at baseline, and months 3, 6, and 12.

Outcomes for the first 6 months of the study have been previously reported.[14] In that article, the primary research question was whether the Medication and CBT treatment groups had better depression and functioning outcomes as compared to the treatment-as-usual group. To answer this question, the data were analyzed using an intent-to-treat random intercept and slope regression substitution model (see Hedeker and Gibbons[15] for an overview in this context). The outcomes reported were HRSD score, SF-36 Social Functioning, and SAS Instrumental Role Performance.

WECare Missing Data

Information on age, ethnicity, income, marital status, number of children, insurance, education, employment, and stressful life events was collected during the screening and the baseline interview. All screening and baseline data were complete except for income, with 10 participants missing data on income. After baseline, the percentage of missing interviews ranged between 16% and 40%. Treating month 6 as the last month of the study, 20% of medication subjects, 21% of CBT participants, and 22% of TAU participants dropped out of the study. Table 1 shows the mean HRSD score, the percentage of missing interviews, and the cumulative measurement dropout rate at each follow-up month. For example, at month 2, 16%

of the month 2 medication interviews were missing, and 7% of the medication participants had dropped out of the study.

THREE GENERALLY VALID APPROACHES FOR ANALYZING LONGITUDINAL DATA WITH MISSING VALUES

In this section, we describe three approaches for analyzing the HRSD scores from the WECare study. Each approach makes a different assumption regarding why data are missing. Although these three approaches for analyzing longitudinal data with missing values are not exhaustive, we hope that they will provide the reader with a flavor of how different assumptions regarding the missing data mechanism can lead to different analyses and different study conclusions. We also note that the WECare data, continuous repeated measures data, is only one type of data that are typically collected in longitudinal designs. Different data types (binary, ordinal, count, etc.) and different statistical models (time-to-event, single endpoint) will result in different approaches for handling missing data.

A MIXED-EFFECTS REGRESSION SUBSTITUTION MODEL

Mixed-effects regression models (MRMs)[16] provide a flexible framework for analyzing longitudinal data with missing values. Because MRMs make use of all observed values, they provide an efficient way to incorporate all available data for a subject. When the missing data is ignorable, and the model describing individual growth patterns is correctly specified, MRMs provide valid inferences in the presence of missing data.[17] For correct specification of the growth trajectory for each individual, the nonmissing data available for each individual must adequately represent that subject's trend over the course of the study.[11,15]

Miranda et al[14] used a MRM to analyze the first 6 months of the WECare data. Specifically, they fit a regression substitution model with a random intercept and random slope so that each subject's data were modeled as the sum of her own linear growth and random deviations from this line. Covariates in the model included baseline HDRS score, month, treatment, ethnicity (black, white, Latina), and a month by treatment interaction.

Multiple Imputation for Missing Data

As discussed earlier, imputation is the method of replacing missing values with plausible values. Single imputation, where each missing value is replaced with one made-up value, does not distinguish between truly observed values and imputed values. As a result, inferences tend to be overly precise because uncertainty caused by missing values is not taken into account. Rubin[13] proposed handling the uncertainty due to missing data through the use of multiple imputation. Multiple imputation refers to the procedure of replacing each missing value with two or more imputed values. Each set of imputed values generates a new complete dataset, each of which can be analyzed using complete-data methods. The final estimates are obtained by combining the results of the analyses on each of the imputed datasets using rules that combine within-imputation and between-imputation variability.[13] Inferences drawn in this manner properly reflect uncertainty due to nonresponse under that model. Multiple imputation provides excellent results 1) if MCAR holds, 2) if MAR holds and those variables that affect missing data are included in the model for imputation, or 3) if the missing data mechanism is NMAR but correctly modeled in the imputation procedure.[1,18]

One advantage of using multiple imputation to handle missing values is that the imputation model can incorporate a variety of variables to help the prediction. In fact, more variables are often used in the imputation model than in the final analytic model. As a result, the MAR assumption is more likely to be satisfied. For example, the MRM used to analyze the WECare data used ethnicity, baseline HDRS score, and treatment condition as predictors. In contrast,

the imputation model for the WECare data included income, age, education, marital status, and number of children, in addition to baseline HDRS score, ethnicity, treatment, month, and treatment by month interaction. Note that all variables that are used in the analytic model are also included in the imputation model, plus auxiliary variables. Collins et al[19] note that missing data procedures that make liberal use of auxiliary variables may result in noticeable gains in terms of increased efficiency and reduced bias.

The WECare data were multiply imputed using a Bayesian multivariate normal model.[20] With this approach, imputations for each variable are drawn from a normal distribution that conditions on all other variables included in the model. To incorporate treatment by month interactions, each treatment group was imputed separately. Twenty imputed values were imputed for each missing value to create 20 imputed datasets.[21] Each dataset was analyzed separately using the same MRM model described in the section on mixed-effects regression substitution models. Inferences across the 20 imputed datasets were combined using the rules described in Rubin.[13]

Pattern-mixture Model: A Non-ignorable Missing Data Method

Both MRMs and most forms of multiple imputation assume that the data are ignorable. That is, it is not necessary to take into account the missing data mechanism. However, this assumption may not always hold in psychiatric research where nonresponse is often related to a participant's mental state and not explained by the observed data. Pattern-mixture models[22-25] are non-ignorable missing data methods that stratify participants based on their missing data pattern. A separate model is fit for each pattern and then typically results are combined across the different patterns to obtain an average estimate of the model parameters. In this way, a model is fit for the joint distribution of the outcome and whether or not the outcome is missing.

In this example, we assume there are two missing data patterns in the WEC-are study: those participants that drop out of the study, and those that do not. The assumption here is that dropouts are potentially systematically different from participants who are observed at every time point or who only have intermittent missing data. See Hedeker and Gibbons[25] for a more in-depth exploration of missing data patterns in a longitudinal psychiatric setting.

The same MRM as before was fit, this time including a main effect indicator variable for whether the participant was a dropout or not. Also included were the two-way dropout by month and dropout by treatment interactions and the three-way dropout by month by treatment interaction. In this way, we can investigate the treatment effect for those participants who dropped out of the study and those who did not.

Results

Results from all three models (MRM, multiple imputation, pattern-mixture) are presented in the first three columns of Table 2. The covariates are coded so that coefficients for Medication and CBT provide contrasts against TAU at baseline (which should be nonsignificant based on randomization), and beneficial intervention effects are reflected by negative coefficients in the MED by Month and CBT by Month interaction coefficients. The parameter estimates are most similar for the MRM and the pattern-mixture model. Because only 21% of the participants in the WECare study were dropouts, we would not expect the pattern-mixture model to differ from the MRM model by very much. However, the intervention effects appear a bit stronger in the pattern mixture model than for MRM. Inspection of the data revealed that the medication and treatment effects for dropouts were indeed greater than that for completers. As a result there is a larger treatment effect (treatment by month interaction effect) under the pattern-mixture model as compared to the MRM.

The multiple imputation results were quite similar to the other results with one notable exception. As mentioned above, one advantage of multiple imputation is that the imputation model can incorporate a variety of variables to help the predictions. It was thought that the amount of treatment that a participant received during the study should be included in the imputation model because it was associated both with the probability that a value was missing and with HRSD score. For medication subjects, amount of treatment was measured using a variable indicating whether the subject received 9 weeks of medication therapy. For CBT subjects, amount of treatment received was measured by number of CBT sessions attended. For TAU subjects, it was the number of mental health visits to a community provider.

The results based on multiple imputation including amount of treatment received are displayed in column 4 of Table 2 (see page 799) and are similar to the other three models with the exception of the CBT treatment effect across time. When amount of treatment received is included in the imputation model, CBT is no longer significant, with its effect almost half the CBT effect based on the MRM. However, the effect size given under "MI-without amount of Tx" is also smaller than MRM estimates, indicating that at least part of the difference from the MRM result is due to MI, irrespective of the inclusion of dosage in the imputation model.

Those participants with missing values tended to attend fewer CBT therapy sessions. When amount of treatment received is not included in the imputation model, the effect of CBT is biased toward those who attended the therapy sessions who had greater improvement over time. By including number of CBT sessions in our multiple imputation model (a variable that is not typically included in an intent-to-treat analysis), we are able to preserve the relationship between number of treatment sessions and HRSD score even if it is not explicitly included in our analysis model. When the amount of treatment received is in the imputation model, the relationship between number of CBT sessions and HRDS score is preserved and leads to larger imputed HRSD scores in the CBT group and as a result, a non-significant CBT effect.

DISCUSSION

Missing data are ubiquitous in longitudinal psychiatric trials, and the failure to adequately handle missing data may result in invalid inferences. Currently, many researchers continue to use ad-hoc procedures, sometimes unknowingly, because complete-case analysis is the default procedure in many statistical software packages. Because of a rich statistical literature on handling missing data and a variety of software packages,[26] it is unnecessary for an investigator to rely on ad-hoc procedures that are inefficient at best and most likely produce biased parameter estimates.

In this article, we described why data are missing, why simple approaches for handling missing data do not work, and why it takes more effort to get valid estimates when there are missing data. By way of illustration, we analyzed data from a depression study using three procedures that typically lead to valid inferences. Most of the results of these different high-quality missing data procedures were consistent; however, the more complex multiple imputation model, which used ancillary information on program exposure for all three conditions, found less evidence of a CBT effect, because of worse trajectories for those who dropped out of this treatment.

There are many more approaches for handling missing data, and choice of what method to use will depend largely on what type of data one has (continuous, ordinal, binary, count) why the data are missing (MCAR, MAR, NMAR) and what study questions are trying to be answered. In most situations, the investigator will not know why data are missing and it is therefore useful to perform several analyses, each of which makes a different assumption regarding the missing data mechanism. In this way, the investigator can get an idea of the range of inferences due to different missing data assumptions.

Researchers and policy makers may be concerned about our example where different models for the missing data led to a different conclusion about an intervention's effect using intent-to-treat. Furthermore, the most complex analysis, which incorporated dosage as auxiliary information, was selected after examining the data rather than being specified in advance. This flexibility in accounting for missing data the way we have done here could be seen as a drawback for both journal editors and regulatory agencies, because it could open a door to applicants selecting the analysis giving the most significant effect. But the current practice of using ad-hoc procedures or methods that fail to fit the observed data are most assuredly coming to wrong conclusions. A strategy for removing much of the subjectivity is to have analysis protocols specify clearly in advance what auxiliary information is to be used in the imputation model, then determine which models best fit the data without regard to the any of the model parameter estimates dealing with intervention effects. This shielding of the step for choosing the best fitting model from the step of making inferences about the scientific questions of interest, preserves the Type I error rate.[27]

We conclude by recommending that when analyzing longitudinal data with missing values, the investigator first starts by thinking carefully about why the data are missing. Then, avoiding the temptation to apply the ad-hoc methods described in the section on ad-hoc approaches for handling missing data, fit statistical models that are consistent with the reason for missing data and answer the study hypotheses. Finally, as a sensitivity analysis, investigate other models that make different assumptions regarding the missing data mechanism. We summarize these recommendations in Table 3 (see page 800).

REFERENCES

1. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. Vol. 2nd ed.. Wiley; New York: 2002.

2. Rubin DB. Inference and missing data. Biometrika 1976;63:581–592.

3. Lavori PW, Brown CH, Duan N, Gibbons RD, Greenhouse JB. Missing Data. Psychiatric Annals 2008;38(12):784–792.Longitudinal Clinical Trials. Part A: Design and Conceptual Issues

4. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med 1991;10(4):585–598. [PubMed: 2057657]

5. Brown CH, Indurkhya A, Kellam SG. Power calculations for data missing by design with application to a follow-up study of exposure and attention. J Am Stat Assoc 2000;95:383–395.

6. 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):540–547. [PubMed: 17942469]

7. Verbeke, G.; Molenberghs, G. Mixed Models for Longitudinal Data. Springer; New York: 2000. Linear.

8. Simpson HB, Petkova E, Cheng J, Huppert J, Foa E, Liebowitz MR. Statistical choices can affect inferences about treatment efficacy: a case study from obsessive-compulsive disorder research. J Psychiatr Res 2008;42(8):631–638. [PubMed: 17892885]

9. Cook RJ, Zeng L, Yi GY. Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation. Biometrics 2004;60(3):820–828. [PubMed: 15339307]

10. Lane P. Handling drop-out in longitudinal clinical trials: a comparison of the LOCF and MMRM approaches. Pharm Stat 2008;7(2):93–106. [PubMed: 17351897]

11. Gibbons RD, Hedeker D, Elkin I, et al. Some conceptual and statistical issues in analysis of longitudinal psychiatric data. Application to the NIMH treatment of Depression Collaborative Research Program dataset. Arch Gen Psychiatry 1993;50(9):739–750. [PubMed: 8357299]

12. Lavori PW. Clinical trials in psychiatry: should protocol deviation censor patient data? Neuropsychopharmacology 1992;6(1):39–48. [PubMed: 1571068]

13. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. Wiley; New York: 1987.

14. Miranda J, Chung JY, Green BL, et al. Treating depression in predominantly low-income young minority women: a randomized controlled trial. JAMA 2003;290(1):57–65. [PubMed: 12837712]

15. Hedeker, D.; Gibbons, RD. Longitudinal Data Analysis. Wiley; New York: 2006.

16. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38(4):963–974. [PubMed: 7168798]

17. Laird NM. Missing data in longitudinal studies. Stat Med 1988;7(12):305–315. [PubMed: 3353609]

18. Rubin DB. Multiple Imputation After 18+ Years. J Am Stat Assoc 1996;91:473–489.

19. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001;6(4):330–351. [PubMed: 11778676]

20. Schafer, JL. Analysis of incomplete multivariate data. Chapman and Hall; London: 1997.

21. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci 2007;8(3):206–213. [PubMed: 17549635]

22. Little RJ, Wang Y. Pattern-mixture models for multivariate incomplete data with covariates. Biometrics 1996;52(1):98–111. [PubMed: 8934587]

23. Little RJA. A class of pattern-mixture models for normal incomplete data. Biometrika 1994;81(3): 471–483.

24. Little RJA. Modeling the drop-out mechanism in repeated-measures studies. J Am Stat Assoc 1995;90:1112–1121.

25. Hedeker D, Gibbons RD. Application of random-effects pattern-mixture models for missing data in longitudinal studies. Psychological Methods 1997;2:64–78.

26. Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Am Stat 2007;61(1):79–90. [PubMed: 17401454]

27. Brown CH, Wang W, Kellam SG, et al. Prevention Science and Methodology Group. Methods for testing theory and evaluating impact in randomized field trials: intent-to-treat analyses for integrating the perspectives of person, place, and time. Drug Alcohol Depend 2008;95(Suppl 1):S74–S104. [PubMed: 18215473]

TABLE 1

Mean HRSD Scores, Percent Missing, and Cumulative Measurement Dropout for Each Time-point

	Mean HRSD Score (% missing, % cumulative measurement dropout)		
Month of Study	Medication (n = 88)	CBT (n = 90)	Care as Usual (n = 89)
Baseline	17.95 (0%, 0%)	16.28 (0%, 0%)	16.48 (0%, 0%)
Month 1	14.00 (20%, 5%)	13.11 (27%, 1%)	12.80 (27%, 1%)
Month 2	10.74 (16%, 7%)	11.42 (27%, 3%)	11.30 (29%, 7%)
Month 3	9.60 (28%, 10%)	10.24 (36%, 7%)	13.05 (27%, 9%)
Month 4	9.54 (31%, 15%)	9.07 (38%, 10%)	11.81 (35%, 11%)
Month 5	8.62 (40%, 18%)	10.47 (34%, 13%)	11.85 (40%, 18%)
Month 6	9.17 (28%, 20%)	10.73 (33%, 21%)	11.92 (29%, 22%)

TABLE 2

Results of WECare Data Analysis

Parameter	MRM			MI without Amount of Tx Received in Imputation Model			Pattern Mixture Model			MI with Amount of Tx Received in Imputation Model		
	Est	SE	P val	Est	SE	P val	Est	SE	P val	Est	SE	P val
Intercept	13.62	0.91	<.001	13.79	0.91	<.001	13.48	0.87	<.001	13.64	0.90	<.001
Medication	-0.10	0.97	0.922	-0.25	1.01	0.805	0.27	1.00	0.790	-0.11	1.06	0.944
CBT	0.85	0.98	0.386	0.59	0.98	0.550	1.18	1.03	0.251	0.18	1.03	0.856
Month	-1.11	0.91	<.001	-1.14	0.44	0.010	-0.96	0.38	0.012	-1.06	0.42	0.066
MEDS x Month	-0.81	0.23	<.001	-0.70	0.23	0.003	-1.12	0.27	<.001	-0.75	0.24	0.012
CBT x Month	-0.60	0.23	0.009	-0.44	0.22	0.044	-0.88	0.30	0.004	-0.35	0.24	0.344

TABLE 3

Summary of Available Options for Handling Missing Data

Action	Pros	Cons	Recommendation
Careful deliberation about why data are missing	Helps determine correct model	Not always clear	Always do this
Last observation carried forward	Easy to do	Unrealistic assumption, overestimates precision	Never do this
Mean imputation	Easy to do, preserves mean	Does not preserve relationships in data. Overestimates precision.	Never do this
Complete-case analysis	Easy to do	Biased estimates unless data are MCAR. Loss of information.	Never do this.
End-point analysis	Missing values no longer an issue	Ignores information, ignores time	Never do this.
Single imputation	Reduces bias	Overestimates precision.	Use multiple imputation.
Mixed-effects regression model	Makes use of all available information	Can be complicated to fit. Assumes MAR.	Often a good choice
Multiple imputation	Allows one to incorporate auxiliary variables into imputation model	Requires expertise. Additional steps for analyzing data.	Do it if MAR assumption is likely to be satisfied
Nonignorable models	Explore the effect of different missing data assumptions	Not clear what is correct model. Can be complicated.	Worth doing, especially as a sensitivity analysis
Fit several different statistical models that make different assumptions regarding why data are missing	Sensitivity of inferences to different assumptions regarding the missing data mechanism.	Additional work. May complicate the overall picture.	Worth doing.