Appendix

B

Imputation Procedures for Missing Data in Clinical Research

Overview

The *MATRICS Consensus Cognitive Battery* (MCCB), building on the foundation of the Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) framework, aims to provide a standardized set of data upon which to make decisions about the efficacy of cognition-enhancing interventions for schizophrenia and related disorders. The tests of the MCCB were selected, in part, because their administrative time is relatively brief and they were perceived to be well tolerated by study participants. For these reasons, study participants are expected to complete the entire battery the vast majority of the time. However, in clinical trials, despite the best efforts of investigators to obtain answers to all questions asked of study participants, individual items occasionally go unanswered, giving rise to "missing data" (Little & Rubin, 2002).

Failure to properly account for missing data in analyses can introduce substantial bias in the estimation of treatment effects. Imputation, which refers to a class of strategies for filling in missing data with plausible values, has become the standard approach for handling missing data and can provide a valid basis for statistical inference (Rubin, 1987; Little & Rubin, 2002). An imputation strategy based on an additive model procedure (as described in Little & Rubin, 2002, pp. 70-71) was initially recommended for use with the MCCB. The additive approach allows for the possibility that certain individuals, tests, treatment groups or measurement occasions might have consistently higher (or lower) than average scores, while avoiding the biases that can arise with overly simplistic strategies such as "person-mean imputation" (i.e., filling in missing values for a particular person with the average of other observed values for that person) or "item-mean imputation" (i.e., filling in missing values of a particular test item with the average of the observed values for that item from other people). However, recent experience with the MCCB as well as new developments in the missing data literature and insights from the 2010 report of the National Research Council's Panel on Handling Missing Data in Clinical Trials (National Research Council, 2010; O'Neill & Temple, 2012) suggest refinements to the original procedure. Specifically, as described in Chapter 6, we now recommend that investigators using the MCCB for clinical trials employ sequential regression multiple imputation (Raghunathan et al., 2001) for handling missing data. This approach has several advantages: it is readily available in standard software packages; easily accommodates covariates to maximize imputation quality; can produce either single or multiple imputations; and can be integrated into any level of the analyses. The procedures recommended below were developed by a MATRICS subcommittee that consisted of Thomas R. Belin, PhD, and Catherine A. Sugar, PhD, of the UCLA Department of Biostatistics and Michael F. Green, PhD, Robert S. Kern, PhD, and Keith Nuechterlein, PhD, of the UCLA Department of Psychiatry and Biobehavioral Sciences. This approach has been endorsed by the MATRICS Neurocognition Committee.

Selection of Values and Variables to Include in Imputation

For simplicity, the original additive imputation procedure used only the measures from the MATRICS battery. However, it is now generally accepted that as much covariate information as possible should be included in imputation procedures (Rubin, 1996; Collins, Schafer, & Kam, 2001; Schafer & Graham, 2002). Ideally, one would include all available measures that might be related to the missing variable to maximize the accuracy and minimize the bias of the imputed scores. The list of measures to be included in the imputation models should be pre-specified as part of the study design and analysis plan and agreed upon with the sponsoring body. As a minimum standardized set, we recommend including age and gender in the imputation models, in addition to the core MCCB test scores, as these are known to be related to neurocognitive performance and will be available in all clinical trials. We note that including these covariates in the imputation procedure is informative even though they are adjusted for in the MCCB scoring program. This is because age and gender may be related to the likelihood of missingness as well as to actual performance and the goal of the imputation procedure is accurate prediction rather than covariate adjustment. In most circumstances, using the raw MCCB test scores in the imputation will yield adequate performance. However, if the analytical plan calls for using transformed versions of the individual measures (e.g., a logarithmic transformation for the *Trail Making Test* time score), then that same transformation should be used in the imputation procedure. (See the technical specifications section below for details.)

Related to the issue of covariate adjustment, it is important to account for treatment group and time in study when performing imputation. Failure to do so could result in significant biases if there are longitudinal trends or treatment effects. We therefore recommend that imputation be done separately for each treatment group at each major study time-point. This simplifies the actual imputation models (avoiding the need for repeated measures or interaction terms) while minimizing bias. It also allows imputations to be performed for interim analyses without breaking the blind since actual group labels would not be needed, nor would group or time effects be included in the output from the imputation models. We note that using the time and treatment group assignments in the imputation does not bias the results in favor of treatment effects; in contrast, failure to include them typically biases the results against treatment effects. If there are reasons (particularly in an interim analysis) why it is not possible to do the recommended stratification of the imputation procedure, the results will in general be conservative. For international studies, we similarly recommend that imputation be done separately by country as long as the resulting subgroups are sufficiently large ($n \geq 30$).

The more observations that are included in the imputation model, the more stable and accurate the imputed values will be. The final imputations should therefore be performed once all assessments are finished and the analysis data set is cleaned and locked, not intermittently as the data are collected. If imputations are performed for interim analyses, they should be redone at the end of the study before the final analyses are performed. Indeed, imputation is fundamentally part of the analytical process rather than part of data collection. It is designed to produce the best possible (e.g., unbiased, maximum likelihood) estimates of parameters of interest, including treatment effects, based on the existing data. Although imputed values should be preserved to allow replication of analyses, they should not be entered into the clinical database as if they were original observations. Different studies may have different amounts and patterns of missing data and may therefore differ in the optimal approach to imputation. At a minimum, it is important that all studies report the amount of and likely reasons for missing data.

If scores from too many tests are missing at a given assessment, it may not be possible to impute values meaningfully. We specifically recommend that values for at least two-thirds of the cognitive domains (i.e., a minimum of 5 of 7) must be available at baseline for it to be counted as a test occasion. For follow-up assessments, at least half of the domains need to be successfully assessed (minimum of 4). We also note that for domains that involve more than one test (Speed of Processing and Working Memory), the *MCCB Computer Scoring Program* automatically computes domain scores based on the available data as long as at least half of the tests were successfully administered. It is therefore unnecessary to perform external imputation if the only missing test scores occur in domains that are adequately represented.

In clinical trials research with new pharmaceutical agents, it is not unusual for some participants to miss entire assessment points, as opposed to lacking data only for certain tests. There are a variety of ways to handle missing assessments, including last observation carried forward and mixed effects (repeated measures) models. In large clinical trials, decisions about the best methods to use in these situations are often the result of discussions between the drug manufacturer and the FDA, so no specific recommendations are made here for instances in which there are entire assessment points missing.

An Updated Framework Based on Sequential Regression Multiple Imputation

In recent years, much research on missing data has centered on the idea of what has been termed “sequential regression multiple imputation” (Raghunathan et al., 2001). Implementations are available in many widely used standard software packages (e.g., SAS [IVEware add-on module], STATA [ice/mi impute chained], SPSS [mi, fully conditional specification] and R [mice]). In this approach, missing values for a particular variable are imputed by regressing it on the other variables in the imputation set. The procedure is iterated sequentially for each variable (here MCCB test scores and covariates) in turn until convergence. The algorithm is initiated using a simple imputation method (e.g., subject or variable mean imputation) to fill in starting values for the missing points.

We recommend the following multi-stage procedure for imputation with the MCCB:

1. Enter the available data into the MCCB scoring program.
2. Export the raw test scores.
3. Run the above sequential procedure to obtain multiple imputations for the missing test scores including age, gender, and other covariates as pre-specified.

4. If any of the imputed scores are outside the valid range for that test, set the imputed value to the minimum or maximum possible score as appropriate.
5. Enter each of the imputed raw test scores into the MCCB scoring program to calculate the composite scores.
6. Run the primary analyses on each of the resulting imputed data sets and combine to obtain the final results.

Note that while it is theoretically possible to perform sequential regression multiple imputation at the *T*-score level, we specifically recommend imputing the raw test scores and then calculating the domain and composite scores using the MCCB scoring program to guarantee consistency of the various components. In particular, if imputation is done at the *T*-score level, it is possible to obtain values that are inconsistent with the *T*-scores corresponding to possible raw test scores; this is much more difficult to detect than an out of range value on the raw test scale. We also recommend the multiple imputation procedure because it correctly accounts for the uncertainty in the missing values as part of the final analysis. It has been well established in the statistics literature (e.g., Rubin 1987) that treating single imputed values as equivalent to observed values in subsequent analyses can substantially understate uncertainty as compared with multiple-imputation procedures where variability in target quantities of interest can be estimated by considering multiple plausible values for each missing item. Specifically, in the multiple imputation setting, the desired analysis is run separately for each imputed data set and the results are combined using standard formulas that adjust the standard errors of the parameter estimates to account for the variation from one analysis to the next. (See Little & Rubin, 2002, for details. Algorithms for combining the individual analysis results are available in all standard statistical packages.)

The procedures listed above currently are considered the optimal approach for performing imputation in large-scale clinical trials using the MCCB. However, we recognize that researchers in some settings will have small data sets or minimal numbers of missing values which may make the sequential regression approach impractical or unnecessary. There is a wide range of available techniques for imputation depending on the amount and pattern of missingness. Methods such as the additive model originally proposed for the MCCB provide a good balance between ease of use and rigor. (See Little & Rubin, 2002, for details of the additive model approach (pages 70-71) and for a general review.)

Sensitivity Analyses

The National Research Council (2010) placed particular emphasis on considering sensitivity of imputations to modeling assumptions in large-scale clinical trials. Sensitivity analyses can be extremely valuable for assessing the effects of missing data and the corresponding choice of imputation procedures. A straightforward paradigm for a sensitivity analysis is to run the primary models using:

1. study participants with complete data
2. a data set with the optimal values filled in for all study participants
3. a data set with the worst values filled in for all study participants

4. a data set with the best values filled in for controls and the worst values filled in for the treatment group, and
5. the data sets obtained using the sequential imputation procedures recommended in this manual for assessment occasions on which partial testing was completed (combining the results using the standard multiple imputation algorithms).

Technical Specifications for Performing Sequential Multiple Imputation

A number of operational choices must be made when performing sequential multiple imputation. Below we provide more detailed recommendations for the most common technical issues.

1. Variable Types: Because sequential imputation procedures treat each of the variables in the imputation set in turn as the outcome in a generalized linear model, it is necessary to specify the type (e.g., continuous, categorical, count, mixed) for each variable so that the appropriate model form (e.g., linear regression, logistic regression, Poisson regression, zero-inflated model) will be used. The MCCB raw test scores should all be treated as continuous. The classification of additional covariates will depend on how they are measured in individual studies.
2. Range Restrictions: The MCCB tests have minimum and maximum possible raw scores that the imputation procedure must respect if the resulting values are to be entered into the MCCB scoring program. Most packages that perform sequential imputation allow the user to specify those bounds and then automatically truncate the values, either by setting all imputations outside the range to the boundary values or by taking draws from a distribution which has been smoothed at the edges. The built-in procedures are generally appropriate, but the users who wish to have complete control of the boundary cases can run the imputation in unrestricted mode and truncate the values themselves.
3. Random Seeds: Sequential multiple imputation procedures involve random draws from appropriate posterior distributions specifying the relationships among the variables of interest. In order to be able to reproduce the imputed data sets, it is important to select and record a fixed random seed which will be used as the starting point for all imputations for the trial. (Each clinical trial should use its own random seed.)
4. Number of Imputed Data Sets: The standard recommendation for the number of imputed data sets is five and this is usually sufficient to achieve good estimates of between imputation variance (the quantity used to adjust the standard errors of the parameter estimates for the uncertainty in the imputed values). However, current computational speed and memory capacity make generating and storing 10, 20 or even 100 imputed data sets and obtaining the combined analysis estimates perfectly feasible, and in some cases this provides additional accuracy.
5. Number of Iterations: Sequential imputation procedures cycle through each of the variables in the imputation set in turn. Manuals for major software packages, such as IVEware, suggest that 10 iterative cycles are sufficient for most imputations. However, as with the number of imputations, there is little cost to running additional cycles.

6. Generating Model Coefficients and Predicted Values (Perturbations): In general, sequential imputation procedures will perturb model coefficients using a multivatiate normal approximation of their posterior distribution and generate the predicted values using the regression model for the current variable based on those coefficients. This is sufficient in most cases. However, there are situations in which the multivariate normal approximation for the posterior distribution of the coefficients is inappropriate. In these cases, a sampling-importance-resampling algorithm can be used.

7. Number of Predictors: As noted above, ideally one would include all available measures that might be related to the missing variable to maximize the accuracy and minimize the bias of the imputed scores. However, in some cases the number of available observations may be small relative to the number of variables in the imputation set, especially if (as recommended above) the imputations are done separately by study arm, time point, and (if applicable) country/language group. We recommend having a minimum of approximately three observations per variable used in the imputation models (which would include the individual MCCB test scores, age, gender, and any additional study-specific covariates; note that not all of these variables need have missing values). Many sequential imputation packages allow use of a stepwise procedure to reduce the number of predictors in each model to an optimal subset of a given size. One can also specify a set of predictors to use for each variable with missing data based on theoretical relationships or empirical correlations.

8. Selection of Additional Imputation Variables: In any individual study there may be additional covariates which are contextually of interest or are known to be related to the MCCB test scores in the study population. For instance, some of the MCCB tests show differences by race or ethnicity which may or may not be relevant depending on the study sample. Such variables should be included in the imputation set whether or not they have missing values. In addition, if interaction terms or similar constructed variables will be used in the final analyses, these terms should be included in the imputation set so that the proper joint relationships among the model variables are respected. (Note that if the imputation is stratified by treatment arm and time, then interactions involving these variables are implicitly already accounted for.) Finally, it is theoretically possible and perhaps even valuable given within subject correlations to use a participant's values for a particular MCCB test at other time points to impute a missing value of that test (lag variables). However, this procedure would introduce considerable complexities into the modeling and is difficult to standardize across trials. We therefore do not recommend inclusion of lag variables as part of the base imputation set, although they could be discussed during the design phase if the planned spacing of observations was tight and autocorrelation was expected to be high.

9. Transformations: Some of the MCCB raw test scores, such as those for the *Trail Making Test*, are known to have skewed distributions and are often transformed when these variables are analyzed individually. In general, the imputation procedures suggested here will be fairly robust to non-normality. Moreover, the MCCB scoring program has built-in transformations for creating the derived *T*-scores. Thus, carrying out transformations before performing an imputation will not usually be necessary unless use of the individual raw scores in the final analyses is planned. If such analyses are planned, then the transformation that will be used in the final models should be used in the imputation. (Note that

in this case it will be necessary to transform the imputed values back to the original scale before entering them in the MCCB scoring program to obtain *T*-scores.) Similarly, some of the scores from MCCB tests have shown curvilinear relationships with age at the extreme ends of the range. However, in the development of the MCCB scoring program, it was found to be sufficient to use linear age in the regression models used to generate the *T*-scores. It is therefore not necessary to include quadratic or other curvilinear age terms in the imputation procedures.

These guidelines should cover the technical specifications necessary to successfully implement the recommended sequential imputation procedures for most standard clinical trials. However, study specific issues can arise that would affect the optimal choice of imputation procedure, and these should be carefully considered and discussed with the sponsoring agency prior to commencing the trial.

References

Collins, L.M., Schafer, J.L., & Kam, C.M. (2001). A comparison of inclusive and restrictive missing-data strategies in modern missing-data procedures. *Psychological Methods*, 6, 330–351.

Little, R.J.A., & Rubin, D.B. (2002). *Statistical Analysis with Missing Data, 2nd edition*. New York: John Wiley & Sons.

National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. Panel on Handling Missing Data in Clinical Trials, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

O'Neill, R.T., & Temple, R. (2012) The prevention and treatment of missing data in clinical trials: An FDA perspective on the importance of dealing with it. *Clinical Pharmacology & Therapeutics*, 91(3), 550–554.

Raghunathan, T.E., Lepkowski, J.M, van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 85–95.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Rubin, D.B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91, 473–489.

Schafer, J.L., & Graham, J.W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7, 147–177.

Siddique, J., & Harel, O. (2009). MIDAS: A SAS macro for multiple-imputation using distance-aided selection of donors. *Journal of Statistical Software*, 29, 1–18.