International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 20(1): 40–49 (2011)
Published online in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.329

Multiple imputation by chained equations: what is it and how does it work?

MELISSA J. AZUR,[1] ELIZABETH A. STUART,[1] CONSTANTINE FRANGAKIS[2] & PHILIP J. LEAF[1]

1 Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA
2 Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MA, USA

Key words
missing data, multiple imputation, analyze

Correspondence
Melissa Azur, Mathematica Policy Research, 600 Maryland Av. SW, Suite 550, Washington D.C., 20024, USA.
Telephone (+1) 202 250 3518
Fax (+1) 202 863 1763
Email:
mazur@mathematica-mpr.com

Received 16 February 2010;
revised 26 July 2010;
accepted 13 October 2010

Abstract

Multivariate imputation by chained equations (MICE) has emerged as a principled method of dealing with missing data. Despite properties that make MICE particularly useful for large imputation procedures and advances in software development that now make it accessible to many researchers, many psychiatric researchers have not been trained in these methods and few practical resources exist to guide researchers in the implementation of this technique. This paper provides an introduction to the MICE method with a focus on practical aspects and challenges in using this method. A brief review of software programs available to implement MICE and then analyze multiply imputed data is also provided. *Copyright © 2011 John Wiley & Sons, Ltd.*

Introduction

Missing data are a common problem in psychiatric research. Multivariate imputation by chained equations (MICE), sometimes called "fully conditional specification" or "sequential regression multiple imputation" has emerged in the statistical literature as one principled method of addressing missing data. Creating multiple imputations, as opposed to single imputations, accounts for the statistical uncertainty in the imputations. In addition, the chained equations approach is very flexible and can handle variables of varying types (e.g. continuous or binary) as well as complexities such as bounds or survey skip patterns. However, despite these benefits, many psychiatric researchers have not yet learned about this approach, there are few practical resources available to assist in its implementation and until recently software limitations inhibited general researchers and practitioners from using the MICE procedure. This paper provides an introduction to implementing MICE with a focus on the practical aspects. A brief review of software available to implement MICE procedures and to analyze data that has been imputed using these procedures is also provided. Readers interested in learning about other methods of addressing missing data can refer to Graham (2009), Lee and Carlin (2010), or Schafer (1999).

There are many different approaches to addressing missing data and the first question researchers might ask is

"why use multiple imputation?" In certain circumstances (e.g. when there is less than 5% missingness and the missingness is totally random and does not depend on observed or unobserved values), complete case analysis may be an acceptable approach to addressing missing data (Graham, 2009; Schafer, 1999). In practice, these circumstances rarely occur. While complete case analysis may be easy to implement it relies upon stronger missing data assumptions than multiple imputation and it can result in biased estimates and a reduction in power (Graham, 2009). Single imputation procedures, such as mean imputation, are an improvement but do not account for the uncertainty in the imputations; once the imputation is completed, analyses proceed as if the imputed values were the known, true values rather than imputed. This will lead to overly precise results and the potential for incorrect conclusions. Maximum likelihood methods are sometimes a viable approach for dealing with missing data (Graham, 2009); however, these methods are primarily available only for certain types of models, such as longitudinal or structural equation models, and can generally be run only using special software such as Amos (SPSS, 2009a) and Lisrel (Scientific Software International, 2006).

Multiple imputation has a number of advantages over these other missing data approaches. Multiple imputation involves filling in the missing values multiple times, creating multiple "complete" datasets. Described in detail by Schafer and Graham (2002), the missing values are imputed based on the observed values for a given individual and the relations observed in the data for other participants, assuming the observed variables are included in the imputation model. Multiple imputation procedures, particularly MICE, are very flexible and can be used in a broad range of settings. Because multiple imputation involves creating multiple predictions for each missing value, the analyses of multiply imputed data take into account the uncertainty in the imputations and yield accurate standard errors. On a simple level, if there is not much information in the observed data (used in the imputation model) regarding the missing values, the imputations will be very variable, leading to high standard errors in the analyses. In contrast, if the observed data are highly predictive of the missing values the imputations will be more consistent across imputations, resulting in smaller, but still accurate, standard errors (Greenland and Finkle, 1995).

The chained equation approach to multiple imputation

MICE is a particular multiple imputation technique (Raghunathan *et al.*, 2001; Van Buuren, 2007). MICE operates under the assumption that given the variables used in the imputation procedure, the missing data are Missing At Random (MAR), which means that the probability that a value is missing depends only on observed values and not on unobserved values (Schafer and Graham, 2002). In other words, after controlling for all of the available data (i.e. the variables included in the imputation model) "any remaining missingness is completely random" (Graham, 2009). Implementing MICE when data are not MAR could result in biased estimates. In the remainder of this paper, we assume that the MICE procedures are used with data that are MAR. Readers interested in general missing data issues, such as data that may not be MAR, should refer to Graham (2009) or Schafer and Graham (2002).

Many of the initially developed multiple imputation procedures assumed a large joint model for all of the variables, such as a joint normal distribution. In large datasets, with hundreds of variables of varying types, this is rarely appropriate. MICE is an alternative, flexible approach to these joint models. In fact, MICE approaches have been used in datasets with thousands of observations and hundreds (e.g. 400) of variables (He *et al.*, 2009; Stuart *et al.*, 2009). In the MICE procedure a series of regression models are run whereby each variable with missing data is modeled conditional upon the other variables in the data. This means that each variable can be modeled according to its distribution, with, for example, binary variables modeled using logistic regression and continuous variables modeled using linear regression. (Software packages do vary somewhat in their implementation of MICE, with some packages also using a multinomial logit model for categorical variables and a Poisson model for count variables.)

MICE steps

The chained equation process can be broken down into four general steps:

Step 1: A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."

Step 2: The "place holder" mean imputations for one variable ("var") are set back to missing.

Step 3: The observed values from the variable "var" in Step 2 are regressed on the other variables in the imputation model, which may or may not consist of all of the variables in the dataset. In other words, "var" is the dependent variable in a regression model and all the other variables are

independent variables in the regression model. These regression models operate under the same assumptions that one would make when performing linear, logistic, or Poison regression models outside of the context of imputing missing data.

Step 4: The missing values for "var" are then replaced with predictions (imputations) from the regression model. When "var" is subsequently used as an independent variable in the regression models for other variables, both the observed and these imputed values will be used.

Step 5: Steps 2–4 are then repeated for each variable that has missing data. The cycling through each of the variables constitutes one iteration or "cycle." At the end of one cycle all of the missing values have been replaced with predictions from regressions that reflect the relationships observed in the data.

Step 6: Steps 2–4 are repeated for a number of cycles, with the imputations being updated at each cycle.

The number of cycles to be performed can be specified by the researcher. At the end of these cycles the final imputations are retained, resulting in one imputed dataset. Generally, 10 cycles are performed (Raghunathan *et al.*, 2002); however, research is needed to identify the optimal number of cycles when imputing data under different conditions. The idea is that by the end of the cycles the distribution of the parameters governing the imputations (e.g. the coefficients in the regression models) should have converged in the sense of becoming stable. This will, for example, avoid dependence on the order in which the variables are imputed. In practice, researchers can check the convergence by, for example, comparing the regression models at subsequent cycles, as discussed in He *et al.* (2009). Different MICE software packages vary somewhat in their exact implementation of this algorithm (e.g. in the order in which the variables are imputed), but the general strategy is the same.

To make the chained equation approach more concrete, imagine a simple example where we have three variables in our dataset: age, income, and gender, and all three have at least some missing values. The MAR assumption would imply that the probability of a particular variable being missing depends only on the observed values, and that, for example, whether someone's income is missing does not depend on their (unobserved) income. In Step 1 of the MICE process, each variable would first be imputed using, e.g. mean imputation, temporarily setting any missing value equal to the mean observed value for that variable. Then in Step 2 the imputed mean values of age would be set back to missing. In Step 3, a linear regression of age predicted by income and gender would be run using all cases where age was observed. In Step 4, predictions of the missing age values would be obtained from that regression equation and imputed. At this point, age does not have any missingness. Steps 2–4 would then be repeated for the income variable. The originally missing values of income would be set back to missing and a linear regression of income predicted by age and gender would be run using all cases with income observed; imputations (predictions) would be obtained from that regression equation for the missing income values. Then, Steps 2–4 would again be repeated for the variable gender. The originally missing values of gender would be set back to missing and a logistic regression of gender on age and income would be run using all cases with gender observed; predictions from that logistic regression model would be used to impute the missing gender values. This entire process of iterating through the three variables would be repeated until convergence; the observed data and the final set of imputed values would then constitute one "complete" data set.

The number of imputed datasets to create

Once the designated number of cycles has been completed, the entire imputation process is repeated to generate multiple imputed datasets. The observed data of course will be the same across the imputed datasets; only the values that had originally been missing will differ. Initial research indicated that 5–10 imputed datasets was sufficient; however, recent research suggests that, depending upon the amount of missing information in the data, increasing that to as many as 40 imputed datasets can improve power (Graham *et al.*, 2007). However, in practice, imputing a large number (e.g. 40) of datasets may not be feasible. Depending on the size of the imputation model and the available computer resources, imputing a single dataset can take minutes or hours. Graham *et al.* (2007) provide simulation results that can be used to guide decision-making regarding the number of imputed datasets to generate. In particular, the size of the dataset, the amount of missing information in the data, and computational resources can help researchers determine how many imputed datasets to generate. For example, creating a single imputed dataset that has hundreds of variables, thousands of cases, and missingness ranging from less than 5% to 80% could take hours to run and therefore, it may be impractical to create 40 imputed datasets. Conversely, creating a single imputed dataset with 20 variables and hundreds of cases could conceivably be run in minutes

and therefore creating 40 imputed datasets would be quite feasible.

Setting up a MICE procedure

When setting up a MICE procedure, one of the first tasks researchers face is to determine which variables to include in the imputation process. This will generally include all variables that will be used in subsequent analyses (whether or not they have missing data), as well as variables (again, whether or not they have missing data) that may be predictive of the missing values. One key point is to include the variables that are likely to satisfy the MAR assumption. Beyond that, there are three specific issues that often come up when selecting variables: (1) creating an imputation model that is more general than the analysis model, (2) imputing variables at the item level versus the summary level, and (3) imputing variables that reflect raw scores versus standardized scores.

There are two aspects to having an imputation model that is more general than the analysis model. As mentioned earlier, all relationships that are going to be investigated in the analysis need to be included in the imputation model. This includes the dependent variable(s) in the research question(s) of interest (Moons *et al.*, 2006) and any potential interactions that will be tested. For example, if the relationship between the interaction of gender and depressive symptoms on inpatient service utilization is of interest, then inpatient service use and the interaction between gender and depressive symptoms (as well as gender and depressive symptoms themselves) should be included in the imputation regression models. If these variables are excluded from the imputation model, when analyses are conducted analysts may not find relationships that actually exist, because the imputations were generated assuming those variables were independent (Graham, 2009). In datasets with many variables that will be analyzed by a number of users, it may not be feasible to include all potential interactions in the imputation model. If this is the case, it is helpful to engage the potential users in a discussion of the interactions likely to be included in analyses and then to include as many of these interactions as possible.

The second aspect of having an imputation model that is more general then the analysis model is including additional ("auxiliary") variables in the imputation process – variables that are not going to be used in the analysis but that can improve the imputations (Collins *et al.*, 2001; Schafer, 2003). As mentioned earlier, MICE procedures assume that the data are MAR. While it is almost always impossible to test this assumption, including auxiliary variables (in the imputation regression model) that are predictive of missingness as well as variables that are correlated with variables that will be used in the data analysis stage, can reduce bias and make the MAR assumption more plausible (Collins *et al.*, 2001; Schafer, 2003). In fact the ability to incorporate additional information through auxiliary variables is a benefit of multiple imputation procedures as compared to standard maximum likelihood methods for handling missing data, which utilize only the variables in the analysis model. Graham (2003) does describe more complex maximum likelihood analyses that can incorporate auxiliary variables but those methods are relatively rare.

An additional consideration when determining which variables to include in the imputation procedure is whether to include the individual items that make up an instrument/scale, or to include a summary measure of the entire scale. In making this decision, it may be helpful to examine the missingness at both the item and summary level. If there is very little item missingness within each scale (i.e. if some items for a scale are observed then they are all observed) and analysts will use only the summary scales, then imputing the summary scales may make sense. Graham (2009) suggests that creating and imputing a scale score is appropriate when at least half of the items are observed, the items have high coefficient alphas, and all of the item-total correlations are similar. In contrast, if these conditions are not met it may make sense to impute the items themselves and then construct the summary scales using the observed and imputed data. This strategy will prevent the loss of observed information on study subjects who responded to some but not all of the items within a scale. This issue is also relevant more generally for variables that are constructed from the observed data, for example calculating the total number of family members from two variables, one giving the number of adults and the other the number of children. In general, when there are relatively few variables used to construct the resulting variable it is more appropriate to impute the original variables and then re-construct the new variable after the imputations are created.

Once a decision has been made to impute items or summary scores, a similar decision may be necessary regarding the imputation of raw or standardized scores. The distribution of these variables may help guide this decision. For example, if the raw scores of a continuous measure such as internalizing behavior problems are more normally distributed than the corresponding standardized scores then researchers may want to use the raw scores in the imputation model, because the raw scores will likely better meet the assumptions of the linear regressions

being used in the imputation process. The availability of syntax to recreate standard scores may also influence decisions as to which type of variable to impute.

Model specifications

To further enhance the imputation model and the creation of valid imputations, bounds and restrictions are useful specifications to impose upon some variables, and these are very easy to specify in some MICE software packages. Bounds provide a range of possible imputed values and are useful when imputing scales that have minimum and maximum values. Restrictions identify conditions under which a variable should or should not be imputed. This could occur if there is missing by design in the data or if there are hierarchical questions (also called skip patterns) in an instrument. For example, there may be instruments that are only administered to (and relevant for) a sub-population of the sample, such as youth report of substance use being asked only of adolescents and not of younger children. Restrictions ensure that missing values on the substance use items would not be imputed for young children. Hierarchical questions are a series of questions on an instrument whereby participants are asked follow-up questions only if the primary question is endorsed. Missing values on the follow-up questions should be imputed only if a positive response is recorded on the primary question. For example, a variable such as "age of first use of heroin" would only be imputed if the previous question "have you ever used heroin?" received a positive response.

Large datasets naturally lead to the possibility of a very large number of variables to include in the imputation regression models, and it may not always be possible to include all of those variables identified for potential inclusion. Selection of which variables to include can be guided by the types of analyses anticipated for use with the data and an awareness that generously incorporating auxiliary variables in the imputation models poses little risk to the precision or bias of estimates (Collins *et al.*, 2001). However, since using hundreds of predictors in each of the regression models is impractical, it may be necessary to have a process for selecting which variables to include in each of the regression models that constitute the MICE procedure. Stepwise regression is one method of identifying variables for inclusion in the individual regression models. With this approach, Steps 1 and 2 of the MICE procedure remain the same: all missing values in the dataset are temporarily replaced with the mean observed values. Those values are then changed back to missing for the first variable that will be imputed. In Step 3, the variable "var" will be regressed upon other variables in the dataset. When stepwise regression procedures are used, a large set of variables are "sent" to the imputation procedure, but not all are used as independent variables in each of the regression models. Stepwise regression chooses the variables that are the most important predictors in each model based upon criteria specified by the imputer, such as indicating the maximum number of variables to include in each regression equation and/or specifying the minimum marginal *r*-squared. A large minimum marginal *r*-squared (e.g. 0.01) leads to fewer variables selected as predictors and a small minimum marginal *r*-squared (e.g. 0.001) leads to more variables selected. In order to check the sensitivity of the imputations, the imputation model can be re-run with multiple *r*-squared values. He *et al.* (2009) provide an example where an *r*-squared of 0.1 led to only one or two predictors in each model, while an *r*-squared of 0.001 led to instability in the regression coefficients because of too many predictors. They chose to use a "compromise" *r*-squared value of 0.01.

Assessing the imputation procedure

Once the imputation model has been specified and the initial imputations created, it is important to check the model and refine as necessary. This can be done in a number of ways. When a stepwise procedure has been used, an examination of the regression models is helpful to gain a sense of the type and quantity of variables selected as predictors. Summary statistics that provide information on the observed values, the imputed values, and the combined values for each variable are also useful for identifying problematic variables and variables that need modification before their use in analyses (e.g. He *et al.*, 2009). These basic statistics are provided by many software programs that perform MICE, but can also be generated by the researcher, if needed. Of course the basic diagnostics we describe here may miss problems. The development of better diagnostics for multiple imputation is a topic of ongoing statistical research (see, for an example, He *et al.*, 2009); more advanced diagnostic methods are not yet fully developed or easily implementable.

To illustrate the type of information imputers may want to examine when reviewing such summary statistics, two variables from a children's mental health services dataset (Manteuffel *et al.*, 2002) are used. These variables were imputed as part of a larger project whereby the MICE procedures were implemented on a dataset with about 9000 cases and 400 variables (Stuart *et al.*, 2009). Stepwise procedures, as described earlier, were used to select the predictors in each imputation regression model.

In the first example presented here, the binary variable *outpatient service use* was imputed (Table 1). The summary statistics show the distributions among the observed and imputed responses separately. Only 649 cases had missing values and thus needed imputation. The distributions for both the imputed and combined values seem reasonable.

The second example is of a variable capturing youth report of the number of times the individual had five or more alcoholic beverages in a row (Table 2). This question was only asked of youth ages 11 years and older who had endorsed a previous question about ever drinking alcohol and thus the imputations were restricted to this group. It was classified as a count variable in the imputation procedure and was modeled using a Poisson distribution. There are two things to note in this example. At first glance it appears as if 8475 cases were imputed. In actuality, many of these values are place holders for missing by design, in this case, youth less than 11 years or youth who did not endorse the previous question and therefore should not have values imputed. These place holders were automatically inserted by the software package used (IVEware, for more details see later) and reflect cases that did not receive valid imputed values. Next, the summary statistics indicate that the imputed values for some cases are unreasonably large (e.g. 4.50×10^{15}). This large value indicates that a problem exists somewhere. When this type of problem occurs, an examination of the regression model and an investigation into which cases received the extreme imputed values may provide insight into the magnitude and potential causes of the extreme values. For this variable, only a few cases received unusual imputed values and it is suspected that the values are a result of the fact that this variable had very little observed data. One option to try to address extreme imputed values is to respecify the imputation model and restrict the maximum number of predictors selected for inclusion in that variable's regression model. Alternatively, bounds could be imposed on the variable to limit the maximum value that could be imputed. In some cases it may make sense to create a list of these "troublesome" variables for data users, so that they are aware of potential problems, as discussed in Stuart *et al.* (2009) and He *et al.* (2009).

Additional graphical and numerical comparisons between the observed and imputed data can also be useful in diagnosing problems (Abayomi *et al.*, 2008). Histograms, quantile-quantile plots, and density plots provide visual representations of the extent that imputed values differ from observed values; however, in large datasets with many variables, it may not be feasible to examine graphical summaries of each variable. Numerical summaries that compare differences in means and standard deviations between the observed and imputed

Table 1 Comparison of observed and imputed values for outpatient service use

	Observed		Imputed		Combined	
Code	*n*	Percentage	*n*	Percentage	*n*	Percentage
0	2435	28.5	222	34.2	2657	28.9
1	6101	71.5	427	65.8	6528	71.1
Total	8536	100.0	649	100.0	9185	100.0

Note: Code indicates how the variable was coded; *n* = number of cases.

Table 2 Comparison of observed and imputed values for number of times consumed more than five drinks in a row

	Observed	Imputed	Combined
Number	710	8475	9185
Minimum	0	0	0
Maximum	30	4.50×10^{15}	4.50×10^{15}
Mean	1.96	5.1×10^{11}	4.90×10^{11}
Standard deviation	4.25	4.89×10^{13}	4.70×10^{13}

values are an additional approach to identifying variables of concern and may be more feasible within the context of large datasets (e.g. He *et al.*, 2009). Stuart *et al.* (2009) suggest identifying variables where the absolute difference in means between observed and imputed values is greater than two standard deviations, or where the ratio of variances is less than 0.5 or greater than 2.0. It is important to recognize that variables with observed versus imputed differences of this magnitude do not necessarily mean that the imputations are inaccurate; they may in fact be reasonable (and in fact may be indicative of the bias that the imputations are trying to address). For example, if older children are more likely to be missing delinquent behavior information and age is associated with delinquency, then the distributions of the imputed values and observed values of delinquent behavior are likely to differ. Imputation diagnostics should be used to identify potentially problematic variables; then information regarding the missingness, along with substantive content-area knowledge, can guide imputers in determining whether the imputations are reasonable or the imputation procedure should be further modified. Further information on imputation diagnostics are described in Abayomi *et al.* (2008), He *et al.* (2009), and Stuart *et al.* (2009).

After the data have been imputed and diagnostics have been conducted, a few remaining steps may be necessary prior to releasing the data for analyses. The data may need to be cleaned and processed. For example, as mentioned earlier, when generating imputations some software programs (such as IVEware) create "place holders" for values that are missing by design. These place holders should be recoded back to missing so that analyses are not unintentionally performed with the placeholder values representing actual values. It is also important to document how the data were imputed (e.g. which auxiliary variables and interactions were included, the method used), what assumptions were made, and which variables (based on the diagnostics) may need further consideration before any analysis.

Software programs

There are a number of software packages available to impute missing data using MICE procedures. These include IVEware, WinMICE, which is designed specifically to impute multilevel missing data, and procedures for Stata (ice), S-Plus (MICE), R (MICE, mi), and SPSS. Methods for implementing MICE in IVEware, R, and Stata are briefly described later. Detailed descriptions and comparisons of these and other software can also be found in Graham (2009), Harel and Zhou (2007), Horton and Kleinman (2007) and Yu *et al.* (2007).

Briefly, IVEware (Raghunathan *et al.*, 2002) is freeware that can either run through SAS or as a stand alone program. It can impute variables using linear, logistic, Poisson, and mixed logistic/linear models. It has many features that are well suited to imputing missing data in large datasets and can accommodate nearly all of the complexities discussed earlier. It is the only package that we know of that has built in stepwise procedures, which eliminate the need for the imputer to manually run stepwise regression or to individually specify each regression equation. This makes IVEware particularly useful for large datasets (Stuart *et al.*, 2009).

For Stata, the *ice* command (Royston, 2005) implements MICE in a similar manner to IVEware. Ice is not a built-in Stata command, but rather is a program that can be downloaded for free from within Stata. Ice imputes variables using linear, logistic, multinomial logistic and ordered logistic regression. With *ice*, all of the variables specified in the imputation model are included in the regression models, unless the imputer specifies the particular subset of variables to be used in each regression equation. This makes the package potentially less useful for very large datasets as it could either result in unstable models with a very large number of predictors, or it requires the imputer to specify each regression model separately. However, there are also add-on functions that can be used to facilitate the use of stepwise models within ice (pred_eq and check_eq).

The *mi* (Gelman *et al.*, 2011; Su *et al.*, in press) and *mice* (van Buuren and Groothuis-Oudshoon, in press) packages implement the MICE procedure within R (R Development Core Team, 2008). The *mi* package calls MICE "multiple iterative regression imputation." It uses linear regression, logistic regression, multinomial log-linear models, or Poisson regression for each variable, as appropriate, and it contains a number of tools to help the procedure run smoothly and for performing diagnostics. The *mice* package also includes capacities for linear multilevel data as well as built-in diagnostics.

Analyzing multiply imputed data

Once the data have been imputed, each imputed dataset is "complete" in the sense that it has no missing values (except those missing by design). Analyzing multiply imputed data involves two steps: (1) running a standard analysis (e.g. regression) on each of the imputed datasets, and (2) combining the estimates from each dataset to obtain the final result. The variance estimates calculated in Step 2 involve both the "within" variance calculated for

each dataset individually, as well as the "between" variance that reflects the uncertainty in the imputations – how variable the results are across the imputed datasets. The formulas for combining coefficients and estimates in Step 2 are provided in Schafer and Graham (2002). While it is possible to write a short computer program to do the combining, many standard statistical software packages include procedures to combine results across datasets automatically. Thus, from the user's perspective doing these two steps and obtaining the final estimates are often no more complicated than running a single regression in a single dataset.

With the exception of WinMice (Jacobusse, 2005), each of the software programs mentioned earlier have the capability to analyze multiply imputed data. There are also software packages that do not implement MICE, but have the capacity to analyze multiply imputed data in a straightforward way (e.g. Mplus, SAS). Briefly, to analyze multiply imputed data in Stata 11.0, the command "mi" (StataCorp, 2009) is used to specify that analyses are conducted on multiply imputed data. The "mim" command, a free program that was developed for use on earlier versions of Stata, can also be used to analyze multiply imputed data (Royston *et al.*, 2009). While there are some differences between mi and mim in data management properties, the two commands are similar in their estimation and post-estimation abilities. In R, the mi and mice packages discussed earlier have built in functions for analyzing multiply imputed data; comparable functions are also found in the mitools package (Lumley, 2008). It is worth nothing that the mice package for R is one of the only packages that allows for model testing using multiply imputed data. In Mplus, the "IMPUTATION" command is specified and a text file is created that contains a column list of the names of each imputed dataset (Muthen and Muthen, 1998–2007). In SAS, the command PROC MIANALYZE (SAS Institute Inc., 2008) is used. Additional details, features, and sample code for each of these, and the other previously mentioned software packages, is available elsewhere (SAS: Yuan, 2000; Stata: Royston, 2005; StataCorp, 2009; R: Lumley, 2008; Mplus: Muthen and Muthen, 1998–2007; SPSS: SPSS Inc., 2009b; IVEware: Raghunathan *et al.*, 2002).

While there have been significant advances in the accessibility of software that can analyze multiply imputed data, there are limitations to the capabilities of the programs. There may be analyses that researchers want to conduct that are not available, either in a particular software program, or that have just not been developed. For example, programs vary in the ability to perform post-estimation commands on imputed data, and to our knowledge none of the programs have graphical capabilities with imputed data. (The standard graphing commands will generally work, but will treat the data as one large dataset rather than recognizing and treating the data as multiply imputed data.) In these situations, analysts can write a program to perform the analyses and combine the results as described earlier. Alternatively, analysts can run the commands on individual imputed datasets and examine the results for consistency across the datasets. While this is not the ideal method, it does provide some information; this may be particularly helpful for graphs. As advances in analyzing multiply imputed data continue to develop, additional options for conducting analyses, running model diagnostics and model checking will hopefully follow.

Discussion

Although the MICE approach is a principled method of addressing missing data, it is important to acknowledge certain complexities and limitations of the approach. While MICE offers great advantage over other missing data techniques in terms of its flexibility, a primary disadvantage is that MICE does not have the same theoretical justification as other imputation approaches. In particular, fitting a series of conditional distributions, as is done using the series of regression models, may not be consistent with a proper joint distribution. Initial research suggests that this may not be a large issue in applied settings (Brand, 1999; Schafer and Graham, 2002); however, further research is needed into the implications for practice. When there are relatively few variables needing imputation and a multivariate normal model would be appropriate (i.e. the variables are continuous and approximately normally distributed), a joint model such as a multivariate normal may be preferable.

There are also a number of data complexities for which standard imputation strategies do not yet exist. For example, clustering is important to address in both data analyses and when imputing missing data, but clustering is not always automatically incorporated by the MICE procedures. Graham (2009) suggests that the imputation procedures used with multi-level data can vary depending upon the types of analyses anticipated. If the analyses will involve cluster-specific intercepts and coefficients (slopes), then the missing data should be imputed within each cluster (e.g. site). Alternatively, if only a random intercepts model will fit to the data, then the clustering variable can be dummy coded and included as a predictor in the imputation procedure. However, if the clusters are small, these options may not be feasible. While this advice is the

current standard, methods for addressing multi-level missing data are an area of on-going statistical research; additional information and state-of-the-art methods are found in Beunckens *et al.* (2007), and Yucel (2008). Data that have sampling weights add another layer of complexity to imputing missing data. We are unfamiliar with software that automatically incorporates sampling weights into the MICE process. Schenker *et al.* (2006) implemented MICE procedures with data from the National Health Interview Survey and included indicators of the sampling weights and the sampling unit as predictors in the imputation model. Longitudinal data offers another challenge. Additional work should investigate the best strategies for imputing longitudinal data, especially when there are large numbers of variables collected at each time point. Information on methods specifically for longitudinal settings, can be found in Demeritas and Hedeker (2008), Li *et al.* (2006), and Nevalainen *et al.* (2009).

This paper has provided an introduction to the chained approach to imputing missing data. Further methodological research is needed to investigate procedures to impute data with particularly complex structures, such as clustering and longitudinal data, to better understand and address the limitations of the MICE approach, and to continue developing imputation diagnostics. However, as research into missing data techniques continues, MICE offers a principled yet flexible method of addressing missing data that is accessible to a wide range of users.

Acknowledgments

This work was supported by the National Institute of Mental Health 1R01MH075828-01A1.

Declaration of interest statement

The authors have no competing interests.

References

Abayomi K., Gelman A., Levy, M. (2008) Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **57**, 273–291, DOI: 10.1111/j.1467-9876.2007.00613.x

Beunckens C., Molenberghs G., Thijs H., Verbeke G. (2007) Incomplete hierarchical data. *Statistical Methods in Medical Research*, **16**, 457–492, DOI: 10.1177/0962280206075310

Brand J.P.L. (1999) Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets, unpublished, Erasmus University, Rotterdam.

Collins L.M., Schafer J.L., Kam C.M. (2001) A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, **6**, 330–351, DOI: 10.1037/1082-989X.6.4.330

Demeritas H., Hedeker D. (2008) An imputation strategy for incomplete longitudinal ordinal data. *Statistics in Medicine*, **27**, 4086–4093, DOI: 10.1002/sim.3239

Gelman A., Hill J., Yajima M., Su Y., Pittau M. (2011) mi: Missing Data Imputation and Model Checking. Package for the R statistical software. http://lib.stat.cmu.edu/R/CRAN/ [27 January 2011]

Graham J.W. (2003) Adding missing-data relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, **10**, 80–100, DOI: 10.1207/S15328007SEM1001_4

Graham J.W. (2009) Missing data analysis: making it work in the real world. *Annual Review of Psychology*, **60**, 549–576, DOI: 10.1146/annurev.psych.58.110405.085530

Graham J.W., Olchowski A.E., Gilreath T.D. (2007) How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, **8**, 206–213, DOI: 10.1007/s11121-007-0070-9

Greenland S., Finkle W.D. (1995) A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, **142**, 1255–1264.

Harel O., Zhou X.H. (2007) Multiple imputation: review of theory, implementation and software. *Statistics in Medicine*, **26**, 3057–3077, DOI: 10.1002/sim.2787

He Y., Zaslavsky A.M., Landrum M.B., Harrington D.P., Catalano P. (2009) Multiple imputation in a large-scale complex survey: a practical guide. *Statistical Methods in Medical Research*, 1–18, DOI: 10.1177/0962280208101273

Horton N.J., Kleinman K.P. (2007) Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, **61**, 79–90, DOI: 10.1198/000313007X172556

Jacobusse G. (2005) *WinMICE User's Manual for WinMICE Prototype Version 0.1*, The Hague, TNO Quality of Life.

Lee K.J., Carlin J.B. (2010) Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*, **171**, 624–632, DOI: 10.1093/aje/kwp425

Li X., Mehrotra D.V., Barnard J. (2006) An analysis of incomplete longitudinal binary data using multiple imputation. *Statistics in Medicine*, **25**, 2107–2124, DOI: 10.1002/sim.2343

Lumley T. (2008) Mitools: tools for multiple imputation of missing data. Package for the R statistical software package. http://cran.r-project.org/web/packages/mitools/ [14 May 2008].

Manteuffel B., Stephens R., Santiago R. (2002) Overview of the national evaluation of the comprehensive community mental health services for children and their families program. *Children's Services: Social Policy, Research, and Practice*, **5**, 3–20, DOI: 10.1207/S15326918CS0501_2

Moons K.G.M., Donders R.A.R.T., Stijnen T., Harrell F.E. (2006) Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, **59**, 1092–1101, DOI: 10.1016/j.jclinepi.2006.01.009

Muthen L.K., Muthen B.O. (1998–2007) *Mplus User's Guide*, 4th edition, Los Angeles, CA, Muthen & Muthen.

Nevalainen J., Kenward M.G., Virtanen S.M. (2009) Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. *Statistics in Medicine*, **28**, 3657–3669, DOI: 10.1002/sim.3731

R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*, Vienna, R Foundation for Statistical Computing.

Raghunathan T.W., Lepkowksi J.M., Van Hoewyk J., Solenbeger P. (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, **27**, 85–95.

Raghunathan T.E., Solenberger P.W., Van Hoewyk J. (2002) *IVEware: Imputation and Variance Estimation Software User Guide*, Ann Arbor, MI, University of Michigan. http://www.isr.umich.edu/src/smp/ive/ [19 May 2008].

Royston P. (2005) Multiple imputation of missing values – update. *The Stata Journal*, **5**, 188–201.

Royston P., Carlin J.B., White I.R. (2009) Multiple imputation of missing values: new features for mim. *The Stata Journal*, **9**, 252–264.

SAS Institute Inc. (2008) *SAS/STAT User's Guide 9.2*, Carey, NC, SAS Institute, Inc. http://support.sas.com/documentation/cdl/en/statug/59654/HTML/default/chap0_toc.htm# [29 May 2008].

Schafer J.L. (1999) Multiple imputation: a primer. *Statistical Methods in Medical Research*, **8**, 3–15, DOI: 10.1177/096228029900800102

Schafer J.L. (2003) Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, **57**, 19–35.

Schafer J.L., Graham J.W. (2002) Missing data: our view of the state of the art. *Psychological Methods*, **7**, 147–177, DOI: 10.1111/1467-9574.00218

Schenker N., Raghunathan T.E., Chiu P-L., Makuc D.M., Zhang G., Cohen A.J. (2006) Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association*, **101**, 924–933, DOI: 10.1198/016214505000001375

Scientific Software International (2006) *Lisrel 8.8*, Lincolnwood, IL, Scientific Software International, Inc.

SPSS Inc. (2009a) *Amos 18.0. SPSS Missing Values 17.0*, Chicago, IL, SPSS Inc.

SPSS Inc. (2009b) *SPSS Missing Values 17.0*, Chicago, IL, SPSS Inc.

StataCorp. (2009) *Stata Multiple-imputation Reference Manual*, College Station, TX, StataCorp., LP.

Stuart E.A., Azur M., Frangakis C.E., Leaf P.J. (2009) Practical imputation with large datasets: a case study of the Children's Mental Health Initiative. *American Journal of Epidemiology*, **169**, 1133–1139, DOI: 10.1093/aje/kwp026

Su Y-S., Gelman A., Hill J., Yajima M. (in press) Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*. http://www.stat.columbia.edu/~gelman/research/published/mipaper.rev04.pdf.

Van Buuren S. (2007) Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, **16**, 219–242, DOI: 10.1177/0962280206074463

Van Buuren S., Groothuis-Oudshoon K. (in press) MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. http://lib.stat.cmu.edu/R/CRAN/web/packages/mice/index.html

Yu L.M., Burton A., Rivero-Arias O. (2007) Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research*, **16**, 243–258, DOI: 10.1177/0962280206074464

Yuan Y.C. (2000) Multiple imputation for missing data: new concepts and development. In *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference* (Paper No. 267), Cary, NC, SAS Institute.

Yucel Y.M. (2008) Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A*, **366**, 2389–2403, DOI: 10.1098/rsta.2008.0038