Multiple Imputation by Chained Equations in Praxis: Guidelines and Review

Jesper N. Wulff[1] and Linda Ejlskov[2]
[1]Aarhus University, Department of Economics and Business Economics, Denmark
[2]Aalborg University, Department of Health, Science and Technology, Denmark

Abstract: Multiple imputation by chained equations (MICE) is an effective tool to handle missing data - an almost unavoidable problem in quantitative data analysis. However, despite the empirical and theoretical evidence supporting the use of MICE, researchers in the social sciences often resort to inferior approaches unnecessarily risking erroneous results. The complexity of the decision process when encountering missing data may be what is discouraging potential users from adopting the appropriate technique. In this article, we develop straightforward step-by-step graphical guidelines on how to handle missing data based on a comprehensive literature review. It is our hope that these guidelines can help improve current standards of handling missing data. The guidelines incorporate recent innovations on how to handle missing data such as random forests and predictive mean matching. Thus, the data analysts who already actively apply MICE may use it to review some of the newest developments. We demonstrate how the guidelines can be used in praxis using the statistical program R and data from the European Social Survey. We demonstrate central decisions such as variable selection and number of imputations as well as how to handle typical challenges such as skewed distributions and data transformations. These guidelines will enable a social science researcher to go through the process of handling missing data while adhering to the newest developments in the field.

Keywords: Multiple imputation by chained equations, MICE, missing data, guidelines, review, R

1. Introduction

Valid tools to handle missing data should be in every data analyst's toolkit. Not having these tools force data analysts to resort to inferior default approaches when analysing incomplete data unnecessarily risking erroneous results. The most often used technique to cope with incomplete data is to apply some form of complete case analysis as e.g. listwise deletion or mean-value imputation. In the case of the former, a respondent is left out if data are missing on one or more of the variables included in the analysis, while the latter approach fills in the missing values with the mean calculated from the observed data. To remove respondents that lack data on certain variables or single-impute missing observations can be harmless, but often it leads to biased estimates and incorrect standard errors (Sterne et al. 2009). Thus, it is not surprising that research methodologists only recommend to use list- / pairwise deletion or single imputation techniques in rare cases (Donders et al. 2006; Greenland & Finkle 1995; King et al. 2001; Newman 2014).

Multiple imputation (MI) is an attractive alternative to help researchers make the best use of their data. While MI is a powerful technique, unleashing its full potential requires its user to make a series of important decisions. If the wrong choices are made, it may compromise the results. This dire potential consequence coupled with the complexity of the decision process may discourage potential users from adopting the technique. In part, it may explain why MI is rarely used in practice in psychology (Roth 1994), political science (King et al. 2001) and management research (Newman 2014), which are fields that often rely on datasets containing missing values. This is unfortunate as very convincing evidence exists suggesting that MI provides more precise and reliable results than simpler methods (Graham & Schafer 1999; Little & Rubin 1987; Rubin & Schenker 1986; Rubin 1996).

We acknowledge that for a researcher unfamiliar with MI getting started may seem like a daunting task. Thus, in this paper, we first offer a non-technical overview of the state of the art and develop concrete guidelines on how to implement MI-techniques with a focus on multiple imputation by chained equations (MICE) in section 2. Our goal is to spread the use of MI-techniques and thereby increase the quality of the research produced. In addition, data analysts who already implement MI-techniques may also benefit from this paper by updating their knowledge with the newest developments in the field as e.g. predictive mean matching, random forest imputation and model selection.

Reference this paper as: Wulff J and Ejlskov L, "Multiple Imputation by Chained Equations in Praxis: Guidelines and Review" *The Electronic Journal of Business Research Methods Volume 15 Issue 1 2017*, (pp41-56) available online at www.ejbrm.com

In section 3 we demonstrate how to implement the guidelines in praxis and overcome common challenges facing the researcher when using MICE to handle missing data. We demonstrate how to handle missing data using MICE and compare the results of different imputation methods. The demonstration is based on data from the European Social Survey using R. R is a free software environment for statistical computing enabling the reader of this text to apply the presented techniques without depending on access to a specific type of commercial software (R Core Team 2015). Further, using R we can demonstrate some of the newest developments in the field not yet accessible through other standard software packages. The R-syntax is available in the appendix. https://www.researchgate.net/publication/316190594_Appendix_with_R_code

2. Multiple imputation

Before describing MI, it makes sense to briefly describe the mechanisms that makes the data *missing*: Missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). MCAR is when the probability of the data missing depends on neither the observed nor the unobserved data. Missing data are MAR when the probability of missing data to some extent depends on the observed data. Finally, data are MNAR when the probability of missing data depends on the missing data values themselves. Missing data are almost never MCAR, but instead somewhere in between MAR and MNAR (Graham 2009). If the missing data are tending towards the MAR category, MI-techniques outperform standard techniques such as listwise deletion (Rubin 1996).

MI was proposed by Rubin (1987) as a statistical technique for handling missing data. At its core, MI uses the distribution of the observed data to estimate a set of likely values of the data that are missing. MI estimates these values *M* times each time incorporating a random component to reflect the uncertainty about the missing values. After running the procedure, we are left with *M* different datasets on which we perform the desired analysis (e.g. linear regression). Next, we take the average of the parameter estimates across the *M* datasets resulting in one unbiased parameter estimate for each parameter in our model. The standard errors are calculated using Rubin's (1987) formula. An illustration of the whole MICE-process based on Rubin (Rubin 1987) is shown in Figure 1 below.

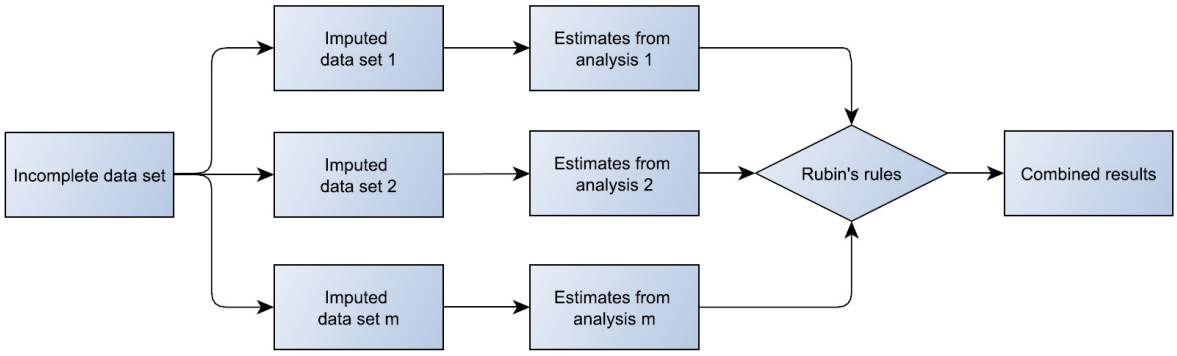

Figure 1: The MICE-process (Rubin 1987)

The power of MI lies in its many imputations. While every single imputation is imprecise, the combination of several imputations takes the uncertainty of each imputation into account. For MAR or MCAR data the pooled parameter estimates are unbiased and standard errors are corrected appropriately (King et al. 2001). In other words, traditional hypothesis testing based on MI-standard errors are more accurate (Newman 2014).

2.1 MICE

In praxis, we often encounter missing values in multiple variables. MICE is a popular adaption of MI and is available for the user through the most commonly used software packages. MICE changes the imputation problem to a series of estimations where each variable takes its turn in being regressed on the other variables. MICE loops through the variables predicting each variable dependent on the others. This procedure provides great flexibility as each variable may be assigned a suitable distribution, e.g. poisson, linear or binomial (White et al. 2011).

MICE runs through an iterative process: In the first iteration, the imputation model for the variable with the least missing values is estimated using only complete data. Next, the variable with the second least missing

values is imputed using the complete data and the imputed values from the last iteration. After each variable has been through this process, the cycle is repeated using the data from the last iteration. Typically, ten iterations are performed where the imputed values after the 10th and final iteration constitutes one imputed data set (Stuart et al. 2009).

As mentioned above, MICE has an important ability to handle different variable types as each variable is imputed using its own imputation model (Bartlett et al. 2014). This provides flexibility and makes it possible to impute data sets that include hundreds of variables (He et al. 2010). For instance, continuous variables may be modeled through linear regression, binary variables through logistic regression etc. (Chevret et al. 2015). One downside of MICE is that it does not have the same theoretical justification as e.g. multivariate normal imputation. In praxis, this does not seem to be an issue (White et al. 2011). Another drawback is the challenge of choosing an appropriate imputation model. Later in this paper, we walk through an example of building an imputation model.

Random forest imputation presents a recently developed alternative to standard MICE procedures. Random forests are an extension of classification and regression trees. Classification and regression trees use a binary splitting approach recursively subdividing the data based on the values of the predictor variables. Random forests build many trees each time varying the sample and the predictors. Consequently, a new bootstrapped sample of observations and predictors are selected for each tree (Hastie et al. 2009).

Under certain circumstances, random forest imputation may offer an attractive alternative to standard MICE. First, in standard MICE omitting important interactions and other non-linear terms may lead to biased results (Seaman et al. 2012). Random forest imputation, on the other hand, is a non-parametric technique well-suited for handling complex non-linear relationships. This is especially beneficial when the imputation model is difficult to specify. Simulation with data sets including interactions have demonstrated that random forests MICE (Shah et al. 2014; Doove et al. 2014) and even MICE with regular regression trees (Burgette & Reiter 2010; Doove et al. 2014) result in less biased parameter estimates than conventional standard MICE.

Random forest imputation is also an attractive alternative in other cases where standard MICE is simply not feasible. Without prior information, standard MICE is not possible in high-dimensional settings when the number of predictor variables exceed the number of observations (Hardt et al. 2012; Zhao & Long 2016) and may not converge if highly correlated predictor variables are included in the imputation model. Because random forest imputation still needs to be tested on a larger range of data, standard MICE may still be preferred in standard cases. However, due to promising results, it is an option in situations where standard MICE is known to produce biased estimates (e.g. when facing severe non-linarites) or is simply unfeasible (e.g. when dealing with high-dimensional data).

2.2 The amount of missing data

Before proceeding with the example, it is relevant to briefly discuss what the partial missing data rate means for whether one should consider MI-techniques. Newman (2014) suggests a rule-of-thumb of using MI when confronting a missing-rate of 10% of higher. At this level, the difference between MI and simpler techniques is expected to be substantial. Multiple studies have investigated the impact of different missing rates, e.g. from 10% - 90% (Janssen et al. 2010) and 2.5% - 30% (Knol et al. 2010), and found that MI outperforms listwise deletion and similar techniques under different missing mechanisms and sample sizes. For instance, it is possible for an odds-ratio estimate to deviate systematically from the true value even for at 5% MAR-missing rate (Knol et al. 2010).

A natural question is whether the missing rate can be too high to use MI. Studies have shown that MI is unbiased to around 50%, but gets unstable for higher rates, especially if the data have skewed distributions (Haji-Maghsoudi et al. 2013; Lee & Carlin 2012). However, this does not imply that listwise deletion should be preferred as MI exhibits superior performance even for a 75% data loss despite biased estimates (Marshall et al. 2010). In praxis, though, high computation times caused by a very high missing data rate may make MI infeasible.

MI-scholars have investigated and debated whether listwise deletion is ever appropriate. The short answer seems to be only in rare cases (Heitjan & Rubin 1990; King et al. 2001). As pointed out by Newman (2014), the 10% rule-of-thumb illustrates an important point despite the arbitrary threshold: When the amount of missing

data is small it will most likely make a very little difference to use MI (Marshall et al. 2010). In summary, research indicates that MI is the superior choice when facing missing rates above 10% and in some cases already around 5%. To sum up the suggestions in the literature, we have developed the following guidelines illustrated in the decision tree in Figure 2. In some cases it might be worth considering MI even though the partial missing rates are below 10%. For example, in cases when including a number of variables in a regression model with low partial missing rates. In such a case this might result in a total missing rate in the full regression model that is substantially higher than in a simple bivariate regression model. Thus, we suggest always to check for both total missing rates as well as partial missing rates.

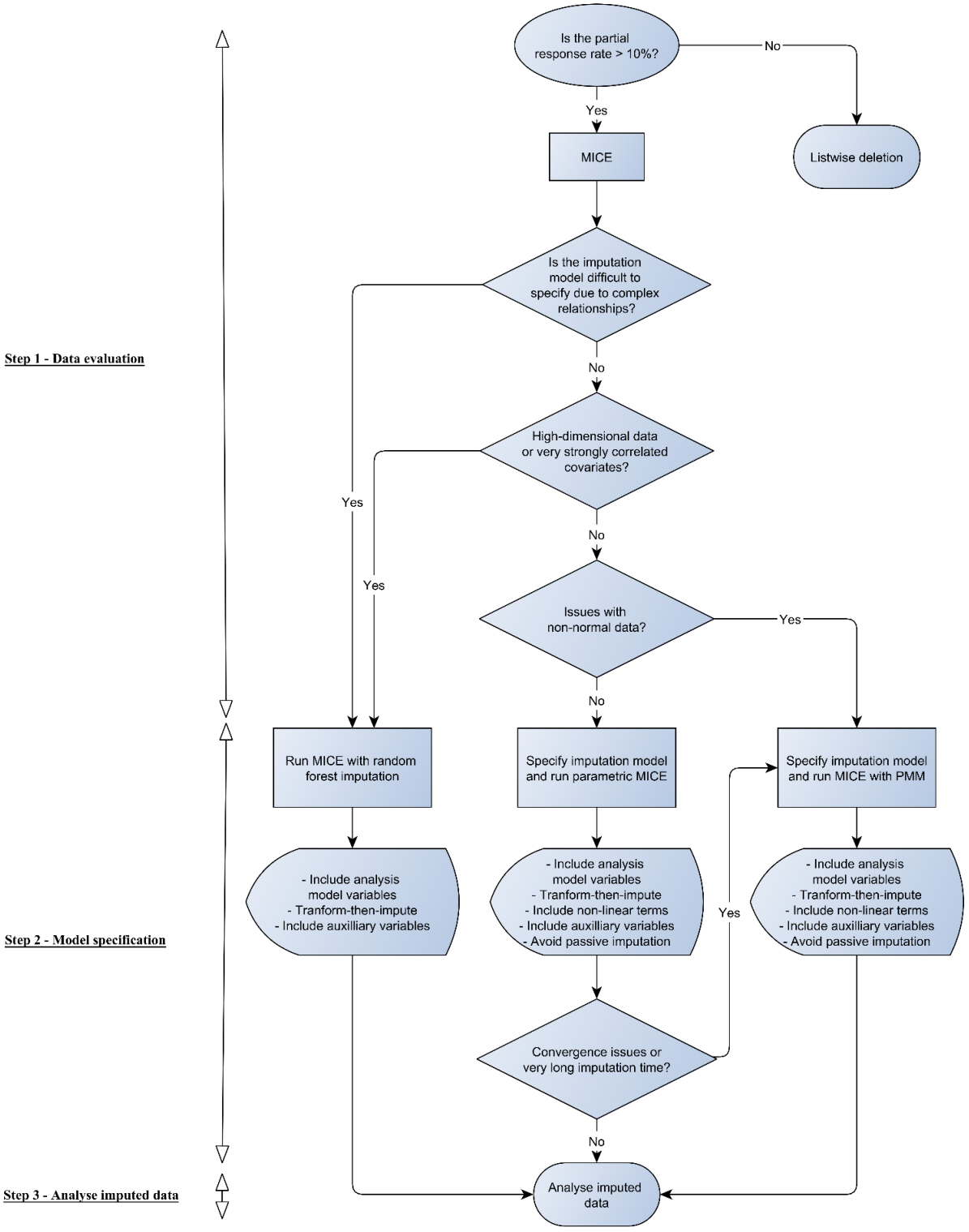

Figure 2: MICE decision tree

3. Multiple imputation: An example

In the following, we walk through an example of using multiple imputation on a real data set following the guidelines we have presented in Figure 2. In our example, we are not going to explicitly choose between the methods but instead compare them in terms of runtime and end results. However, for each method we do detail its appropriateness in relation to the real-life example. As we explain below, the first phase in the specification part in step 2 demands more from the analyst when using standard MICE than when using random forest MICE. The analysis procedure of the imputed data, however, does not differ depending on the chosen method of imputation.

3.1 Data

The real life data used in the example stem from The European Social Survey (ESS). The ESS monitors and charts a long range of individual attitudes, beliefs and behaviour patterns in Europe and is a cross-national repeated cross-sectional survey with the first round occurring in 2002 and has since then been repeated every second year (ESS Round 6: European Social Survey Round 6 Data 2012). In this paper, we have used the 6th round of ESS, which in addition to the core modules has a special emphasis on personal and social wellbeing. The target population for the survey consists of persons aged 15 and over that are resident within private households and in total 54,673 participants from 29 surveyed countries answered the interview-administered questionnaire. For this example, we are interested in investigating how different attitudes towards immigration are associated with happiness among the elderly population in Europe controlled for a range of covariates. After excluding people below the age of 71, 7,582 participants were included in the shown example.

3.2 Step 1: Data evaluation

The sample is summarized in Table I. As it can be seen in Table I, the partial response rate is above 10% for one of the variables and above 5% for five variables. This means that using a MICE framework to handle missing data is appropriate. The next phase is choosing the appropriate imputation method. This choice is discussed throughout the remainder of section three.

Table I: Summary of subsample

Names	Description	Number missing	Proportion missing	Range	Mean(sd)
Analysis model variables					
happy	Level of happiness	102	1.35	1-11	7.85(2.32)
flapppl	Feel appreciated by people you are close to	140	1.85	1-11	8.86(1.91)
agea	Age	133	1.75	71-103	77.54(5.29)
gndr	Gender	3	0.04	1-2	1.58(0.49)
health	Self-rated health	17	0.22	1-5	2.9(0.95)
hinctnta	Household's total net income	1372	18.1	1-10	3.54(2.26)
grdfincc	Government seeks to reduce income differences	550	7.25	1-11	5(2.89)
eduyrs	Education in years	127	1.68	0-42	9.97(4.51)
imueclt	Cultural life undermined or enriched by immigrants	701	9.25	1-11	6.03(2.54)
imwbcnt	Immigrants make country worse or better place to live	689	9.09	1-11	5.52(2.39)
imbgeco	Immigration bad or good for country's economy	658	8.68	1-11	5.52(2.52)
cntry	Country of residence				
Auxilliary variables					
fltlnl	Loneliness	77	1.02	1-4	1.69(0.9)
rehlppl	Receive help and support from people you are close to	112	1.48	1-7	5.99(1.33)
prhlppl	Provide help and support to people you are close to	149	1.97	1-7	5.96(1.39)
inprdsc	How many people with whom you can discuss intimate and personal matters	185	2.44	1-7	3.41(1.54)

Names	Description	Number missing	Proportion missing	Range	Mean(sd)
sclact	Take part in social activities compared to others of same age	349	4.6	1-5	2.54(1.09)
hincfel	Feelings about income	75	0.99	1-4	2.27(0.95)
dweight	Design weights	0	0	0.03-4	0.87(0.38)
pweight	Population weights	0	0	0.03-5.5	1.06(1.3)

3.3 Step 2: Model specification

Below, we discuss selected considerations and choices that are central in the specification of the imputation model when using MICE. As it will become evident, this step is often the most cumbersome. We provide examples under each point using the data described above and the R-package mice (van Buuren & Groothuis-Oudshoorn 2011).

3.3.1 Choosing variables

MICE needs an imputation model and an analysis model. While the imputation model is used for filling in the missing values, the purpose of the analysis model is to analyse the imputed data afterwards. These models need to be compatible. This means that every relationship included in the analysis model needs to be included in the imputation model regardless of whether they contain missing values or not (von Hippel 2009). Note that this includes all the variables in the analysis model including the dependent variable (Graham 2009). If the intent is to use more than one analysis model, the imputation model needs to include all the variables included in all of the analysis models (White et al. 2011). Variables that are mutually exclusive should be included as one combined categorical variable to avoid that a respondent may belong to several categories simultaneously. If the analyst is conducting survival analyses, the censoring variable and a variable containing the Nelson-Aalen estimate of cumulative hazard function should be included as well (White & Royston 2009). If the variable selection makes the imputation infeasible due to too many variables relative to the observations or unstable due to highly correlated predictors, random forest imputation may be considered (Shah et al. 2014).

Our analysis models in the current example include the following variables (also see Table I): Level of happiness (outcome variable), feeling appreciated by the people close to you, age, gender, self-rated health, income, government actions to reduce income differences, years of education, three questions about opinions on immigration and country of residence.

3.3.2 Interactions and other transformations

If the analysis model contains interactions and the imputation is performed through standard MICE, the analyst must specify these in the imputation model as well to avoid bias in the analysis (Seaman et al. 2012). This principle holds for other nonlinear terms, e.g. log-transforms or polynomials, and may be summed up as *transform-then-combine* (Enders et al. 2014; von Hippel 2009). All other methods involving correcting values after the imputation are not advised (von Hippel 2009). For instance, it might be tempting to adjust the imputed values for a squared variable so they correspond exactly to the actual squared values of the imputed variable. Even passively generating interaction terms as it is possible in some statistical packages may lead to biased estimates (von Hippel 2009). If the data analyst suspects conceptually or empirically that important nonlinear terms are left out, random forest imputation may offer an alternative where the researcher does not specify the interactions in the imputation model (Stekhoven & Bühlmann 2012).

To illustrate the principle of incorporating non-linear terms, we include the interaction between the two income-related variables (government seeks to reduce income differences (grdfincc) and the household's total net income (hinctnta)) into the standard MICE model. To reduce collinearity, we mean-center the variables before the generating the interaction term (von Hippel 2009).

3.3.3 Composite scales

When composite scales generated from other variables are in the analysis model, the researcher may choose to form the scale before or after the imputation. Graham (2009) suggests to generate the scale before the imputation when at least half of the items are observed, the items exhibit high coefficient alphas and have high item-total correlations. In other cases, the recommendation is to first impute and then construct the scales. Even though both approaches results in unbiased estimates, the most efficient results is generally achieved by imputing first and then combining the items (Gottschall et al. 2011).

After our imputation, we generate a scale that will reflect the attitude towards immigration. We will do this by summing three immigration-related variables: The degree you believe cultural life is enriched by immigrants; immigrants make your country a better place to live; immigration is good for the economy. A higher score on the index reflects a more positive attitude towards immigration.

3.3.4 Auxiliary variables

Auxiliary variables are not used in the analysis model but are included in the imputation model as they are suspected to contain information about the missing values. Including auxiliary variables that are correlated with the analysis variables or are good predictors of the missing values reduces bias and make the MAR-assumption more reliable (Schafer 2003). Studies have shown that a more loose strategy where extra variables are added on a looser basis is preferred over a more restrictive strategy as the marginal costs of adding variables is often low (Collins et al. 2001). Still, the ratio of variables to cases with complete data should not fall below 1:3 to avoid downward bias in regression coefficients and precision decrease (Hardt et al. 2012). Variables that are highly correlated with an incomplete variable should be included in the imputation model (Little 1995; Graham 2009). Finally, if the data contain survey weights these should be included as a covariate in the imputation model (Kim et al. 2006).

The included auxiliary variables can be seen in Table I. We follow a relatively loose strategy but only include one variable in the imputation of another if it has a correlation of at least 0.2 to shorten the computation time. The ESS provide survey weights why these are also included in the imputation model[1].

3.3.5 Handling skewed and non-normal variables

As mentioned above, the rule for choosing an imputation model for a given variable in MICE is to choose the regression model that would also have been appropriate for ordinary analyses. For instance, linear regression for continuous variables, logistic (logit) for binary, poisson for cardinal variables etc. Before the imputation, it is advantageous to run a model with each variable as a dependent variable and check the model's assumptions. If the distribution of a variable is non-normal and the imputation model assumes normality, the distribution of the imputed values may not match the distribution of the observed values well (Morris et al. 2014). However, simulations have shown that a normal imputation model with non-normal data works surprisingly well (Graham & Schafer 1999). Besides turning to random forest imputation, predictive mean matching offers an alternative imputation modeling approach through MICE. This approach is described below.

3.3.6 Predictive mean matching

Predictive mean matching (PMM) is an increasingly popular tool in the MICE-toolbox. PMM produces imputed values that resemble the observed values better than methods based on the normal distribution (White et al. 2011). If the original variable is right-skewed, PMM will produce imputed values that follow the same distributional pattern. The reason for this is that PMM uses the predicted value for a given observation to identify similar observations. Using the identified observations, a matching set is created containing *k* matches. Next, PMM draws from this set at random. Consequently, PMM uses the real values from individuals with real data. This prevents unwanted extrapolation beyond the range of the data (Little 1988).

The analyst can specify the size of the matching set, i.e. how many similar cases a set should contain. The default *k*=1 in Stata and SPSS has been shown to lead to estimated standard errors that are too low resulting in t-statistics that are too large (Morris et al. 2014). Some research has found a small advantage of *k*=3 over *k*=10 (Schenker & Taylor 1996) while the literature contains recommendations of *k*=10 (Morris et al. 2014). A large *k* is likely to be more effective in larger samples and may have poor performance in small samples as the most similar observations may become too different.

In praxis, PMM has shown performance close to correctly specified parametric models and better than poorly specified parametric models characterized by non-normality (Morris et al. 2014; Schenker & Taylor 1996) and moderate skewness (Marshall et al. 2010) despite that the method lacks a formal mathematical justification

[1] Survey sampling weighting adds another layer of complexity to multiple imputation (Azur et al. 2011). Our sample code contains an example of how one might combine multiple imputation and survey weights. However, the reader should be aware that this is an area of on-going research where few reliable guidelines exists. For more details on the matter, we refer to Kim et al. (2006) and Schenker et al. (2006).

(Kenward & Carpenter 2007). Especially the speed of PMM seems to provide it with an advantage for large datasets, but may run into problems for small samples due to the matching set issue.

3.3.7 Random forests

If the decision falls on random forest imputation, there is little to specify for the analyst besides the number of trees. The recommended number is 10 (Shah et al. 2014), which is also the default in the mice package. Similar to standard MICE, it is preferable to include auxiliary variables in the imputation. As mentioned above, the researcher does not compute interaction and higher-order terms beforehand. It remains to be investigated whether other transformations such as log-transformations follow the *transform-then-combine* rule-of-thumb in the case of random forest. Situations when random forest imputation may be preferred are discussed above and an overview of the decision process is available in Figure 2. Below, we compare results from MICE (using software defaults), MICE with some variables imputed through PMM, MICE using full PMM and MICE using random forests.

3.3.8 Comparing methods

In Table II below, we show four different imputation approaches. For MICE Default, we let the software choose the imputation method for each variable. For numeric variables like agea, it chooses PMM, while factor variables with more than two levels are imputed using multinomial logistic regression (polyreg) and two-level factors are imputed using logistic regression (logreg). For MICE Mixed, we have coded factor variables with more than 10 levels as numeric, thus letting them be imputed by PMM. For MICE Full PMM, we impute every variable using PMM, while MICE RF imputed every variable using random forest imputation.

In Figure 3, we graphically compare the distributions of the observed (blue) and imputed (red) composite immigration variable (immi) and one of the immigration-related variables (imueclt) across imputation models. The distributions are very similar even though MICE Default uses multinomial logit while MICE mixed and Full PMM use PMM (Table II). Random forest seems to deviate a little more from the observed data. The close similarities are also apparent when observing one of the variables used to generate the composite variable.

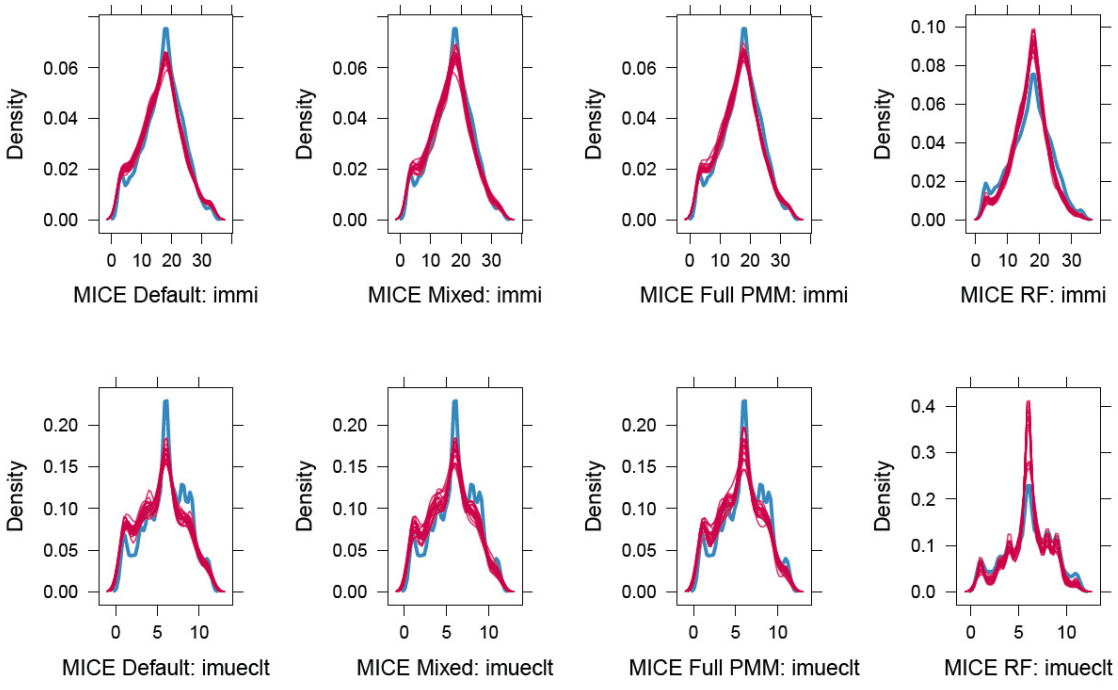

Figure 3: Distributions of observed and imputed values

3.4 Running the imputation model

Before the analysis of the imputed data, we briefly discuss some practicalities considering running the imputation model. The subjects include the number of imputations, model convergence and computation time.

3.4.1 Number of imputations

The data analyst specifies the number of imputations. Earlier work on MI has focused on efficiency and recommends that 3 or 5 imputations are sufficient (Schafer 1997) and that 10 are more than enough (Fichman & Cummings 2003; Schafer 1999). Recent research has turned its focus from efficiency to statistical power instead recommending that the number of imputations should be at least equal to the percent of missing data, i.e. 20% missing data requires 20 imputations (Graham et al. 2007; Bodner 2008; White et al. 2011). This rule-of-thumb minimizes the loss in statistical power in most situations (White et al. 2011).

Table II: Overview of imputation methods for different variables

Variable	MICE Default	MICE Mixed	MICE Full PMM	MICE RF
Analysis model variables				
happy	polyreg	pmm	pmm	rf
flapppl	polyreg	pmm	pmm	rf
agea	pmm	pmm	pmm	rf
gndr	logreg	logreg	pmm	rf
health	polyreg	polyreg	pmm	rf
hinctnta	pmm	pmm	pmm	rf
grdfincc	pmm	pmm	pmm	rf
eduyrs	pmm	pmm	pmm	rf
imueclt	polyreg	pmm	pmm	rf
imwbcnt	polyreg	pmm	pmm	rf
imbgeco	polyreg	pmm	pmm	rf
cntry				
income_int	pmm	pmm	pmm	
Auxilliary variables				
fltlnl	polyreg	polyreg	pmm	rf
rehlppl	polyreg	polyreg	pmm	rf
prhlppl	polyreg	polyreg	pmm	rf
inprdsc	polyreg	polyreg	pmm	rf
sclact	polyreg	polyreg	pmm	rf
hincfel	polyreg	polyreg	pmm	rf
dweight				
pweight				
Imputation time				
Minutes	256.6	71.9	6.7	64.9

Notes: The blank cells indicate that the variable was not imputed either due to non-missingness (cntr, dweight, pweight) or because the variable was not included in the imputation model (income_int for MICE RF). Computation time is based on an average of three runs on an Intel® Core™ i7-5600U CPU 2.60 GHz with 16 GM RAM running 64-bit Windows 7 Enterprise and R version 3.2.4 (2016-03-10).

In our data example, the highest single-item missing rate is 18.1%. In addition, we have relatively high missing rates on the immigration variables (see Table I). Following the rule-of-thumb above, we run 20 imputations.

3.4.2 Model convergence

As explained earlier, MI usually runs 10 iterations with the 10^{th} iteration constituting one imputed dataset. The first 9 iterations are called the burn-in period and are usually sufficient for the process to stabilize (Stuart et al. 2009). Model convergence can be examined by plotting the mean and standard deviation (sd) as a function of the iterations for each variable. In Figure 4 below, we assess model convergence for three variables by plotting the mean and variance of the imputations against the iteration number, respectively. That is, each colored line represents an imputation and for each imputation we plot the mean (left side of figure 4) and sd (right side of figure 4) for each of the 10 imputation iterations. Convergence should be assessed for each variable in the imputation. If a clear trend emerges, the researcher should increase the burn-in period. Below, we show

iteration plots for the mean and standard deviation of our interaction term variable and two of the immigration variables (MICE Default).

From Figure 4, we learn that in our case the means and standard deviations stabilize rather quickly. The interaction term (income_int) shows the most exemplary convergence while the two immigration items take around five iterations to stabilize. Because we know that the two items are highly correlated, this is to be expected. The user can always increase the burn-in period to observe if the chains indeed stabilize. With little reason to increase the burn-in period in our case, we move on to the analysis of the results after a brief note on computation time.

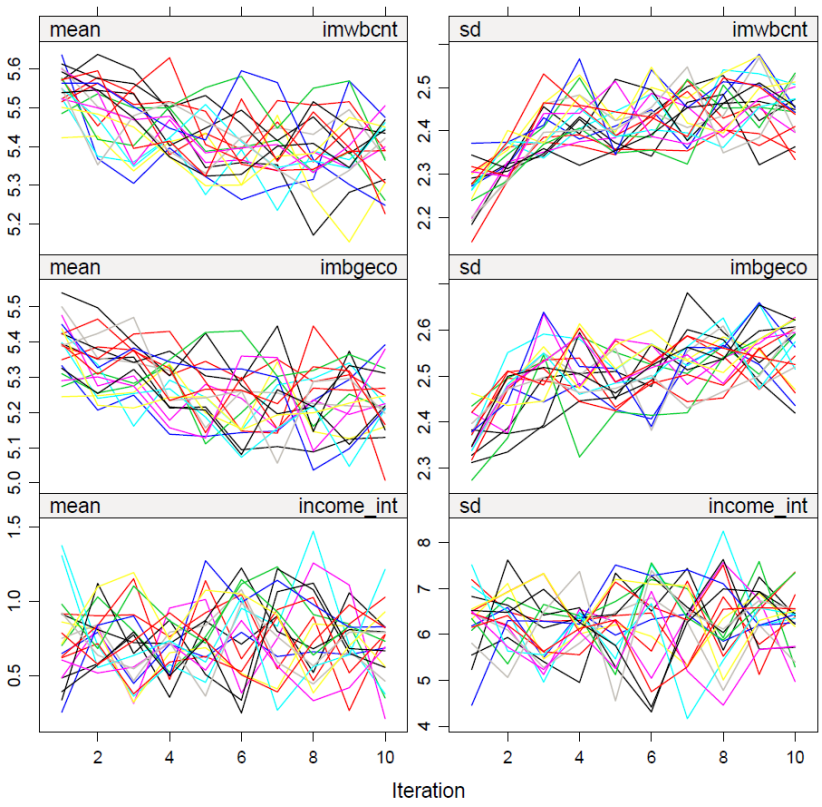

Figure 4: Convergence plots

3.4.3 Computation time

Even though the distributions presented in Figure 3 are very similar, the difference in time it took to impute the data is not: While the MICE Default took 256.6 minutes to run, it took only a about fifth as long for MICE Mixed and just a total of 6.7 minutes for the full PMM (Table II). This illustrates an important point about the enormous speed advantage PMM can have over competing methods: As imputation models grow increasingly complex or in cases of large amounts of missing data, PMM becomes ever more attractive. If the imputation process becomes unreasonably long, factor variables with several levels become obvious candidates for reducing imputation time through PMM. In our case, random forest completes almost as fast as MICE mixed, but still not nearly as fast as full PMM. This illustrates that random forest imputation may be viable alternative in cases where lowering the computation becomes valuable, but a full PMM solution for some reason is undesirable.

3.5 Step 3: Analysis of imputed data

After the imputation, we have *m* complete datasets. We estimate the model (e.g. linear regression) on each of the *m* datasets and combine the estimates to one combined result. Several steps of analysing these datasets is often very similar to running a regression on a single dataset as the process is automated in modern statistical software. This combination step is the same regardless of the imputation procedure selected above. As we discuss below, some issues as e.g. variable selection become quite complicated when using MI.

Below in Table III, we display the estimated coefficients for the analysis model while comparing our four procedures used to impute the missing data as we described above. In addition, we compare the results to a complete case analysis where the missing data are removed using listwise deletion. The complete case results deviate notably from the imputed results. The effect size of household income (hinctnta) is approximately 11% larger when not considering the missing data. It is also noteworthy that the interaction term is smaller when running the regression on the imputed data compared to the complete data. When conducting traditional hypothesis testing, the coefficient on the complete data would be significant at an alpha of 1% (t = -2.78, p = 0.005), while it would only be significant at an alpha of 5% on the imputed data (t = -2.09, p = 0.037). Thus, results may appear more convincing than they really should be when not considering the uncertainty of the missing data.

The results from the default method and the full PMM are very similar. Recall from Table II, however, that the default method took almost 40 times longer to run because of the much more computationally heavy multinomial logit. This repeats the point about the PMM being an attractive alternative when the imputation process becomes unreasonably long.

As the interpretation of the coefficients and standard errors are exactly as in regular regression, we will not dwell further on these. Instead, we move on to areas where some subtle differences to analysing non-imputed data exist.

Table III: Comparing linear regression results from MICE and complete case

Variable	Complete case	MICE Default	MICE Mixed	MICE Full PMM	MICE RF
Intercept	3.869 (0.487)	3.692 (0.393)	3.700 (0.389)	3.662 (0.389)	3.654 (0.389)
flapppl	0.309 (0.014)	0.293 (0.012)	0.296 (0.012)	0.297 (0.012)	0.287 (0.012)
agea	0.013 (0.005)	0.012 (0.004)	0.011 (0.004)	0.012 (0.004)	0.012 (0.004)
gndr	-0.044 (0.053)	-0.012 (0.045)	-0.011 (0.045)	-0.015 (0.045)	-0.020 (0.045)
health2	-0.379 (0.107)	-0.348 (0.092)	-0.341 (0.092)	-0.343 (0.092)	-0.347 (0.092)
health3	-0.777 (0.106)	-0.819 (0.091)	-0.815 (0.091)	-0.816 (0.091)	-0.828 (0.091)
health4	-1.492 (0.122)	-1.573 (0.102)	-1.571 (0.102)	-1.571 (0.102)	-1.572 (0.102)
health5	-2.284 (0.168)	-2.271 (0.138)	-2.258 (0.135)	-2.269 (0.135)	-2.271 (0.135)
eduyrs	-0.010 (0.007)	-0.005 (0.006)	-0.005 (0.006)	-0.006 (0.006)	-0.003 (0.006)
hinctnta	0.098 (0.013)	0.088 (0.011)	0.087 (0.011)	0.087 (0.011)	0.084 (0.011)
grdfincc	0.071 (0.01)	0.076 (0.009)	0.076 (0.009)	0.078 (0.009)	0.068 (0.009)
income_int	-0.011 (0.004)	-0.008 (0.004)	-0.008 (0.004)	-0.008 (0.004)	-0.009 (0.004)
immi	0.030 (0.004)	0.033 (0.004)	0.032 (0.004)	0.033 (0.004)	0.034 (0.004)

Notes: Standard errors in parentheses. Country fixed effects are included in all models.

3.5.1 Fit statistics

Statistics that are not estimators cannot be combined using Rubin's (1987) rules. White et al. (2011) provide an overview of the most common statistics and whether they may be combined, transformed or should not be combined. For linear models, the mean of adjusted R^2 across imputed datasets may be supplemented by information about the percentiles. Fischer's r to z transformation has been suggested in the literature (Harel 2009), but if it is not substantially different, the simple mean may be preferred. For likelihood-based models, as e.g. logistic regression, the likelihood ratio is popular when analysing regular non-imputed data. A method for calculating the likelihood ratio has been suggested (Meng & Rubin 1992), but is has been shown to be less accurate than the easily obtained Wald test statistic (Wood et al. 2008).

Here, we focus on the results from the default MICE method. When using Fischer's r to z transformation, the adjusted R^2 for our model is 0.382 (low 95: 0.364; high 95: 0.399). In our case, there is not much variability in the adjusted R^2 across the 20 imputed datasets. If we compute the adjusted R^2 for a model excluding the interaction term, this simpler model would even fit a little better (Adj. R^2: 0.385; low 95: 0.368; high 95: 0.403). We can compute a Wald test statistic and compare the model with to the model without the interaction term. This results in a test statistic of 4.386 with an associated p-value of 0.037. Using an alpha of 5%, traditional statistical inference suggests that the H_0 of no difference between the models should be rejected. However, these model comparisons consider neither model uncertainty nor do they penalize model complexity adequately. To do this, we move on to a brief discussion of model selection and averaging in the context of imputed data.

3.5.2 Model selection and model averaging

Model selection considers choosing the single best model among a set of candidate models. This preferred model would often be the one that minimizes the generalization error commonly approximated through cross-validation or information-based criteria as e.g. Akaike's Information Criterion (AIC). Model averaging deals with the uncertainty about the model selection process and acknowledges that there may be many good models that describe the data.

Schomaker and Heumann (2014) suggested and tested different ways of integrating model selection and averaging approaches into the MI process. The different algorithms are made available through the MAMI-package (Schomaker 2015). The package allows for several combinations of selection criteria for both model selection and averaging. Below, we show results from their stepwise variable selection based on AIC (Posada & Buckley 2004). This procedure performs stepwise variable selection on each imputed data set. If a given variable is selected its estimate is different from zero while it is zero if the variable is not selected. Thus, coefficients of variables selected in only a few imputed data sets are shrunk towards zero when combined. We continue with the imputed data from the MICE Default below.

Table IV presents the results from the stepwise model selection procedure after multiple imputation. When performing AIC-based model selection on each imputed dataset, the procedure never selects gender (gndr) which is why its coefficient is shrunk completely to zero. This contrasts with a small non-zero coefficient in the standard imputation-analysis and an even larger coefficient in the complete case analysis (Table III). Education (eduyrs) is also of little importance when predicting our outcome variable and is shrunk very close to zero. Notably, the coefficient of the interaction term is no longer significant at an alpha of 5% as the 95% confidence interval includes zero.

Table IV: Combining stepwise selection and multiple imputation

Variable	Estimate	Std.Error	Lower CI	Upper CI
Intercept	3.62484	0.38672	2.86673	4.38296
flapppl	0.29250	0.01182	0.26932	0.31567
agea	0.01202	0.00419	0.00381	0.02023
gndr	0.00000	0.00000	0.00000	0.00000
health2	-0.34627	0.09233	-0.52723	-0.16530
health3	-0.81544	0.09090	-0.99360	-0.63728
health4	-1.56644	0.10127	-1.76493	-1.36795
health5	-2.26402	0.13737	-2.53335	-1.99469
eduyrs	-0.00042	0.00229	-0.00505	0.00422
hinctnta	0.08577	0.01064	0.06489	0.10664
grdfincc	0.07598	0.00858	0.05916	0.09281
income_int	-0.00742	0.00392	-0.01516	0.00032
immi	0.03221	0.00359	0.02516	0.03925

Notes: Country fixed effects are included. Calculated confidence intervals are 95% intervals.

Model selection and averaging on imputed data is an area of on-going research. It is not advised to perform model selection on the complete cases, but rather use the desired selection method based on Rubin's (1987) rules (Chen & Wang 2013). In our example, performing stepwise selection on the complete data does indeed exclude gender, but estimates education to have a coefficient around 23 times larger (β = -0.01, p = 0.14) than on the imputed data (β = -0.0004, p = 0.85). As especially model averaging may be impractical for complicated models, e.g. with numerous possible interactions, stacking the imputed data sets into a single data set using weights at the model-building state may be a pragmatic alternative (Wood et al. 2008). Simply averaging model selection criteria such as AIC across imputed data set is temptingly simple, but not supported in the MI literature (Schomaker & Heumann 2014; White et al. 2011).

3.5.3 Reporting results

Reporting the regression results based on imputed data is not different from regression results based on regular data. However, when using multiple imputation to account for missing data, the data analyst must supply additional information either in the main paper / report or in the supplement materials. Because many of the decisions we make when imputing data may affect the end-result, it is reasonable to state these decisions. In praxis, there seems to be deficiencies in the documentation of missing data and the details about

the imputation (Hayati Rezvan et al. 2015). Sterne et al. (Sterne et al. 2009) provide a set of guidelines about which information to report. Below we condense those that should be made as a minimum and refer to the paper for further details.

(1) The software used including the key settings described in this paper as e.g. number of imputed datasets. (2) A list of the variables used in the imputation model incl. auxiliary variables (e.g. Table I). (3) Handling of non-normally distributed variables and non-linear terms. (4) Comparison of observed and imputed values for variables with a high level of missing rates (e.g. Figure 3). (5) Discussion of notable differences between analyses of complete cases and imputed data. Finally, (6) a summary of the number and fraction of missing values and additional relevant information of the way the data are missing.

4. Concluding remarks

MI is among the most prominent methods to handle missing data and its superiority has been documented in a long range of studies. One of the hurdles of getting started with MI is that it may be difficult navigating around the pitfalls. With this overview, we aim to help more researchers to get started with implementing MI-techniques such as MICE instead of inferior approaches. We hope the data analysts who already actively apply MICE may use it to review some of the newest developments.

We chose a practical focus on fewer techniques in one selected software environment to keep the guidelines simple. We focused on MICE but fully recognize maximum likelihood routines as e.g. full maximum likelihood. MICE and full maximum likelihood produce identical results making the choice between them a matter of personal preference (Enders et al. 2014; Schafer 2003; Collins et al. 2001). We also focused on the R statistical environment in the provided example. For an overview of alternative software packages, we recommend consulting Horton and Kleinman (2007). Finally, we also focused on a subset of commonly used models in the analysis example. After having been introduced to our suggested guidelines, we hope that readers are motivated to dig into some of the details with regard to more complicated analyses of e.g. hierarchical (Zhao & Yucel 2009; van Buuren 2011; Enders et al. 2016; Grund et al. 2016) or longitudinal data structures (Ibrahim & Molenberghs 2009; Jansen et al. 2006).

We hope that our synthesis of the literature into guidelines may help improve current standards of handling missing data. We end with the caveat that our practical guidelines are no more than the name suggests. We have tried to condense the statistical correct praxis into discrete rules-of-thumb to make it accessible to a larger audience. While nuances and details are often lost in a summary, we believe that accessible guidelines are preferable to the praxis that is currently dominating. More detailed and complex decision-rules for praxis are certainly possible but in our opinion, their contribution will be marginal if the basics are ignored.

References

Azur, M.J. et al., 2011. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1), pp.40–9.

Bartlett, J.W. et al., 2014. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical methods in medical research*, 24(4), pp.462–487.

Bodner, T.E., 2008. What Improves with Increased Missing Data Imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, 15(4), pp.651–675.

Burgette, L.F. & Reiter, J.P., 2010. Multiple imputation for missing data via sequential regression trees. *American journal of epidemiology*, 172(9), pp.1070–6.

van Buuren, S., 2011. Multiple imputation of multilevel data. In J. Hox & J. K. Roberts, eds. *The Handbook of Advanced Multilevel Analysis*. Milton Park: Routledge, pp. 173–196.

van Buuren, S. & Groothuis-Oudshoorn, K., 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), pp.1–67.

Chen, Q. & Wang, S., 2013. Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in medicine*, 32(21), pp.3646–59.

Chevret, S., Seaman, S. & Resche-Rigon, M., 2015. Multiple imputation: a mature approach to dealing with missing data. *Intensive care medicine*, 41(2), pp.348–50.

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(4), pp.330–51.

Donders, A.R.T. et al., 2006. Review: a gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10), pp.1087–91.

Doove, L.L., Van Buuren, S. & Dusseldorp, E., 2014. Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72, pp.92–104.
Enders, C.K., Baraldi, A.N. & Cham, H., 2014. Estimating interaction effects with incomplete predictor variables. *Psychological methods*, 19(1), pp.39–55.
Enders, C.K., Mistler, S.A. & Keller, B.T., 2016. Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, 21(2), pp.222–240.
ESS Round 6: European Social Survey Round 6 Data, 2012. *Data file edition 2.1*,
Fichman, M. & Cummings, J.N., 2003. Multiple imputation for missing data: Making the most of what you know. *Organizational Research Methods*, 6(3), pp.282–308.
Gottschall, A.C., West, S.G. & Enders, C.K., 2011. A Comparison of Item-Level and Scale-Level Multiple Imputation for Questionnaire Batteries. *Multivariate Behavioral Research*, 47(1), pp.1–25.
Graham, J.W., 2009. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, pp.549–576.
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(3), pp.206–213.
Graham, J.W. & Schafer, J.L., 1999. On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle, ed. *Statistical strategies for small sample research*. Thousand Oaks, CA: Sage, pp. 1–29.
Greenland, S. & Finkle, W.D., 1995. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American journal of epidemiology*, 142(12), pp.1255–64.
Grund, S., Lüdtke, O. & Robitzsch, A., 2016. Multiple imputation of missing covariate values in multilevel models with random slopes: a cautionary note. *Behavior Research Methods*, 48(2), pp.640–649.
Haji-Maghsoudi, S. et al., 2013. Influence of pattern of missing data on performance of imputation methods: an example using national data on drug injection in prisons. *International journal of health policy and management*, 1(1), pp.69–77.
Hardt, J., Herke, M. & Leonhart, R., 2012. Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC medical research methodology*, 12(1), p.184.
Harel, O., 2009. The estimation of R2 and adjusted R2 in incomplete data sets using multiple imputation. *Journal of Applied Statistics*, 36(10), pp.1109–1118.
Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd ed., Springer.
Hayati Rezvan, P., Lee, K.J. & Simpson, J.A., 2015. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC medical research methodology*, 15(1), p.30.
He, Y. et al., 2010. Multiple imputation in a large-scale complex survey: a practical guide. *Statistical methods in medical research*, 19(6), pp.653–70.
Heitjan, D.F. & Rubin, D.B., 1990. Inference from Coarse Data via Multiple Imputation with Application to Age Heaping. *Journal of the American Statistical Association*, 85(410), pp.304–314.
von Hippel, P.T., 2009. How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, 39(1), pp.265–291.
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(1), pp.79–90.
Ibrahim, J.G. & Molenberghs, G., 2009. Missing data methods in longitudinal studies: a review. *TEST*, 18(1), pp.1–43.
Jansen, I. et al., 2006. Analyzing Incomplete Discrete Longitudinal Clinical Trial Data. *Statistical Science*, 21(1), pp.52–69.
Janssen, K.J.M. et al., 2010. Missing covariate data in medical research: to impute is better than to ignore. *Journal of clinical epidemiology*, 63(7), pp.721–7.
Kenward, M.G. & Carpenter, J., 2007. Multiple imputation: current perspectives. *Statistical methods in medical research*, 16(3), pp.199–218.
Kim, J.K. et al., 2006. On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3), pp.509–521.
King, G. et al., 2001. Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review*, 95(1), pp.49–69.
Knol, M.J. et al., 2010. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *Journal of Clinical Epidemiology*, 63(7), pp.728–736.
Lee, K.J. & Carlin, J.B., 2012. Recovery of information from multiple imputation: a simulation study. *Emerging themes in epidemiology*, 9(1), p.3.
Little, R.J.A., 1988. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3), pp.287–296.
Little, R.J.A., 1995. Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *Journal of the American Statistical Association*, 90(431), pp.1112–1121.
Little, R.J.A. & Rubin, D.B., 1987. *Statistical analysis with missing data*, New York: John Wiley.
Marshall, A., Altman, D.G. & Holder, R.L., 2010. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. *BMC medical research methodology*, 10(1), p.112.
Meng, X.-L. & Rubin, D.B., 1992. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79(1), pp.103–111.
Morris, T.P., White, I.R. & Royston, P., 2014. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC medical research methodology*, 14(1), p.75.

Newman, D.A., 2014. Missing data: Five practical guidelines. *Organizational Research Methods*, 17(4), pp.372–411.
Posada, D. & Buckley, T.R., 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology*, 53(5), pp.793–808.
R Core Team, 2015. R: A Language and Environment for Statistical Computing.
Roth, P., 1994. Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47(3), pp.537–560.
Rubin, D.B., 1987. *Multiple imputation for nonresponse in surveys*, New York: John Wiley and Sons.
Rubin, D.B., 1996. Multiple imputation after 18+ years. *American Statistical Association*, 91(434), pp.473–489.
Rubin, D.B. & Schenker, N., 1986. Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association*, 81(394), pp.366–374.
Schafer, J.L., 1997. *Analysis of incomplete data*, London: Chapman & Hall.
Schafer, J.L., 1999. Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(3), pp.3–15.
Schafer, J.L., 2003. Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, 57(1), pp.19–35.
Schenker, N. et al., 2006. Multiple Imputation of Missing Income Data in the National Health Interview Survey. *Journal of the American Statistical Association*, 101(475), pp.924–933.
Schenker, N. & Taylor, J.M.G., 1996. Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, 22(4), pp.425–446.
Schomaker, M., 2015. MAMI: Model averaging and model selection after multiple imputation.
Schomaker, M. & Heumann, C., 2014. Model selection and model averaging after multiple imputation. *Computational Statistics & Data Analysis*, 71, pp.758–770.
Seaman, S.R., Bartlett, J.W. & White, I.R., 2012. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC medical research methodology*, 12(1), p.46.
Shah, A.D. et al., 2014. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American journal of epidemiology*, 179(6), pp.764–74.
Stekhoven, D.J. & Bühlmann, P., 2012. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics (Oxford, England)*, 28(1), pp.112–8.
Sterne, J.A.C. et al., 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ (Clinical research ed.)*, 338, p.b2393.
Stuart, E.A. et al., 2009. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *American journal of epidemiology*, 169(9), pp.1133–9.
White, I.R. & Royston, P., 2009. Imputing missing covariate values for the Cox model. *Statistics in medicine*, 28(15), pp.1982–98.
White, I.R., Royston, P. & Wood, A.M., 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4), pp.377–399.
Wood, A.M., White, I.R. & Royston, P., 2008. How should variable selection be performed with multiply imputed data? *Statistics in medicine*, 27(17), pp.3227–46.
Zhao, E. & Yucel, R.M., 2009. Performance of sequential imputation method in multilevel applications. In *American Statistical Association*. Alexandria, pp. 2800–2810.
Zhao, Y. & Long, Q., 2016. Multiple imputation in the presence of high-dimensional data. *Statistical Methods in Medical Research*, 25(5), pp.2021–2035.

