Official Journal of the Asian Pacific Society of Respirology
Respirology

INVITED REVIEW SERIES:
MODERN STATISTICAL METHODS IN RESPIRATORY MEDICINE
SERIES EDITORS: RORY WOLFE AND MICHAEL ABRAMSON

Introduction to multiple imputation for dealing with missing data

KATHERINE J. LEE[1,2] AND JULIE A. SIMPSON[3]

[1]*Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute,* [2]*Department of Paediatrics, The University of Melbourne,* [3]*Centre for Molecular, Environmental, Genetic & Analytic Epidemiology, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia*

ABSTRACT

Missing data are common in both observational and experimental studies. Multiple imputation (MI) is a two-stage approach where missing values are imputed a number of times using a statistical model based on the available data and then inference is combined across the completed datasets. This approach is becoming increasingly popular for handling missing data. In this paper, we introduce the method of MI, as well as a discussion surrounding when MI can be a useful method for handling missing data and the drawbacks of this approach. We illustrate MI when exploring the association between current asthma status and forced expiratory volume in 1 s after adjustment for potential confounders using data from a population-based longitudinal cohort study.

Key words: experimental study, missing data, multiple imputation, observational study.

Abbreviations: FEV_1, forced expiratory volume in 1 s; MI, multiple imputation; SD, standard deviation; TAHS, Tasmanian Longitudinal Health Study.

INTRODUCTION

Missing data are common in both experimental and observational studies of respiratory health. The problem of missing data can arise in a variety of forms, from item non-response, where an individual has missing data on a particular measure (e.g. when a participant refuses to provide a blood sample for cytokine measurement), to unit non-response, where a participant does not have data for a range of measures (e.g. a participant missing a wave of data collection in a longitudinal study). It can also occur in any number of variables, including the outcome, the exposure of interest and in confounding variables required for adjustment, all of which can lead to issues for the analysis of interest (see Wolfe and Abramson[1], and Kasza and Wolfe[2] which provide a prelude to this paper). For example, the study by Vermeulen *et al*. considers the scenario where there are missing data on health-related quality of life at various follow-up times within a longitudinal study of lung transplantation.[3] This would mean missing outcome data if the research question was around quality of life following lung transplantation, or missing covariate data if the research question was around the association between quality of life following lung transplantation and longer term outcomes.

The most common method of dealing with missing data is to exclude participants who have one or more missing values, a so-called complete case analysis. Excluding participants in this way can introduce selection bias as participants who have complete data may be different to those with missing data. This means that the sample used for analysis may no longer be representative of the population of interest. Complete case analysis can also be inefficient as it can mean excluding a large number of participants. A number of methods have been suggested in the literature which enable participants with missing data to be included in the analysis, such as last observation

Correspondence: Katherine Lee, Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute, Flemington Road, Parkville, Melbourne, Vic. 3052, Australia. Email: katherine.lee@mcri.edu.au

The Authors: Dr Katherine Lee is a biostatistician with 11 years of experience in clinical and statistical research and over 65 peer-reviewed publications. Her main areas of expertise are clinical trials and the method of multiple imputation for dealing with missing data. Associate Professor Julie Simpson is a biostatistician with 20 years of experience in clinical and population health research and over 130 peer-reviewed papers. Her main area of expertise is the application of nonlinear mixed effects models to pharmacokinetic-pharmacodynamic data and statistical methods for handling missing data in longitudinal cohort studies.

Received 6 October 2013; accepted 13 October 2013

carried forward, the missing indicator method, mean substitution, and regression imputation. However, these unprincipled, simple approaches have also been shown to lead to biased results.[4,5] Multiple imputation (MI) is an alternative approach whereby missing values are imputed based on the observed data, repeating this process a number of times to account for the uncertainty in the imputed values. This approach is becoming increasingly popular to deal with missing data as it has the potential to correct the bias in the complete case and alternative analyses.[6,7]

Medical journals now often request that researchers indicate the amount of missing data in their study, assess differences between those individuals with complete data and those with missing data, and explain how the missing data were handled in the statistical analyses.[8] Furthermore, leading journals are starting to request justification for the statistical approach chosen by the authors to deal with the missing data, and state that additional sensitivity analyses may be requested when there are extensive missing data.[9] This increased awareness regarding the impact of missing data means that researchers need to consider the best way to handle missing data in their analyses, and justify their decision regarding the approach chosen to deal with missing data.

In this paper, we provide an introduction to MI as a statistical method for dealing with missing data, illustrating the approach using data from a population-based longitudinal cohort study of respiratory health. We also discuss scenarios where MI can be a useful method for handling missing data and present some of the drawbacks of this approach.

MOTIVATING EXAMPLE

As an example, we consider the question of whether current asthma status is associated with forced expiratory volume in 1 s (FEV_1), after adjustment for age, gender, socio-economic status, smoking status, height and waist circumference using multivariable linear regression (see Kasza and Wolfe[2] for a description of multivariable linear regression). We use data (a random sample) from the fifth decade of follow-up (commencing in 2004) from the Tasmanian Longitudinal Health Study (TAHS). TAHS is a population-based longitudinal cohort study of 8683 children born in 1961 and attending school in Tasmania in 1968.[10] In this study, waist circumference was not available for approximately one quarter of the participants, meaning that when we adjust for this covariate using the standard complete case analysis, those participants are excluded. We present MI as an alternative approach for dealing with these missing data.

For illustrative purposes, we restrict our analysis to 316 TAHS participants who have complete data on all variables in the analysis aside from waist circumference. We note that the estimated associations are for illustrative purposes only, and should not be interpreted as a definitive analysis of this study.

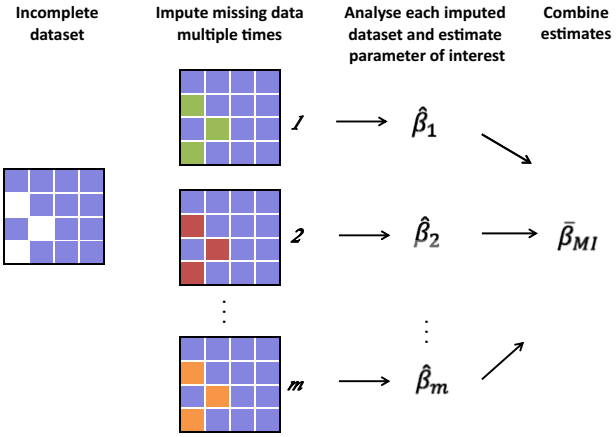

Figure 1 Illustration of the method of multiple imputation. Each box represents a data value where the columns are variables and the rows are individuals. Blank spaces represent the missing values. In this figure, $\hat{\beta}_i$ is the estimate of interest from the completed dataset number *i*, in our case the regression coefficient for current asthma status in a multivariable linear regression model for forced expiratory volume in 1 s, $\bar{\beta}_{MI}$ is the estimate obtained from multiple imputation, and m is the number of imputed datasets.

WHAT IS MI?

MI is a two-stage process. In stage 1, the missing values are imputed by sampling from an imputation model. This imputation model should include all variables that are in the analysis model (outcome, exposure, confounders), as well as additional (at least partially) observed variables that are not included in the analysis model, but are associated with the variable(s) with missing data. These additional variables are known as auxiliary variables. The imputation process is repeated a number of times, that is, a number of completed datasets are created (Fig. 1), to capture the uncertainty in the missing values.

In the second stage, the epidemiological analysis of interest is performed on each of the 'completed' datasets (observed plus imputed values) (Fig. 1).[11] The final MI estimate is simply the average of the estimates (see Kasza and Wolfe[2] for a description of regression parameter estimates) derived from each of the completed datasets. The standard error of the MI estimate incorporates both the uncertainty in the estimate within the completed datasets and the uncertainty across the completed datasets due to the missingness (see Appendix for details). This ensures that MI produces a valid 95% confidence interval and *P*-value for the MI estimate of the regression parameter of interest.[11]

In the simplest case, when there are missing data in a single (continuous) variable, the imputation model consists of a linear regression model for the variable that has missing values (e.g. waist circumference in our example), regressed on the other variables to be used for imputation (e.g. the other variables to be used in the analysis plus the auxiliary variables). When there are missing data in a number of variables, there are currently two approaches for imputing the

missing values. One approach is to impute the missing values using a series of conditional regression models. Under this approach, a regression model is set up for each variable with missing data as described above, cycling through the regression models sequentially to impute missing values for each variable conditional on the imputed values for the other variables with missing data.[12,13] The second approach is to impute all of the variables with missing data simultaneously using a joint normal distribution.[14] Both of these MI approaches are available in standard computerized statistical packages (e.g. Stata[15] and SAS[16]).

WHEN MIGHT MI OFFER BENEFITS OVER A COMPLETE CASE ANALYSIS?

MI can offer gains over a complete case analysis in terms of reducing bias and/or improving precision.

Reducing bias

In some scenarios, there may be differences between participants with and without complete data, for example, those with asthma and/or allergies may be more motivated to attend follow-up visits for a respiratory study. In this context, carrying out a complete case analysis can lead to biased results as it will be based on a non-representative sample from the population.

If the reasons for missingness are known, that is, the missingness depends on observed data (e.g. age in our data example), but not on data that are unobserved, we refer to the data as missing at random. If data are missing at random, MI can use the observed data to fill in the missing values. Filling in the missing values enables all participants to be included in the analysis, hence correcting the bias in the complete case analysis. This bias correction is, however, only possible if there are auxiliary variables that can be included in the model used to impute the missing values (otherwise the imputation and analysis models are analogous). Data may also be missing not missing at random, that is the missingness may depend on unobserved data (e.g. unmeasured genetic factors). In this context, MI can still reduce bias compared with complete case analysis if there are auxiliary variables that are strong predictors of missingness, although some bias may remain as the imputed values are estimated from the observed data only.

Improving Precision

If the missingness occurs completely at random, for example, if the spirometer was not available to measure lung function for a random set of appointments, then a complete case analysis will be unbiased since it includes a random sample of the original study participants and hence a random sample from the population (assuming that the original sample was a random sample from the population). However, performing a complete case analysis still means throwing away information from study participants simply because they have missing data in that one (or more) variable(s), hence can be inefficient (i.e. can mean wide confidence intervals around parameter estimates). MI can improve efficiency (i.e. results in narrower confidence intervals) as it allows all participants to be included in the analysis.

The largest potential for efficiency gains from MI over the complete case analysis is when the variable(s) of interest are fully observed, such as in our motivating example, where the exposure of interest (asthma) and the outcome (FEV_1) are fully observed, but there are missing values in important confounders (waist circumference). In this scenario, excluding incomplete cases means that we are losing information about the exposure–outcome relationship in cases where the covariate is missing, information that can be recovered using MI. In contrast, if there are missing exposure or outcome data, then we are less likely to gain information about the association between exposure and outcome from MI, unless there are auxiliary variables that are very highly correlated with the incomplete variable, for example a similar outcome measured at a previous wave of data collection.[17,18]

The gain in efficiency from MI is due to the inclusion of auxiliary variables in the imputation model (see Section 2). The stronger the association between the incomplete and auxiliary variables, the more accurate the imputed values will be, and the larger the potential gains in precision from MI. In practice, there needs to be a reasonably strong correlation between the incomplete and auxiliary variable for MI to have notable gains over a complete case analysis,[19] and sometimes there are not many (if any variables) with such strong correlation.[20]

MI IS NOT A MIRACLE CURE

Although MI is intuitively appealing, it is not a miracle cure. Importantly, MI can introduce bias over a complete case analysis if not carried out appropriately.[17] Specifically, there are a number of decisions which must be made when setting up the imputation model (stage 1), which can affect the validity of the resulting inference:

1 How much missing data are there?—If there is a lot of missing data, any bias introduced by the decisions made in setting up the imputation model will be inflated as a large amount of data is being imputed from a potentially mis-specified model.[21]

2 Which variables to include in the imputation model?—It is important to include all variables that are in the analysis model in the imputation model, including any interaction terms and terms that describe a nonlinear association, for example quadratic or logarithm transformations.[19] Leaving these variables out of the imputation model can lead to biased results. As discussed in the previous section, it is also important to include auxiliary variables that aid information recovery.

3 How to include non-normally distributed continuous variables?—Both approaches to MI (described briefly under "What is MI?") assume normality for

continuous variables (at least conditionally on the other variables in the imputation model). This means that including the raw (i.e. original) scale values of any non-normally distributed variables (e.g. cytokine levels pmol/l, which often follow a highly skewed distribution) in the imputation model can produce imputed values that are quite different to the observed values. This, in turn, can have flow on effects to the inference obtained.[22] In contrast, it has been suggested that transforming data to improve normality prior to imputation can also lead to biased results in some scenarios.[23]

4 How to impute categorical variables when using a joint normal distribution?—Since this approach assumes normality for all variables in the imputation model, it is unclear how best to impute missing values for a categorical variable under this framework.[24,25]

5 How to impute and analyse variables that have a restricted range of values?—Some variables are restricted in the range of values that they can take, for example FEV_1 must be greater than 0 by definition and is also usually < 8 L. Again there are various approaches that have been suggested for the imputation and analysis of restricted range variables.[23,26,27]

All of the above issues (see Graham[19] for a detailed discussion) should be considered prior to imputation, and with respect to the dataset under investigation. The fact that this approach is so flexible means that it is desirable to have some expertise in the methodology of MI prior to using this approach and making these decisions. Furthermore, it is important to explore the sensitivity of the results of the epidemiological analysis of interest to the decisions made in the imputation process, which is reassuring if all imputation models lead to the same overall conclusion.

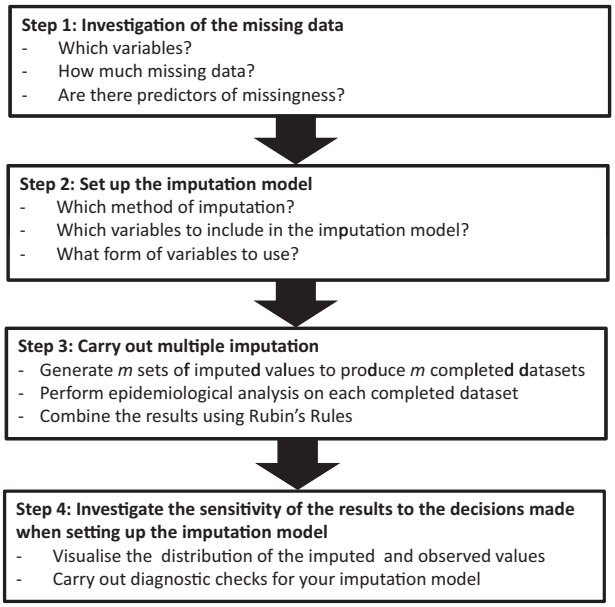

Figure 2 Process for carrying out multiple imputation.

RESULTS FROM OUR CASE STUDY

Figure 2 shows a flow chart of the steps that should be taken when carrying out MI.

In our case study, waist circumference was missing in 81 of the 316 participants in the analysis dataset (26%). Participants with available data on their waist circumference were slightly younger than those missing waist circumference (mean age 42.5 (standard deviation (SD) 0.5) years compared with 42.7 (SD 0.2) years, note at the fifth decade of follow-up participants were aged between 41 and 44 years), suggesting that a complete case analysis may lead to biased results.

In the observed data, mean waist circumference was greater for participants who had high blood pressure (99.3 (SD 14.9) cm) compared with those without high blood pressure (92.3 (SD 13.0) cm), and was greater for participants with high cholesterol (95.7 (SD 11.4)) compared with those without high cholesterol (93.2 (SD 14.2)). These findings suggest that high blood pressure and high cholesterol may be useful auxiliary variables for imputing waist circumference.

We present the results from our case study using a complete case analysis and MI. Analyses were conducted using Stata Release 12.[15] MI was carried out using the Stata command, 'mi impute regress'. Note that there was no need to decide whether to use sequential imputation or a joint normal model in this example as there was only one variable with missing data, hence imputation was carried out using a univariate linear regression model. Since waist circumference was approximately normally distributed, it was imputed using the original scale (i.e. no transformation was required). All variables in the analysis model were included in the imputation model, along with indicators for high blood pressure and high cholesterol. Other categorical variables were also included in the imputation model as a series of indicators (single indicators for current asthma, female and ever smoker, and four indicators for the socioeconomic status—a five-level ordinal variable). Twenty imputed datasets were generated, with the resulting estimates of the model parameters (from separate analyses of each completed dataset) combined using the Stata command, 'mi estimate'.[15]

Table 1 shows the results from our case study. In the complete case analysis, asthma was associated with a lower FEV_1 (regression coefficient –266 mL, standard error 84 mL). The estimate from MI shows a slightly stronger relationship, regression coefficient –294 mL, but importantly has decreased uncertainty (standard error = 74 mL), and hence increased precision, compared with the complete case analysis due to the recovery of the 81 cases with missing waist circumference measurements. This means a narrower 95% confidence interval and smaller *P*-value from the MI analysis. In contrast, the estimate of the association between waist circumference and FEV_1 (and the corresponding standard error) are similar in the two analyses, since there is less to be gained from MI (in terms of bias or precision) regarding this relationship.

Table 1 Results from the analysis of the Tasmanian Longitudinal Health Study

	Asthma			Waist circumference		
	Estimate (SE)	95% CI	*P*	Estimate (SE)	95% CI	*P*
Complete case analysis	−266 (84)	−431, −102	0.002	−6.0 (2.9)	−11.6, −0.3	0.04
Multiple imputation	−294 (74)	−439, −149	<0.001	−5.8 (2.9)	−11.6, 0.0	0.05

Estimates are regression coefficients from a multivariable linear regression model. The outcome variable is forced expiratory volume in 1 s (mL), and the covariates are asthma (exposure of interest), age (years), gender, socio-economic status, smoking status, height (cm) and waist circumference (cm; variable with 26% missing data).

CI, confidence interval; SE, standard error.

CONCLUDING REMARKS

We have presented MI as a useful approach for dealing with missing data. In particular, it can reduce bias and increase precision compared with a complete case analysis when there are additional observed variables associated with the variable with missing data. However, we note that this may not always be the case, particularly when the variables of interest (e.g. the outcome or the exposure of interest) have missing data.[17,28]

Although MI can be useful, we do not recommend it be used as a blanket approach for dealing with all missing data. In particular, MI can introduce bias not present in a complete case analysis if not carried out appropriately. Instead, we highlight the importance of MI as a sensitivity analysis surrounding the robustness of the results to the missing data. As demonstrated in our case study, it is reassuring when the complete case analysis and MI, both of which make different assumptions about the missingness, result in the same overall conclusion. We also recommend that the researcher performing MI be suitably familiar with the methodology of this approach before using MI.

When considering MI, it is important to first assess whether MI is likely to offer gains, in terms of either reducing bias or increasing precision, over a complete case analysis. Once it has been decided to use MI, the analyst needs to consider carefully the most appropriate imputation model, and the sensitivity of the approach to the decisions made during imputation. There is an increasing focus on diagnostic checks of the imputation model to help guide the researcher in building an appropriate model.[5,29] Finally, MI assumes that data are missing dependent on observed variables. In practice, missingness may depend on unobserved data. Given it is not possible to assess whether missingness depends on unobserved data, it may be important to assess the sensitivity of the analysis to the missing at random assumption when using MI.[30,31]

Acknowledgements

We thank the TAHS Steering Committee for providing us with a random subset of the data from the fifth decade of follow up of the TAHS cohort which was funded by the NHMRC (ID 299901). This work was supported by funding from the National Health and Medical Research Council: Career Development Fellowship ID 1053609 (K.J.L.), a Centre of Research Excellence grant, ID 1035261, awarded to the Victorian Centre for Biostatistics (ViCBiostat), and project grant 607400. Research at the Murdoch Childrens Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program.

REFERENCES

1 Wolfe R, Abramson M. Modern statistical methods in respiratory medicine. *Respirology* 2014; **19**: 9–13.
2 Kasza J, Wolfe R. Statistical regression models: interpretation of commonly-used models. *Respirology* 2014; **19**: 14–21.
3 Vermeulen KM, Post WJ, Span MM, van der Bij W, Koëter GH, TenVergert EM. Incomplete quality of life data in lung transplant research: comparing cross sectional, repeated measures ANOVA, and multi-level analysis. *Respir. Res.* 2005; **6**: 101–10.
4 Molenberghs G, Kenward MG. *Missing Data in Clinical Studies*. John Wiley and Sons Ltd, Chichester, 2007.
5 Van Buuren S. *Flexible Imputation of Missing Data*. CRC Press, Hoboken, 2012.
6 Sterne JAC, White IR, Carlin JB, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009; **338**: b2393.
7 Mackinnon A. The use and reporting of multiple imputation in medical research—a review. *J. Intern. Med.* 2010; **268**: 586–93.
8 von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche P, Vandenbroucke JP, STROBE Initiative. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies in epidemiology. *BMJ* 2007; **335**: 806–8.
9 Ware JH, Harrington D, Hunter DJ, D'Agostino RB. Missing data. *N. Engl. J. Med.* 2012; **367**: 1353–4.
10 Wharton C, Dharmage S, Jenkins M, Dite G, Hopper J, Giles G, Abramson M, Walters EH. Tracing 8,600 participants 26 years after recruitment at age seven for the Tasmanian Asthma Study. *Aust. N. Z. J. Public Health* 2006; **30**: 105–10.
11 Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, 1987.
12 Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.* 2001; **27**: 85–95.
13 VanBuuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat. Med.* 1999; **18**: 681–94.
14 Schafer JL. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London, 1997.
15 StataCorp. *Stata Statistical Software: Release 12*. StataCorp LP, College Station, TX, 2011.
16 SAS Institute Inc. *PROC MI. SAS Procedures Guide, Version 92*. SAS Institute Inc, Cary, NC, 2008.
17 Lee KJ, Carlin JB. Recovery of information from multiple imputation: a simulation study. *Emerg. Themes Epidemiol.* 2012; **9**: 3.

18 Marshall A, Altman D, Royston P, Roger L. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Med. Res. Methodol.* 2010; **10**: 7.
19 Graham J. *Missing Data: Analysis and Design*. Springer, New York, NY, 2012.
20 Karahalios A, Baglietto L, Lee KJ, English DR, Carlin JB, Simpson JA. The impact of missing data on analyses of a time-dependent exposure in a longitudinal cohort: a simulation study. *Emerg. Themes Epidemiol.* 2013; **10**: 6.
21 Rubin DB. Multiple imputation after 18+ years. *J. Am. Stat. Assoc.* 1996; **91**: 473–89.
22 Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *Am. J. Epidemiol.* 2010; **171**: 624–32.
23 von Hippel PT. Should a normal imputation model be modified to impute skewed variables? *Sociol. Methods Res.* 2013; **42**: 105–38.
24 Galati JC, Seaton KA, Lee KJ, Simpson JA, Carlin JB. Rounding non-binary categorical variables following multivariate normal imputation: evaluation of simple methods and implications for practice. *J. Stat. Comput. Simul.* 2012; **iFirst**: 1–14.
25 Lee KJ, Galati JC, Simpson JA, Carlin JB. Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study. *Stat. Med.* 2012; **31**: 4164–74.
26 Enders C. *Applied Missing Data Analysis*. The Guilford Press, New York, 2010.
27 Royston P, Carlin JB, White IR. Multiple imputation of missing values: new features for 'mim'. *Stata J.* 2009; **9**: 252–64.
28 White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat. Med.* 2010; **29**: 2920–31.
29 Abayomi K, Gelman A, Levy M. Diagnostics for multivariate imputations. *Appl. Stat.* 2008; **57**: 273–91.
30 Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Stat. Methods Med. Res.* 2007; **16**: 259–75.
31 Carpenter J, Kenward MG. *Multiple Imputation and Its Application*. WIley, Chichester, 2013.

APPENDIX—RUBIN'S RULES

The MI estimate of a parameter which we will denote β, in our example the regression coefficient for current asthma status in a multivariable linear regression model for FEV_1, is simply the average of the estimated β from the analysis of each of the m completed datasets (observed plus imputed data, Fig. 1):

$$\bar{\beta}_{MI} = \frac{1}{m}\sum_{i=1}^{m}\hat{\beta}_i$$

where $\hat{\beta}_i$ is the estimate of the parameter in the i^{th} ($i = 1, \ldots, m$) completed dataset. In our example with 20 completed datasets, the estimate of β (the regression coefficient obtained from the multivariable regression model) from the first completed dataset = –298, and then $\hat{\beta}_2 = -267$, and so on up to $\hat{\beta}_{20} = -287$, so that:

$$\bar{\beta}_{MI} = \frac{1}{20}\times((-298)+(-267)+\cdots+(-287)) = -294.$$

The standard error of the MI estimate, used to calculate the 95% confidence interval (CI) and P-value for $\bar{\beta}_{\text{MI}}$, is derived using the formula:

$$\text{SE}(\bar{\beta}_{\text{MI}}) = \sqrt{\bar{\text{V}} + \left(1+\frac{1}{m}\right)\text{B}}$$

where $\bar{\text{V}}$ is a measure of the uncertainty of the estimate within each of the imputed datasets, calculated as the average of the square of the standard errors of the $\hat{\beta}_{\text{i}}$'s derived from each of the i imputed datasets:

$$\bar{\text{V}} = \frac{1}{\text{m}}\sum\nolimits_{\text{i}=1}^{\text{m}}\left(\text{SE}\left(\hat{\beta}_{\text{i}}\right)\right)^2.$$

In our example, $\text{SE}\left(\hat{\beta}_1\right) = 72$, $\text{SE}\left(\hat{\beta}_1\right) = 73$, ... $\text{SE}\left(\hat{\beta}_1\right) = 73$, so that

$$\bar{\text{V}} = \frac{1}{20}(72^2 + 73^2 + \cdots + 73^2) = 5361.$$

And B is a measure of how far the estimates of β from each of the imputed datasets are from the overall MI estimate ($\bar{\beta}_{MI}$), representing the between imputation variability. This is calculated from the sum of the squared differences between $\hat{\beta}_i$ and $\bar{\beta}_{MI}$:

$$\text{B} = \frac{1}{m-1}\sum\nolimits_{i=1}^{m}\left(\hat{\beta}_i - \bar{\beta}_{MI}\right)^2.$$

In our example, $\hat{\beta}_1 = -298$, $\hat{\beta}_2 = -267, \ldots,$ $\hat{\beta}_{20}$ –287, and $\bar{\beta}_{MI} = -294$ so that:

$$\text{B} = \frac{1}{20-1}(-298-(-294))^2 + (-267-(-294))^2 + \cdots + (-287-(-294))^2 = 85.$$

Combing these gives the standard error of the MI estimate:

$$\text{SE}(\bar{\beta}_{MI}) = \sqrt{5361 + \left(1+\frac{1}{20}\right)\times 85} = 74.$$