STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:3227–3246
Published online 17 January 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3177

How should variable selection be performed with multiply imputed data?

Angela M. Wood[1,*,†], Ian R. White[2] and Patrick Royston[3]

[1]*Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Worts Causeway, Cambridge CB2 8RN, U.K.*
[2]*MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB1 2SR, U.K.*
[3]*MRC Clinical Trials Unit, 222 Euston Road, London NW1 2DA, U.K.*

SUMMARY

Multiple imputation is a popular technique for analysing incomplete data. Given the imputed data and a particular model, Rubin's rules (RR) for estimating parameters and standard errors are well established. However, there are currently no guidelines for variable selection in multiply imputed data sets. The usual practice is to perform variable selection amongst the complete cases, a simple but inefficient and potentially biased procedure. Alternatively, variable selection can be performed by repeated use of RR, which is more computationally demanding. An approximation can be obtained by a simple 'stacked' method that combines the multiply imputed data sets into one and uses a weighting scheme to account for the fraction of missing data in each covariate. We compare these and other approaches using simulations based around a trial in community psychiatry. Most methods improve on the naïve complete-case analysis for variable selection, but importantly the type 1 error is only preserved if selection is based on RR, which is our recommended approach. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: multiple imputation; variable selection; stepwise; multiply imputed data; stacked data

1. INTRODUCTION

The need for adequate handling of missing data in medical research is increasingly recognised, with implications both for main analyses and for sensitivity analyses [1]. Multiple imputation (MI) [2, 3] is a statistical technique for analysing such data and has become more popular because of its generality and recent software developments [4–11]. Most MI developments have focused on obtaining parameter estimates with an appropriate measure of uncertainty [2]. However, data analysts commonly perform other procedures including model selection, prediction, and diagnostic tests and graphical checks for model mis-specification. There are currently no guidelines for

*Correspondence to: Angela M. Wood, Strangeways Research Laboratory, Worts Causeway, Cambridge CB2 8RN, U.K.
†E-mail: amw79@medschl.cam.ac.uk

Received 21 December 2006
Accepted 7 November 2007

performing such procedures in multiply imputed data. In this paper we investigate methods for multivariable model-building involving selection of variables according to statistical criteria.

MI involves replacing each missing value with $M>1$ plausible values estimated from the observed data, resulting in M multiply imputed data sets. Commonly, $M=5$. Each imputed data set is analysed identically by complete-data methods. The multiple parameter estimates and their standard errors are combined to give an overall estimate and an overall standard error that includes both within-imputation and between-imputation components of variation ('Rubin's rules', RR [2]). Between-imputation variability reflects the uncertainty due to missing information.

Standard implementations of MI rely on an assumption that missing data are missing at random (MAR); that is, the missing values may depend on the observed data but not on the unobserved data [12]. A special case of MAR is missing completely at random (MCAR); that is, the missing values are independent of both observed and unobserved data. However, MI may be extended to allow for data that are not MAR [4, 13].

Variable selection is typically concerned with finding important predictors of an outcome variable. Several variable selection procedures have been proposed and used in the literature. In this paper we focus on the *backward stepwise selection* approach, which is appealing due to its popularity and availability in statistical software. However, it is often criticised [14]; problems include falsely narrow confidence intervals for effects and predicted values [15] and multiple hypothesis testing inflating risks of capitalising on chance features of the data [16], such as noise covariates gaining entry into the model when the number of candidate variables is large [17]. Such problems are exacerbated in small samples and give rise to the notion that selection of variables according to statistical criteria is inevitably flawed and dangerous. For 'reasonable' sample sizes and 'strong' effects of covariates, this interpretation is incorrect [18]. However, when selecting a set of variables, consideration should always be given to issues such as clinical importance, confounding, collinearity, and model stability [19], as well as statistical significance.

When some data are missing, as is increasingly likely as more variables are considered, MI may be used, but variable selection is typically performed only on the complete cases (CC). This approach is inefficient when there are many missing values and may bias type 1 error rates unless the data are MCAR. This paper compares CC variable selection with several alternatives including performing variable selection in each of the multiply imputed data sets and performing variable selection based on the parameter estimates and their standard errors combined according to RR [2]. We also consider model selection on the *stacked* data set of the MIs using weighted regression, a computationally easier approach. Bayesian methods of model selection [20] are not commonly used and are not considered in this paper. We describe the methods in Section 2. In Section 3, we illustrate the methods using data from the UK700 randomised controlled trial of psychiatric case management and assess the methods through simulation studies in Sections 4 and 5. Section 6 provides a discussion.

2. METHODS FOR MODEL SELECTION IN MI DATA SETS

2.1. Notation

Let Z_i be one of p complete or incomplete variables ($i=1,\ldots,p$) observed on a set of subjects, and let $Z=(Z_1,\ldots,Z_p)$. Denote the observed and missing values of Z_i by Z_i^{obs} and Z_i^{mis}, respectively, and let $Z^{\text{obs}}=(Z_1^{\text{obs}},\ldots,Z_p^{\text{obs}})$ and $Z^{\text{mis}}=(Z_1^{\text{mis}},\ldots,Z_p^{\text{mis}})$. Let R_i be the response indicator for

DOI: 10.1002/sim

Z_i, with $R_i=1$ if Z_i is observed and $R_i=0$ if Z_i is missing, and $R=(R_1,\ldots,R_p)$. We assume that the aim of the statistical analysis is to build a model in which one of the Z's will be a single outcome variable, Y, and the rest are potential predictor variables, X.

2.2. *Multiple imputation*

MI typically comprises three main stages:

Stage 1: *Generating multiply imputed data sets*: The unknown missing data Z^{mis} are replaced by independent simulated values drawn from the posterior predictive distribution $P(Z^{\text{mis}}|Z^{\text{obs}},R)$. Under MAR this is the same as $P(Z^{\text{mis}}|Z^{\text{obs}})$. Imputation is commonly done under a multivariate normal distribution for Z [3] or by iterated chained equations [4].

Stage 2: *Analysing multiply imputed data sets*: Once the MIs have been generated, each imputed data set is analysed identically by complete-data methods to estimate quantities of scientific interest. M analyses are performed and the results differ only because the imputations are different.

Stage 3: *Combining estimates from multiply imputed data sets*: The multiple estimates are combined into an overall inference that incorporates both within- and between-imputation variability using RR [2]. The rules were derived from a Bayesian framework. Suppose $\hat{\theta}_k$ is an estimate of a univariate or multivariate quantity of interest (e.g. a regression coefficient) obtained from the imputed data set k (for $k=1,\ldots,M$) and V_k is the estimated variance of $\hat{\theta}_k$. The combined estimate is the average of the individual estimates:

$$\bar{\theta}=\frac{1}{M}\sum_{k=1}^{M}\hat{\theta}_k \tag{1}$$

The overall variance of $\bar{\theta}$ is given by

$$\text{Within-imputation variance:}\quad \bar{W}=\frac{1}{M}\sum_{k=1}^{M}V_k \tag{2}$$

$$\text{Between-imputation variance:}\quad B=\frac{1}{M-1}\sum_{k=1}^{M}(\hat{\theta}_k-\bar{\theta})(\hat{\theta}_k-\bar{\theta})^{\mathrm{T}} \tag{3}$$

$$\text{Total variance:}\quad \text{Var}(\bar{\theta})=\bar{W}+(1+M^{-1})B \tag{4}$$

Further details can be found in Shafer (Chapter 4) [3]. The Wald test statistic for testing the null hypothesis $\theta=0$ is $\bar{\theta}^{\mathrm{T}}\,\text{var}(\bar{\theta})^{-1}\bar{\theta}$, but the likelihood ratio test cannot be directly implemented. Meng and Rubin [21] have, however, proposed an approximation for the likelihood ratio test, which is based on less information and is less accurate than the Wald test.

Each stage of the MI process is distinct and may be performed separately. RR for combining estimates were, however, derived under the assumption that the imputation and analysis stages are conditioned on the same set of observed data. This implies that all variables included in the analysis stage (Stage 2) should also be included in the imputation model (Stage 1). Failure to do so can lead to combined estimates biased towards the null. On the other hand, if variable relationships are correctly excluded from the analysis stage but not from the imputation model, the combined estimates will still be valid, but combined interval estimates will be wider to reflect the extra degree of uncertainty in the imputations [3].

DOI: 10.1002/sim

2.3. Variable selection in complete data

Variable selection procedures are required when there is no known or established model or when a large complex model would be unstable or difficult to interpret. There are two main classes of model-building procedures. 'Classical' variable selection [22] is based on hypothesis tests between nested models using Wald tests, likelihood ratio tests, and F statistics. Variables are selected or eliminated according to their statistical significance in the current model. Penalised estimation procedures attempt to identify a model that optimises a likelihood penalised for model complexity, typically using the Akaike or Bayesian information criterion.

Classical variable selection is usually implemented through computational procedures such as *forward*, *backward*, and *stepwise* selection. These approaches are most appropriate when the initial set of potential variables is fairly large, despite being restricted on the basis of clinical knowledge. *Forward selection* is the simplest approach. Starting with a null model, each variable that is not in the model is tested for inclusion in the model. The most significant of these variables is added to the model, so long as it meets some pre-specified significance level. The process continues until no remaining variable is significant when added to the model. A drawback is that some included variables in the model may become non-significant on the addition of a new variable. An alternate approach is *backward selection*. Starting with a model with all potential variables, each variable in the model is tested for exclusion from the model. The least significant variable in the model is dropped so long as it is not significant at some pre-specified critical level. The process continues until all variables in the model are significant. However, some variables may be dropped that would be significant when added to the final reduced model. *Stepwise selection* provides a compromise between forward and backward selection methods. The *backward stepwise selection* procedure consists of backward selection, followed by forward selection and iterates if necessary. Starting with a model with all potential variables, the backward selection process removes from the model the variable that is least statistically important (at α per cent significance). The forward selection process checks whether removed variables should be added back into the model (at α_2 per cent significance, usually $\alpha_2 = \alpha(1-\varepsilon)$ for ε small). These selection procedures use multiple testing and are data driven. The importance of a variable whose effect is relatively weak in a selected model may be overestimated (selection bias). However, estimates for 'strong' variables (i.e. with large standardised regression coefficients) are typically unbiased. Correlated variables may cause problems in stepwise procedures due to an inflated probability that the 'wrong' variable is selected.

2.4. Variable selection in multiply imputed data

Variable selection procedures need adaptation when applied to multiple imputed data. The Rubin inferential framework provides Wald tests for average parameter estimates obtained at Stage 3. Thus, each model selection step involves fitting the model under consideration to all imputed data sets (MI Stage 2) and combining estimates across imputed data sets (MI Stage 3). For large data sets and large M, this process may not be computationally feasible. We, therefore, compare this 'RR' approach with four simpler alternatives described below.

2.4.1. Complete cases (CC). The CC data set, consisting of all individuals with non-missing values for all potential predictors, may be used for variable selection. Using a single data set reduces the number of models fitted and avoids having to combine estimates across imputations.

DOI: 10.1002/sim

2.4.2. Single stochastic imputation (Single). A single-imputed data set may be used for variable selection, not only maintaining the convenience of dealing with a single data set but also avoiding the inefficiency in the CC approach caused by excluding individuals. We use the first data set created by the MI procedure.

2.4.3. Separate imputations (S1, S2, and S3). Rather than choosing one of the imputed data sets for model selection, better use may be made of the data by performing model selection separately in each imputed data set. This approach will typically result in models with different selected predictors. We propose three strategies to provide a single selected model across the imputed data sets:

S1: select predictors that appear in any model;
S2: select predictors that appear in at least half of the models;
S3: select predictors that appear in all models.

In comparison to the RR approach, roughly the same number of models needs to be fitted, but the models are only combined once at the end of model selection, rather than at each step.

2.4.4. Stacked imputed data sets with weighted regression (W1, W2, and W3). Stacking the M imputed data sets for the n individuals yields one large data set of length Mn. Fitting models to this single stacked data set yields valid parameter estimates (see Appendix) but standard errors that are too small. A simple way to correct the standard errors is to apply a fixed weight to all individuals. Denote this weight by w_i in an analysis assessing inclusion or exclusion of variable X_i. We consider three possible sets of weights (see Appendix A):

W1: $w_i = 1/M$. These weights scale the log likelihood for the stacked data to the equivalent of a data set of length n but ignore the degree of missing information.
W2: $w_i = (1-f)/M$ where f, the average fraction of missing data across all variables, is calculated as (total number of missing values across all variables) divided by pn.
W3: $w_i = (1-f_i)/M$ where f_i, the fraction of missing data for variable X_i, is calculated as (number of missing values for variable X_i) divided by n.

Using weighted regression, variable selection may be performed using suitable hypothesis tests (likelihood ratio or Wald). A possible advantage of the stacked approach over the RR approach is that likelihood ratio test statistics, which are usually preferred to Wald statistics for non-linear regressions and/or small samples, are easy to obtain [23].

2.5. Properties of methods

The method RR can be considered as the gold standard approach but is the more computationally intensive method. CC is the most commonly adopted approach due to its simplicity but will produce biased estimates in the case of MAR and will be inefficient in cases of MCAR. Single will not produce biased estimates, but the standard errors will be too small, resulting in an increased chance of accepting noise variables into the model. A further disadvantage of this method is that a different model may be selected depending on which imputation is used. Methods S1–S3 are also fairly computationally intensive; S1 is likely to select too many variables; S3 may select too few; and S2 should be more reasonable and give solutions comparable to Single. W1 is also similar to Single and S1 in that it is unlikely to be biased, but the standard errors will be too small because

DOI: 10.1002/sim

it ignores the uncertainty caused by the missing information. Finally, W3 (and W2 in very specific conditions) should give solutions comparable to RR in case of MCAR.

2.6. Practical implementation

Once a model has been selected, the final model is fitted to all imputed data sets and the results combined using RR (MI Stage 3). It would be possible to update the imputation model to account for the change in the analysis model. We return to this issue in the Discussion. The merits of the MI variable selection approaches described above are assessed in our case study in Section 3 and via a simulation study in Sections 4 and 5.

2.7. Software implementation

All procedures were implemented using STATA 9.2. MI was performed using the **ice** package [6–8], which uses the iterated chained equations approach [4]. This requires specification of conditional models for each incomplete variable given all other variables. In total, 50 imputed data sets were created, each using 10 cycles of regression switching (results are later shown using all 50 imputations and a random subset of 5 imputations). Imputations for each variable were drawn from a Normal approximation to the posterior distribution from the corresponding conditional model. Parameter estimates from any regression model were combined across imputations by the **micombine** command which applies RR.

Model selection approaches CC, Single, S1, S2, and S3 were performed using the built-in **stepwise** procedure in STATA using the default Wald test option. Weighting methods W1 and W2 were also implemented using **stepwise**; the weights were specified as 'importance' weights. For the RR method, **stepwise** was modified to use the Wald test statistics from **micombine**. The W3 weight scheme could not be implemented using **stepwise** because the weights depend on the variable being tested. Instead, we used **mpfmi**, a user-written procedure described by Royston *et al.* [23]. Although **mpfmi** is designed to extend the multivariate fractional polynomial approach [24] for use with multiply imputed data, we used it to fit linear terms only. Variables are tested using weighted likelihood ratio tests, but in the sample sizes used we would expect similar results using Wald tests.

In practice, various values for α might be used (e.g. $\alpha = 5, 10, 15.7$ or 20 per cent, etc.), chosen depending on sample size and the aim of the model-building process. For our main method comparison, we chose significance levels $\alpha = 5$ per cent for backward steps and $\alpha_2 = 4.9$ per cent for forward steps. As a sensitivity analysis we also consider $\alpha = 15.7$ per cent and $\alpha_2 = 15.6$ per cent: these values select a model maximising the Akaike information criterion when all variables have 1 degree of freedom [18].

3. CASE STUDY: THE UK700 TRIAL

3.1. Motivation

We use data from the UK700 trial to illustrate the variable selection approaches for multiply imputed data. The data have small-to-moderate proportions of missing values in the covariates (up to 20 per cent missing) and no unusual or special features.

DOI: 10.1002/sim

Table I. Summary of variables in the UK700 trial.

Variable	Data type	Code	Observations ($n=708$)
Outcomes			
The Comprehensive Psychopathological Rating Scale (CPRS) at 2 years	Continuous	`cprs`	585
(Dis)satisfaction with case management at 2 years	Continuous	`sat`	490
Time to loss of contact with case manager	Survival	`contactloss`	708 (500 censored)
Baseline variables			
Comprehensive Psychopathological Rating Scale	Continuous	`cprs0`	705
(Dis)satisfaction with case management	Continuous	`sat0`	571
Centre	Categorical	`centre` (centres 1–4)	708
Total disability	Continuous	`distot`	659
Time from onset of psychosis to study entry (months)	Continuous	`onset`	705
Age (years)	Continuous	`age`	708
Sex	Binary	`sex` (female = 0, male = 1)	708
Ethnic group	Binary	`ethnic` (Afro-Caribbean = 1, other = 0)	708
Randomised group	Binary	`group` (standard = 0, intensive = 1)	708
Outpatient status	Binary	`status` (recruited in hospital = 0, recruited as outpatient = 1)	707
Father's occupation at birth	Ordered categorical	`occgp` (scores 1–6)	576

3.2. Description of the data

The UK700 trial was a multi-centre randomised controlled trial conducted in four inner-city areas [25]. Participants were aged 18–65 with a diagnosed psychotic illness and had at least two psychiatric hospital admissions, the most recent within the previous 2 years. In the U.K., such patients are typically managed in the community by a case manager. In the trial, 708 participants were randomly allocated to a case manager with either a case load of 30–35 patients (standard case management) or a case load of 10–15 patients (intensive case management). The trial outcomes and baseline characteristics used in this paper are summarised in Table I. The main trial findings have been previously reported [25].

We aim to identify predictors of satisfaction with case management after 2 years and illustrate the differences between the MI variable selection approaches. We apply each MI variable selection procedure described in Section 2.4 to the following baseline predictors chosen for their potential clinical relevance: baseline satisfaction, baseline CPRS, centre, total disability, time since onset of psychosis, age, sex, father's occupation, ethnic group, randomised group, and outpatient status. MI is a sensible approach for these data because of the substantial proportions of missing data in several baseline variables (up to 30 per cent missing values) (Table I).

DOI: 10.1002/sim

3.3. Multiple imputation

To ensure that all continuous variables were approximately normally distributed (required for the imputation models), we transformed `distot` and `onset` logarithmically before imputation, and transformed imputed values back to the original scale. We specified linear, logistic, and ordered logistic models for imputing variables identified in Table I as continuous, binary, and ordered categorical, respectively. Each imputation model included the outcome variable and all baseline variables, except the one being imputed, from all 708 individuals. Continuous and ordered categorical variables were entered linearly and dummy variables used for binary and unordered categorical variables.

3.4. Variable selection

Cases with missing outcome Y contain no information on the parameters of the regression model. Thus we performed the MI variable selection procedures on the sub-sample of 490 individuals with observed outcomes, with the randomised group kept in the model. Table II illustrates the variables selected by each different MI variable selection for $M=5$ and 50. The parameter estimates shown were obtained by linear regression on complete cases for CC and by using Rubin' approach on M imputed data sets for all other procedures.

The variables selected were fairly similar across the different MI variable selection methods (Table II). Variables `onset`, `occgp`, `ethnic`, `status`, and `distot` were eliminated from all models. A model with variables `sat0`, `age`, `centre`, and `cprs0` was favoured by most approaches. The directions of the effects suggest that those more satisfied at baseline, younger,

Table II. Parameter estimates (SEs) of baseline variables selected* into a linear model for satisfaction at 2 years using 409 individuals with observed outcome.

		Baseline variables						
					centre			
		group	sat0	age	2	3	4	cprs0
Observed *n*		490	410	490	490	490	490	490
Method†	*M*							
CC	—	−0.54	0.25	−0.06				0.05
		(0.41)	(0.05)	(0.02)				(0.02)
Single/S1/S2/	5	−0.48	0.25	−0.06	−1.38	0.25	0.35	0.03
W1/W2/W3/RR‡		(0.41)	(0.05)	(0.02)	(0.61)	(0.60)	(0.61)	(0.02)
Single/S1/S2/	50	−0.45	0.25	−0.06	−1.35	0.25	0.36	0.04
W1/W2/W3RR‡		(0.41)	(0.05)	(0.02)	(0.60)	(0.59)	(0.61)	(0.02)
S3	5	−0.35	0.26	−0.07	−1.49	0.36	0.49	
		(0.41)	(0.05)	(0.02)	(0.60)	(0.59)	(0.61)	
S3	50	−0.38	0.27	−0.07	−1.51	0.35	0.46	
		(0.41)	(0.05)	(0.02)	(0.61)	(0.60)	(0.61)	

*Baseline variables `onset`, `occgp`, `ethnic`, `status`, and `distot` were not selected by any approach.
†As defined in Section 2.
‡Methods grouped together all selected the same model.

DOI: 10.1002/sim

from centres 1, 3, or 4 and a higher baseline CPRS tend to be more satisfied with case management at 2 year follow-up. The model without `cprs0` was selected by S3, and the model without a centre effect was selected using CC.

4. SIMULATION PROCEDURE

The aim of our simulation studies is to compare and assess the merits of the variable selection methods described in Section 2.4. Unlike in the case study, the procedures are assessed on simulated data for which we know the exact true model. Thus, each model selection method is formally assessed by its ability to correctly select variables from the *true* model and its failure to wrongly select variables not in the *true* model.

Our simulation strategy is the following:

1. Start with a set of complete explanatory variables denoted by X.
2. Simulate the outcome variable Y (continuous, binary, or time to event) from a pre-defined *true* model consisting of a subset of X.
3. Induce missing data in X under a variety of missing data patterns and assumptions.
4. Multiply impute missing data in X.
5. Perform the model selection procedures.
6. Repeat Steps 2–5 1000 times, and compare results across the simulations.

Results are presented in Section 5 from seven simulation scenarios, which differ according to the outcome variable type (continuous, binary, or time to event), sample size, and pattern of missing data. These scenarios and each of the strategy steps are described in detail in the subsections below. In all our simulation studies Y is fully observed.

4.1. Complete data: the UK700 trial

The simulations are based on the UK700 trial outcomes (A) CPRS at 2 years and (B) time to loss of contact with case manager. We used nine baseline variables from Table I: baseline CPRS, centre, total disability, time from onset of psychosis to study entry, age, sex, ethnic group, randomised group, and outpatient status. The outcomes and corresponding baseline variables were chosen because they are reasonably complete.

A single completed data set was formed by imputing missing values in variables `cprs`, `cprs0`, `distot`, `onset`, and `status` using a single run of STATA's `ice` program as described in Sections 2.7 and 3.3. All baseline variables and the CPRS outcome at 2 years were included in the imputation models. Note that the method used to impute the missing values at this stage has little impact on the simulation results, although the completed data may be slightly more normalised compared with real data. The resulting completed data set forms the basis of our simulation studies.

4.2. Simulation models for outcome Y

Y values were simulated from three different models, each of which was derived from the completed UK700 data set. The first model is based on a linear model for CPRS outcome. Backward stepwise linear regression model selection identified six baseline variables. The fitted model was as

DOI: 10.1002/sim

follows:

Linear model:

$$\begin{aligned}\texttt{cprs} = {} & 9.9 + 0.33\texttt{cprs0} - 0.14\texttt{age} + 1.7\texttt{onset} + 2.8\texttt{distot} - 2.1\texttt{sex} \\ & - 0.26\texttt{centre2} + 4.6\texttt{centre3} - 1.3\texttt{centre4} + \mathrm{N}(0, 12) \end{aligned} \quad (5)$$

The second model is based on a logistic model for dichotomised CPRS outcome. A cut-off value of CPRS>25 was used to produce a binary outcome with 25 per cent of individuals assigned a value of one. Using the six baseline variables selected above yielded the fitted model:

Logistic CPRS model:

$$\begin{aligned}\mathrm{logit}(P(\texttt{cprs}>25)) = {} & -1.69 + 0.03\texttt{cprs0} - 0.03\texttt{age} + 0.21\texttt{onset} + 0.58\texttt{distot} \\ & - 0.20\mathrm{sex} - 0.20\texttt{centre2} + 0.90\texttt{centre3} - 0.90\texttt{centre4} \end{aligned} \quad (6)$$

Thirdly, a backward stepwise Cox regression model selection identified four predictors of the time to loss of contact with the case manager. The fitted model was as follows:

Survival model:

$$\begin{aligned}\mathrm{Hazard}(\texttt{contactloss}) = {} & (\text{baseline hazard}) \times \exp(-0.03\texttt{age} + 0.07\texttt{centre2} \\ & - 2.02\texttt{centre3} - 0.21\texttt{centre4} + 0.55\texttt{group} \\ & + 0.56\texttt{status}) \end{aligned} \quad (7)$$

Thus, each simulated data set consists of observed/imputed values for baseline CPRS, centre, total disability, time from onset of psychosis to study entry, age, sex, ethnic group, randomised group and outpatient status and simulated values for continuous CPRS, dichotomised CPRS or time to loss of contact. The observed outcome values are no longer used.

4.3. Inducing missing data

To fully explore the variable selection proposals we consider four patterns of missing data.

4.3.1. Independent MCAR, equal fractions of missing data. We deleted a random 10 per cent of the data in each baseline variable, independent of the data values and independent of missingness in other variables.

4.3.2. Independent MCAR, unequal fractions of missing data. This is the same as above, except that the fraction of missing data is increased to 50 per cent in variables `ethnic` and `onset`. This pattern of missing data allows us to specifically explore the assumptions required for the stacked approaches W1–W3.

4.3.3. Monotone MCAR. We created three random subgroups of size 10 per cent of the data. We delete data on `onset` in the first subgroup, data on `ethnic` in the first two subgroups, and data on `distot` in all three subgroups. This is the same missing data pattern as for MAR (below) and enables a clear comparison of variable selection approaches between MCAR and MAR assumptions.

DOI: 10.1002/sim

4.3.4. Monotone MAR. We constructed an MAR missing data model based on a logistic model for the probability of a further variable, father's occupation (`occgp`) being missing. `occgp` was missing in 19 per cent of patients in the study. Missing data in covariates `onset`, `ethnic`, and `distot` were imposed in the following stages:

(1) Construct a logistic model for missing `occgp` regressing on all variables except `onset`, `ethnic`, `distot`, and dummy variables for `centre`. Impose missing data in `onset`, `ethnic`, and `distot` for individuals with fitted probabilities falling in the top tenth of probabilities.
(2) Construct a logistic model for missing `occgp` regressing on all variables except `ethnic`, `distot`, and dummy variables for `centre` and excluding individuals with missing values imposed by stage 1. Impose missing data in `ethnic`, and `distot` for individuals with fitted probabilities falling in the top ninth of probabilities.
(3) Construct a logistic model for missing `occgp` regressing on all variables except `distot` and dummy variables for `centre` and excluding individuals with missing values imposed by stages 1 and 2. Impose missing data in `distot` for individuals with fitted probabilities falling in the top eighth of probabilities.

Centre was excluded from the missing data model because it was highly predictive of missing data in `occgp` and we wanted to produce an MAR pattern more equally dependent on the other covariates. Our strategy produced the same pattern of missing data as 4.3.3 above, but with a different mechanism.

4.4. MI using chained equations

MIs were created using the methods described in Sections 2.7 and 3.3. Each imputation model included all nine candidate baseline variables except the one being imputed, and the corresponding outcome variable i.e. `cprs`, I(`cprs`>25), or `contactloss`. In the latter case the time-to-event outcome was log-transformed and the censoring indicator was included in the imputation model [4].

4.5. Model selection methods

Each model selection method described in Section 2.4 was performed on each simulated data set. For comparison, we also performed model selection on the *full* data set before missing data were induced.

4.6. Summary of the simulation studies

The seven simulation scenarios are summarised in Table III. Scenarios 1–3 will illustrate the performance of the variable selection procedures for linear and common non-linear regression. Scenario 4 will illustrate the potential problems in the proposed stepwise methods in small sample sizes. Scenarios 5–7 will compare and assess the variable selection approaches under different missing data assumptions. For each scenario, we ran each model selection on 1000 simulated data sets and recorded which baseline variables were selected or rejected by each model selection method. The model selection methods are assessed and compared using the percentages of simulated data sets in which each baseline variable is selected into the model. We use the terms 'power' to indicate the probability that a method will correctly select a given variable from the true model and 'type 1 error' to indicate the probability that a method will wrongly select a given variable not

DOI: 10.1002/sim

Table III. Summary of the seven simulation studies.

Simulation scenario	Model	Missing data	Sample size
1 (lin)	Linear model: equation (5)	Independent MCAR, equal fractions of missing data	708
2 (bin)	Binary model: equation (6)	Independent MCAR, equal fractions of missing data	708
3 (surv)	Survival model: equation (7)	Independent MCAR, equal fractions of missing data	708
4 (lin: 100)	Linear model: equation (5)	Independent MCAR, equal fractions of missing data	100
5 (lin: unequal f)	Linear model: equation (5)	Independent MCAR, unequal fractions of missing data	708
6 (lin: mono MCAR)	Linear model: equation (5)	Monotone MCAR	708
7 (lin: mono MAR)	Linear model: equation (5)	Monotone MAR	708

from the true model. For example in simulation scenarios 1 and 4–7, of interest is the method's power defined as the probability of correctly selecting each of the six baseline variables in true linear model (5) and the method's type 1 error defined as the probability of wrongly selecting each of the three baseline variables (ethnic group, randomised group, and outpatient status) not in the true linear model (5). Summarising further, we obtain the average power (e.g. mean powers over the six variables in the true linear model (5)) and the average type 1 error (e.g. mean type 1 errors over the three variables not in the true linear model (5)).

A good method should have type 1 error close to that observed using the full data (which may differ from the nominal significance level used in the model selection procedures). It should also have high power. We thus compare our approaches with the results using the full data in the following section.

5. SIMULATION RESULTS

5.1. Simulation scenario 1 (lin)

The percentage of the simulations in which each baseline variable was selected by each model selection method is shown in Table IV. The average power and average type 1 error for each method ($M = 5$) are displayed in Figure 1(a). Using the full data set produces an average power of 0.84 (range = 0.61–1) and an average type 1 error of 0.052 (range = 0.050–0.054). In comparison, model selection on CC has low power (average = 0.55, range = 0.31–1) and slightly inflated type 1 error (average = 0.060, range = 0.044–0.073) due to the exclusion of on average 61 per cent of individuals. In Figure 1(a), all other proposed methods fall approximately on a curve running from the bottom left corner (average power = average type 1 error = 0) to the top right corner (average power = average type 1 error = 1), like a receiver operating characteristic curve and showing a

DOI: 10.1002/sim

Table IV. Percentages of simulations in which each variable was selected into the linear model: independent MCAR and equal fractions of missing data (simulation scenario 1).

		Variables in true model				Variables not in true model				
Method*	*M*	cprs0	centre	distot	onset	age	sex	ethnic	group	status
Averaged observed *n*		637	637	637	637	637	637	637	637	637
Full	—	100	98.8	98.8	74.6	71.8	60.7	5.4	5.2	5.0
CC	—	99.9	66.0	68.4	34.6	33.0	30.8	6.2	4.4	7.3
Single	1	100	97.7	97.2	71.3	69.2	58.4	7.9	7.8	10.3
S1	5	100	99.3	99.0	85.2	84.3	76.1	16.7	16.1	19.7
	50	100	99.9	99.8	94.0	93.5	85.4	31.8	30.8	33.9
S2	5	100	98.2	97.5	72.8	69.9	60.7	7.4	7.2	7.7
	50	100	98.4	98.1	75.3	73.7	59.9	6.9	7.0	8.1
S3	5	100	94.3	93.2	54.8	51.8	43.2	2.5	3.5	3.1
	50	100	88.0	87.3	38.1	37.0	27.7	0.5	1.2	1.1
W1	5	100	97.9	97.5	74.1	72.5	60.3	7.4	7.7	8.4
W2	5	100	96.4	97.1	70.6	69.0	56.1	6.5	6.2	6.2
W3	5	100	96.6	96.8	68.7	67.6	56.4	6.1	6.4	6.5
RR	5	100	96.4	96.6	67.9	65.0	55.1	5.1	5.1	5.8

*As defined in Section 2.

trade-off between average power and average type 1 error. The RR approach produces type 1 error rates that are closest to the results obtained using the full data, although the results from the weighting methods W2 and W3 are also reasonably close. Results from the methods S2, W1, and Single all have average powers close to 0.84 but have inflated average type 1 errors. The two methods at the extremes are S3 and S1. The S3 method is too stringent in selecting variables, whereas the S1 method selects too many variables. The performance of these two approaches decreases as the number of imputations increases (Table IV). No other method considered was affected by the number of imputations (results not shown).

As a sensitivity analysis, we repeated this simulation using a nominal type 1 error $=0.157$. The results are shown in Figure 2. As expected, the overall pattern of results is similar to that found in Figure 1(a), with the powers shifted towards 1.

5.2. Simulation scenario 2 (bin)

Similar patterns are found in the results from the binary model (Figure 1(b)), although average powers are generally lower. Note methods W2 and W3 appear identical, as do methods S2 and W1, although on closer inspection the powers and type 1 error probabilities for the separate variables differ slightly.

5.3. Simulation scenario 3 (surv)

The results from the survival model (Figure 1(c)) follow a similar pattern to those seen from the linear and binary model. However, an increased type 1 error rate is observed in all methods because of a strong correlation ($r=0.62$) between a variable in the true model (`age`) and one not in the true model (`onset`) frequently leads to `onset` being selected instead of age (Table V). We take

DOI: 10.1002/sim

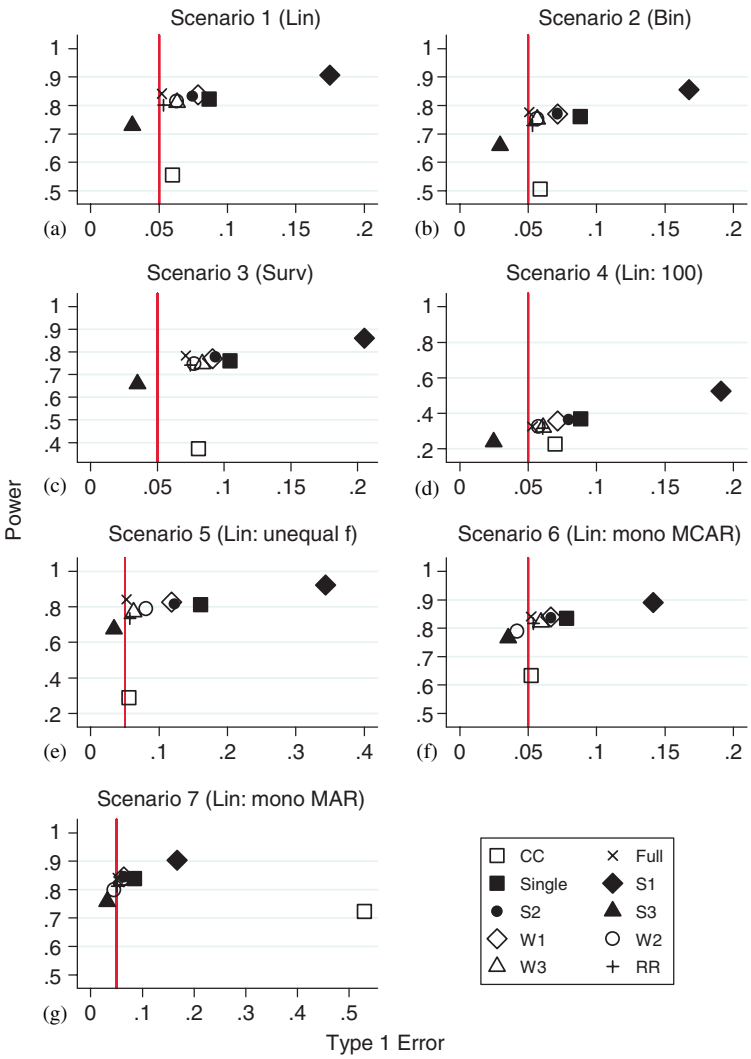

Figure 1. Type 1 error rate and power of the MI variable selection approaches under seven different simulation scenarios ($M=5$) (see Table III for a description of the simulation scenarios). Nominal type 1 error rate $=0.05$.

the (increased) type 1 error rate seen with the full data as a benchmark and note that methods RR, W2, and W3 have good type 1 error rates.

5.4. Simulation scenario 4 (lin: 100)

Reducing the sample size of the data set to $n=100$ increases the estimated parameter standard errors in all considered methods, reducing the average power (Figure 1(d)).

DOI: 10.1002/sim

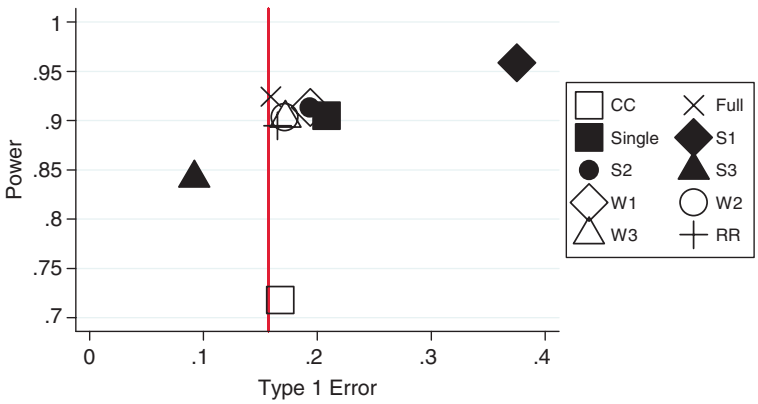

Figure 2. Type 1 error rate and power of the MI variable selection approaches under simulation scenario 1 (see Table III) ($M=5$). Nominal type 1 error rate $=0.157$.

Table V. Percentages of simulations in which each variable was selected into the survival model: independent MCAR and equal fractions of missing data (Simulation Scenario 3).

		Variables in true model				Variables not in true model				
Method*	M	centre	age	group	status	onset	distot	sex	cprs0	ethnic
Averaged observed n		637	637	637	637	637	637	637	637	637
Full	—	98.5	79.6	68.1	67.4	12.9	6.4	6.2	5.8	4.3
CC	—	40.0	48.6	33.2	27.1	14.8	6.5	4.9	7.9	6.4
Single	1	97.6	76.9	66.5	63.6	17.5	10.5	7.1	7.9	9.2
S1	5	98.5	88.9	80.0	77.8	31.7	21.6	16.1	17.3	15.7
	50	98.8	91.3	83.7	80.7	39.1	26.2	20.0	23.2	18.8
S2	5	97.2	79.7	68.1	66.2	14.5	9.6	7.9	7.9	6.8
	50	97.0	80.4	69.3	68.1	16.0	10.4	8.4	7.9	7.0
S3	5	96.3	63.0	53.5	50.8	5.9	3.7	2.2	3.6	2.1
	50	95.7	57.8	48.9	45.2	3.9	2.1	1.5	2.5	1.5
W1	5	97.6	77.6	66.4	66.6	16.2	8.6	6.9	7.8	6.2
W2	5	97.5	76.5	62.8	62.1	15.0	7.3	5.0	6.5	5.0
W3	5	99.7	73.9	63.6	62.3	18.0	6.8	5.0	7.2	4.8
RR	5	96.1	76.5	63.0	61.4	14.3	7.4	4.8	5.9	4.9

*As defined in Section 2.

5.5. *Simulation scenario 5 (lin: unequal f)*

Figure 1(e) shows the effect of increasing the amount of missing data to 50 per cent in variables `onset` and `ethnic`. Owing to the decreased sample size amongst the complete-case analysis, this method has lower average power than seen in simulation study 1 (Figure 1(a)). Of most interest is the difference in the results between weighting methods W2 and W3. W2 produces standard errors too large for variables with more missing data and standard errors too small for variables with less missing data. As a result, variable `ethnic` is selected into the model on more occasions,

DOI: 10.1002/sim

Table VI. Percentages of simulations in which each variable was selected into the linear model: independent MCAR and unequal fractions of missing data (simulation scenario 5).

		Variables in true model				Variables not in true model				
Method*	*M*	cprs0	centre	distot	onset	age	sex	ethnic	group	status
Averaged observed *n*		637	637	637	354	637	637	354	637	637
Full	—	100	98.8	98.8	74.6	71.8	60.7	5.4	5.2	5.0
CC	—	85.0	22.4	26.7	13.6	12.8	13.5	6.3	5.7	4.8
Single	5	100	96.9	96.9	66.5	67.6	60.3	26.4	9.9	11.9
S1	5	100	99.4	99.0	90.5	87.3	77.8	62.1	17.9	23.0
	50	100	99.9	99.8	99.5	97.3	87.0	99.1	33.1	42.7
S2	5	100	97.5	97.8	66.6	67.6	60.3	20.6	7.5	8.7
	50	100	98.1	98.0	69.4	69.6	61.6	18.5	7.3	8.3
S3	5	100	93.3	93.1	36.2	40.9	42.5	4.0	2.7	3.5
	50	100	85.8	84.6	12.4	18.7	24.7	0.3	1.0	0.7
W1	5	100	97.6	97.7	70.6	71.0	59.5	19.5	6.8	9.1
W2	5	100	95.4	96.2	65.4	65.5	51.5	14.2	4.5	5.5
W3	5	100	96.6	97.0	51.7	60.8	56.9	5.8	5.0	8.0
RR	5	100	95.7	96.1	43.4	49.2	56.9	5.0	5.1	7.1

*As defined in Section 2.

whereas variables rand and status are selected on fewer occasions (Table VI) when compared with simulation study 1 (Table IV). The type 1 errors for W3 are close to nominal.

5.6. Simulation scenario 6 (lin: mono MCAR)

Results for Scenario 6 (Figure 1(f)) are similar to those for Scenario 5, again showing the inability of weighting method W2 to allow for unequal fractions of missing data. Standard errors for variables without missing data are substantially increased, thus reducing sensitivity and type 1 error below those observed from using RR and breaking the consistent pattern in the results found in scenarios 1–5.

5.7. Simulation scenario 7 (lin: mono MAR)

Results under the MAR assumption (Figure 1(g)) are comparable to those under the MCAR assumption (Figure 1(f)) with one exception. Unlike the MI approaches, the complete-case analysis does not allow for MAR and thus produces biased parameter estimates and large standard errors, leading to grossly inflated type 1 error (Figure 1(g)).

6. DISCUSSION

In this paper, we have described and compared various methods for variable selection using stepwise procedures in multiply imputed data. A good method should have the power to detect true predictors and should have type 1 error close to what would be achieved if there were no missing data (which may not equal the nominal type 1 error rate if important correlations are present, as

DOI: 10.1002/sim

in scenario 3). Our recommended approach for performing stepwise methods in multiply imputed data is based on RR, as it is the only approach to preserve the type 1 error. We recommend against using CC, as it fails to detect important predictors due to a lack of power and, when missing data are not MCAR, may select unimportant variables due to biased regression estimates. In other methods, there is a trade-off between power and type 1 error. Since RR is a multi-stage iterative process it may be impractical when multiple outcomes are of interest or numerous possible interaction terms are to be assessed. Our proposed stacking method with appropriate weights is a pragmatic alternative. Next we consider its properties in more detail.

A regression analysis of a single stacked data set, consisting of M imputations, produces unbiased estimates of regression coefficients and, with appropriate weights, may produce variance estimates that approximate those produced according to MI theory. Our simulation results showed that compared with Rubin's approach, the stacking approach for MI variable selection generally has more power compromising a slightly inflated type 1 error. However, we do not advocate that the stacking approach replace Rubin's procedure for a simple analysis; rather, we believe it to be sensible alternative when repeated analyses are required, as at the model-building stage. In such procedures, weighting by $1/M$ is reasonable for large data sets with small fractions of missing data. Weighting by (fraction of observed data)$/M$ is suitable for data sets with equal fractions of missing data in all variables. This is rarely the case, and our results showed this weighting strategy performed less well when the fraction of missing data differs substantially between variables. Thus, weighting by (fraction of observed data in covariate X)$/M$ may be used in most cases. Even this choice of weight, however, is only weakly justifiable in multivariate situations, because it assumes that the fraction of missing data in X_i equals the fraction of missing information about X_i. This may be shown to be approximately true, provided that correlations between predictor variables are small, correlations between the outcome and the predictor variables are small, and data are MCAR [26]. However, if correlations are large, the stacking method is likely to have substantially inflated type 1 error, especially if there are substantial amounts of missing data in other variables. It follows that these procedures will tend to slightly overstate the statistical significance of terms in the model. Since these weighting schemes are generally liberal, the significance levels for model selection could be made somewhat more stringent (and similarly for the Single approach). The weighting schemes are likely to be adversely affected by incomplete Y, and we recommend that they only be applied after excluding observations with missing Y, as in Section 3.3: further work is required to explore this issue.

It may be useful to summarise the models selected in different imputations, such as in our methods S1–S3. Variable selection on each single imputation is done under the assumption that the imputed observations are real data, resulting in estimated standard errors being too small and an increased chance of selecting noise variables. Obtaining a final model by selecting variables appearing in either all or any model is unlikely to lead to a useful model and both selected models are dependent on the number of imputations used. Selecting variables appearing in at least 50 per cent of the imputations is similar to our stacking procedure W1 with weights $1/M$, as both approaches attempt to average across imputations and also similar to variable selection in a single imputed data set: all these methods fail to take into account the uncertainty caused by missing information.

Another possibility for variable selection is available cases analysis in which the working data set is updated at each step of the model selection process to include individuals with missing values only on variables that are not currently in the model. This approach would be useful if a single highly incomplete covariate could be excluded at an early stage, but in general it could lead

DOI: 10.1002/sim

to instability and loops in the stepwise algorithm, and it would not be expected to be as efficient as using the multiply imputed data.

A referee suggested a practical two-step procedure in which the variable selection methods (CC, Single, S1–S3, or W1–W3) are followed by a backward selection based on RR to remove any final non-relevant variables. The rationale for this procedure is that (i) the alternative variable selection methods tend to have inflated type 1 errors and (ii) in practice, the final model will typically be based on a MI analysis; hence, it seems unnatural to present a model containing predictors that may be non-significant in this final model. We repeated our simulation study for the linear model (simulation scenario 1) adopting this two-step procedure. Our results showed that type1 errors were closer to the nominal level in all methods with the exception of CC and method S3 in which the type1 error and power were grossly underestimated. This procedure may be most useful following the computationally simple approaches, such as Single or W1–W2 methods, although further exploration is needed.

In this paper the imputation model and analysis model were considered separately, but the two models could be built simultaneously. This is especially relevant when the outcome of interest is incomplete, since not knowing the analysis model implies not knowing the best imputation model for the outcome. Omitting variables from the imputation model causes downward biases in those terms in the analysis model [27]. Thus, the only safe rule is that the imputation model should, as a minimum, include all candidate predictors for the analysis model [4].

Selection of non-linear and interaction terms would present further difficulties. To assess an interaction in the analysis model, for example, terms for the interactions between the outcome and interacting variables should be included in the imputation model. Allowing for all possible interactions might make the imputation model impractically large. One possible *ad hoc* approach would be to (1) perform a first imputation without interactions; (2) select an analysis model; (3) explore interactions using a liberal significance level (to allow for downward bias from their non-inclusion in the imputation model); (4) update the imputation model to allow for the interactions detected; and (5) repeat the selection of variables and interactions with the desired significance level. This is a topic for future research.

In the UK700 trial case study, we restricted model selection to the sample of individuals with observed outcome. Since individuals with missing outcome contribute no information about the regression model, this seems a sensible way to reduce the play of chance on the imputed data; it also avoids possible bias due to mis-specifying the imputation model for Y. However, if the imputation model for Y included some auxiliary variable not in the analysis model (perhaps an associated trial outcome), then individuals with missing outcome would contribute information and should not be excluded.

In this paper we have focused on traditional non-Bayesian variable selection approaches for multiply imputed data, because we believe that this is of greatest practical relevance to most data analysts. An alternative approach described by Yang *et al*. [4] draws on the Bayesian frameworks of MI and variable selection. These authors show that our two-step method of imputing and then selecting variables using RR has a natural Bayesian extension, and they compare it with conducting Bayesian model selection and MI simultaneously within one Gibbs sampling scheme. Their simulation results show that such methods outperform the complete-case analysis but that their integrated strategy only slightly outperforms the two-step Rubin's approach.

MI is increasingly used due to the recognition of its generality and recent developments in several mainstream statistical packages. As a consequence, and in our experience, there is a demand for using multiply imputed data for more than just a simple regression analysis. One closely related

DOI: 10.1002/sim

area is model building, such as determining non-linear terms for the analysis model (e.g. fractional polynomials). Diagnostic procedures also need development, such as detecting influential points, making predictions, performing diagnostic tests, and graphical checks for model mis-specification [28–30]. This paper is a step towards demonstrating to users that although using multiply imputed data is not necessarily straightforward and involves some compromises, sensible and practical starting points do exist.

APPENDIX A: JUSTIFICATION FOR STACKING METHOD

Point estimate: Denote the kth set of imputed values for Z^{mis} as $Z^{(k)}$ for $k=1,\ldots,M$ and let $S^{(k)}(\theta)$ denote the complete-data log-likelihood for the kth imputed data set. The overall estimate of θ calculated using the stacked data set is the solution $\hat{\theta}_{\text{ST}}$ to the score equation $\sum_{k=1}^{M} S^{(k)}(\theta)=0$, while the estimate $\hat{\theta}_k$ from the kth imputed data set solves $S^{(k)}(\theta)=0$. Since in large samples $S^{(k)}(\theta)$ is likely to be approximately linear in θ with slope independent of k, it follows that

$$\hat{\theta}_{\text{ST}} \approx \frac{1}{M}\sum_{k=1}^{M}\hat{\theta}_k$$

Variance: The variance of $\hat{\theta}_{\text{ST}}$ calculated using the stacked data set can be expressed as

$$V_{\text{ST}} = \left[\sum_{k=1}^{M} I^{(k)}(\hat{\theta}_{\text{ST}})\right]^{-1} \approx \left[\sum_{k=1}^{M} \bar{W}^{-1}\right]^{-1} = \frac{1}{M}\bar{W}$$

where $I^{(k)}(\theta)$ is the information matrix from the kth imputed data set and W is the *within-imputation* variance component in (2). The approximation assumes that $I^{(k)}(\theta)$ is reasonably constant between imputations and over a small range of θ. To get the correct variance given in (4), we need to post-multiply by the matrix

$$M\bar{W}^{-1}[\bar{W}+(1+1/M)B]$$

Now if we are interested in a scalar parameter, then we can define the fraction of missing information as $f=B/(W+B)$. The correction factor above then simplifies to

$$M\left(\frac{1+f/M}{1-f}\right) \approx \frac{M}{1-f}$$

One practical way to use this correction factor when testing explanatory variable X_i is to apply common weights $w_i=(1-f_i)/M$ to all observations in the stacked regression. Finally, if the data are MCAR, then the fraction of missing information f_i approximately equals the fraction of missing data in X_i. This yields weight W3 as defined in the text. Weight W2 is a further approximation obtained by assuming a common value of f_i across variables X_i, which has the computational advantage that the weight does not need to be changed as the stepwise procedure progresses. Weight W1 additionally ignores the factor $(1-f)$ and has the intuitive rationale of correcting the size Mn of the stacked data set to the size n of the observed data set but is clearly only valid with small missing data fractions f.

DOI: 10.1002/sim

REFERENCES

1. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials* 2004; **1**:368–376.
2. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
3. Schafer JL. *Analysis of Incomplete Multivariate Data*. Chapman & Hall: London, 1997.
4. Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999; **18**:681–694.
5. Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician* 2001; **55**:244–254.
6. Royston P. Multiple imputation of missing values. *Stata Journal* 2004; **4**:227–241.
7. Royston P. Multiple imputation of missing values: update. *Stata Journal* 2005; **5**(2):188–201.
8. Royston P. Multiple imputation of missing values: update. *Stata Journal* 2005; **5**(4):527–536.
9. Van Buuren S, Brand JPL, Groothuis-Oudshoorn K, Rubin DB. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* 2006; **76**(12):1049–1064.
10. Van Buuren S, Oudshoorn CGM. *Multivariate Imputation by Chained Equations*: *MICE V1.0 User's Manual. PG/VGZ/00.038*. TNO Preventie en Gezonheid, Leiden, 2000.
11. Yu L, Burton A, Rivero-Arias O. Evaluation of available multiple imputation programmes for dealing with incomplete semi-continuous data. *Statistical Methods in Medical Research* 2007; **6**(3):243–258.
12. Little R, Rubin D. *Statistical Analysis with Missing Data*. Wiley: New York, 1987.
13. Carpenter JR, Kenward MG. Sensitivity analysis after multiple imputation under missing at random—a weighting approach. *Statistical Methods in Medical Research* 2007; **6**(3):259–275.
14. Cohen J, Cohen P, West SG, Aiken LS. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Chapter 5 (3rd edn). Erlbaum: Mahwah, NJ, 2003.
15. Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine* 1989; **8**:771–783.
16. Altman DG. *Practical Statistics for Medical Research*, Chapter 12. Chapman & Hall: London, 1991.
17. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology* 1992; **45**:265–282.
18. Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. *Applied Statistics* 1999; **48**:313–329.
19. Royston P, Sauerbrei W. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Statistics in Medicine* 2003; **22**:639–659.
20. Yang X, Belin TR, Boscardin WJ. Imputation and variable selection in linear regression models with missing covariates. *Biometrics* 2005; **61**:498–506.
21. Meng XL, Rubin DB. Performing likelihood ratio tests with multiply imputed data sets. *Biometrika* 1992; **79**:103–111.
22. Efboymson MA. Multiple regression analysis. In *Mathematical Methods for Digital Computers*, Ralston A, Wilf HS (eds). Wiley: New York, 1960; 191–203.
23. Royston P, White IR, Wood AM. Building non-linear multivariable models in multiply imputed data, in preparation.
24. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society* (*Series A*) 1999; **162**:71–94. Corrigendum: *Journal of the Royal Statistical Society, Series A* 2002; **165**:399–400.
25. Burns T, Creed F, Fahy T, Thompson S, Tyrer P, for the UK700 trial group IW. Intensive versus standard case management for severe psychotic illness: a randomised trial. *Lancet* 1999; **353**:2185–2189.
26. Little RJA. Regression with missing X's: a review. *Journal of the American Statistical Association* 1992; **87**:1227–1237.
27. Fay RE. When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Sections*. American Statistical Association, Alexandria, VA, 1992; 227–232.
28. Gelman A, King G, Liu C. Not asked and not answered: multiple imputation for multiple surveys. *Journal of the American Statistical Association* 2004; **93**:846–857.
29. Gelman A, Van Mechelen I, Verbeke G, Heitjan DF, Meulders M. Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics* 2005; **61**:74–85.
30. Kenward M, Carpenter J. Multiple imputation: current perspectives. *Statistical Methods in Medical Research* 2007; **6**(3):199–218.

DOI: 10.1002/sim