SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2016, Vol. 25(6) 2541–2557

Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280214526216
smm.sagepub.com
SAGE

Comparison of imputation variance estimators

RA Hughes, JAC Sterne and K Tilling

Abstract

Appropriate imputation inference requires both an unbiased imputation estimator and an unbiased variance estimator. The commonly used variance estimator, proposed by Rubin, can be biased when the imputation and analysis models are misspecified and/or incompatible. Robins and Wang proposed an alternative approach, which allows for such misspecification and incompatibility, but it is considerably more complex. It is unknown whether in practice Robins and Wang's multiple imputation procedure is an improvement over Rubin's multiple imputation. We conducted a critical review of these two multiple imputation approaches, a re-sampling method called full mechanism bootstrapping and our modified Rubin's multiple imputation procedure via simulations and an application to data. We explored four common scenarios of misspecification and incompatibility. In general, for a moderate sample size ($n = 1000$), Robins and Wang's multiple imputation produced the narrowest confidence intervals, with acceptable coverage. For a small sample size ($n = 100$) Rubin's multiple imputation, overall, outperformed the other methods. Full mechanism bootstrapping was inefficient relative to the other methods and required modelling of the missing data mechanism under the missing at random assumption. Our proposed modification showed an improvement over Rubin's multiple imputation in the presence of misspecification. Overall, Rubin's multiple imputation variance estimator can fail in the presence of incompatibility and/or misspecification. For unavoidable incompatibility and/or misspecification, Robins and Wang's multiple imputation could provide more robust inferences.

Keywords

bootstrap confidence intervals, imputation inference, missing data, multiple imputation, variance estimator

1 Introduction

Multiple imputation (MI)[1] is the most commonly used technique for analysing incomplete data, which are frequently encountered in health-related research. Imputations are repeatedly and independently drawn from the posterior predictive distribution of the missing data given the observed data, under a Bayesian model, to generate $m(\geq 2)$ imputed datasets. These imputed datasets are then analysed separately using complete data methods, and the m sets of results are combined using a simple set of rules. Appropriate imputation inference requires both an unbiased imputation estimator and an unbiased variance estimator.

School of Social and Community Medicine, University of Bristol, Bristol, UK

Corresponding author:
Rachael Hughes, School of Social and Community Medicine, University of Bristol, 39 Whatley Road, Bristol, BS8 2PS, UK.
Email: rachael.hughes@bristol.ac.uk

MI was first developed for analysing large incomplete public shared datasets, for which there may be many analysts, with a wide range of scientific questions, who may not have the specialized knowledge required to generate statistically valid inferences when data are incomplete. It was envisioned that the task of generating multiple imputed datasets would be undertaken by someone with specialized knowledge of missing data methods.[2] The analysts need only apply standard complete data statistical procedures to each imputed dataset separately and then combine the multiple estimates according to Rubin's rules.[1]

At the time of imputation, it may not be possible to anticipate all analysis procedures that will be applied to the imputed datasets. Consequently, the assumptions of a given analysis procedure may differ to those of the imputation model. We define an analysis procedure to be incompatible with an imputation model when one or more assumptions of the imputation model contradict with those made by the analysis procedure. For example, the imputer may assume that two subgroups have the same population mean of an incomplete variable (where missingness is independent of subgroup status) whilst the analyst estimates the population mean in one subgroup only.

As with any modelling or statistical procedure, the imputation model may be misspecified; that is, the imputation model's assumptions about the missing data mechanism (MDM) or the complete data are incorrect. For example, the imputation model could incorrectly assume homoscedastic errors. Similarly, for the analysis procedure.

The MI literature has focused primarily on generating methods and guidelines for reducing the bias of the imputation estimator. However, correct imputation inference also requires an unbiased imputation variance estimator and an efficient interval estimator with at least nominal coverage. In certain settings of misspecification and incompatibility of the models, for an unbiased imputation estimator, Rubin's variance estimator was either upwardly or downwardly biased, which led to conservative or anti-conservative confidence intervals, respectively.[2–7]

There are alternatives to Rubin's MI. Robins and Wang have proposed imputing under a frequentist procedure,[5] that fixes the imputation model parameters at the observed data maximum likelihood estimate, and an imputation variance estimator that allows for misspecification and incompatibility of the imputation model and the analysis procedure, and for non- or semi-parametric analysis procedures. However, it is unclear whether the Robins and Wang MI procedure is an improvement over Rubin's MI in situations typical of the use of imputation in practice. Also, the Robins and Wang imputation variance estimator is considerably more complex to compute than Rubin's and importantly is more technically difficult for the analyst, although the greater burden of calculating the variance estimator is still placed on the imputer. With ever faster computers, a simpler solution could be to calculate the variance of the imputation estimator using re-sampling methods. Full mechanism bootstrapping (FMB) is a bootstrapping approach to imputation that can be applied to parametric problems.[8,9]

In this paper, we have conducted the first comparative evaluation of Rubin's MI, our modified version of Rubin's MI, Robins and Wang's MI and FMB with respect to variance estimation and interval estimation. We have compared these four methods of imputation inference in four scenarios of incompatibility and misspecification of the imputation and analysis models that can occur in practice, via a data example (see Sections 2 and 4) and simulations (see Sections 5 and 6). In Section 2, we describe a motivating example for seeking an alternative to Rubin's MI. Section 3 describes all four methods considered in the context of a specific example. Section 4 revisits the motivating example, applying these methods to data from the British Child and Adolescent Mental Health Study 1999 (B-CAMHS99).[10] In Sections 5 and 6, we present our simulation study and conclude with a general discussion in Section 7.

2 Motivating example

There are several examples in the literature, where observations used for estimation at the imputation stage of MI are not used or available at the analysis stage. For example, confidential information may be used to improve estimation of the imputation model but cannot be disseminated to the analysts. Or, when external data are used to account for measurement error using MI and the external data are not included at the analysis stage because it may not be representative of the target sample.[11] We briefly describe a case study where external data were used to impute measurements, which were not collected for any of the subjects of the target study.

The B-CAMHS99 measured conduct, hyperactive and emotional problems in 15 year olds using the strengths and difficulties questionnaire (SDQ).[12] Researchers wished to compare the results of this study to previous UK population studies, which had used the Rutter A scale.[13] A calibration study was undertaken, which measured both the Rutter A scale and the SDQ scale on an independent sample of adolescents.[10] These external data were used to impute the Rutter A subscales (for conduct, hyperactive and emotional problems) in the B-CAMHS99 study, but were not included at the analysis stage.

In this scenario, the imputations were generated using extra information that was not utilized by the analyst; i.e. that the target and external data were assumed to be identically distributed. Rubin calls such imputation *superefficient* from the analyst's perspective.[2] Superefficient imputations can cause Rubin's MI variance estimator to be positively biased.[3,5]

3 Methods for imputation inference

We describe Rubin's MI and its modified version, Robins and Wang's MI and FMB. Robins and Wang have described their complex method for a general missing data setting.[5] To aid further understanding of this method we restrict ourselves to a more simplified setting, which we describe in Section 3.1. The description of Rubin's MI and FMB remains the same for a more general missing data setting.

3.1 Notation and setup

Suppose g random variables $\boldsymbol{Y} = (Y_1, \ldots, Y_g)$ are intended to be observed on n subjects. We use subscripts i and j to index subjects and variables, respectively, $(i = 1, \ldots, n;\ j = 1, \ldots, g)$. Let $\boldsymbol{y} = (y_{ij})$ denote a $(n \times g)$ matrix, whose i, j element is y_{ij}. Row i of matrix $\boldsymbol{y}$ is denoted by $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ig})$. The rows of the matrix are assumed to be independent and identically distributed. In practice, some subjects have missing observations, and we write $\boldsymbol{y} = (\boldsymbol{y}^{obs}, \boldsymbol{y}^{mis})$ with *obs* and *mis* denoting the observed and missing parts, respectively. In our simplified example, missingness is confined to one variable and without any loss of generality, we assume variable Y_g is incompletely observed for t $(t < n)$ subjects. The MDM is assumed to be ignorable for Bayesian inference,[14, p.120] and hence is also ignorable for likelihood inference.

Let, $\widetilde{\boldsymbol{Y}}$ denote $p - 1$ $(1 \le p - 1 \le g - 1)$ variables in the set $Y_1, \ldots, Y_{g-1}$ and $\widetilde{\boldsymbol{y}}_i$ the row vector of observations on $\widetilde{\boldsymbol{Y}}$ for subject i. The imputation model is the normal linear regression $y_{ig}|\widetilde{\boldsymbol{y}}_i \sim N(\widetilde{\boldsymbol{y}}_i\boldsymbol{\mu}, \sigma)$, where $\boldsymbol{\mu}$ is a column vector of dimension $p - 1$. Let $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma)^T$, a $(p \times 1)$ vector of unknown parameters.

Let, y_{ig}^k denote the kth imputed value of variable Y_g for subject i, and $\boldsymbol{y}_i^k$ is row i of the kth imputed dataset $\boldsymbol{y}^k = (\boldsymbol{y}_1^k, \ldots, \boldsymbol{y}_n^k)^T$ $(k = 1, \ldots, m)$. If y_{ig} is observed then $y_{ig}^k = y_{ig}$, and so $\boldsymbol{y}_i^k = \boldsymbol{y}_i$, for all k. Let $\ddot{\boldsymbol{Y}}$ denote q $(1 \le q \le g - 1)$ variables in the set $Y_1, \ldots, Y_{g-1}$ and $\ddot{\boldsymbol{y}}_i^k$ the set of observations on $\ddot{\boldsymbol{Y}}$ for subject i of the kth imputed dataset. Similarly, $\tilde{\boldsymbol{y}}_i^k$ is the set of observations on $\tilde{\boldsymbol{Y}}$ for subject i of the kth imputed dataset. Note, in this example, $\ddot{\boldsymbol{y}}_i^k = \ddot{\boldsymbol{y}}_i$ and $\tilde{\boldsymbol{y}}_i^k = \tilde{\boldsymbol{y}}_i$ for all k. In the interest of completeness, we have included the superscript k.

The analysis procedure is the normal linear regression model $y_{ig}^k | \ddot{\boldsymbol{y}}_i^k \sim N(\ddot{\boldsymbol{y}}_i^k \boldsymbol{\beta}, \omega)$, where $\boldsymbol{\beta}$ is a column vector of dimension q. The imputation estimate for the vector of coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$ is denoted by $\hat{\boldsymbol{\beta}}^I = (\hat{\beta}_1^I, \ldots, \hat{\beta}_q^I)^T$ and, for $j = 1, \ldots, q$, the variance estimate for $\hat{\beta}_j^I$ is $\hat{V}_j^I$, with $\hat{\boldsymbol{V}}^I = (\hat{V}_1^I, \ldots, \hat{V}_q^I)^T$.

3.2 Rubin's MI inference

Missing values are imputed independently m times, to generate m imputed datasets. Imputations are drawn independently from the posterior predictive distribution of the missing data given the observed data under a Bayesian model. For $k = 1, \ldots, m$, the analysis model is fitted to each imputed dataset $\boldsymbol{y}^k$ separately, generating regression coefficients estimate $\hat{\boldsymbol{\beta}}^k = (\hat{\beta}_1^k, \ldots, \hat{\beta}_q^k)^T$ with associated variance estimates $\hat{W}_1^k, \ldots, \hat{W}_q^k$, where $\hat{W}_j^k$ is the variance estimate for coefficient estimate $\hat{\beta}_j^k$. The set of m coefficient estimates is combined into one MI inference using a simple set of rules. When the estimand of interest is a set of regression coefficients it is simpler to apply these rules separately to each regression coefficient as follows

$$\hat{\beta}_j^I = m^{-1} \sum_{k=1}^{m} \hat{\beta}_j^k, \quad \bar{W}_j = m^{-1} \sum_{k=1}^{m} \hat{W}_j^k$$

$$B_j = (m-1)^{-1} \sum_{k=1}^{m} (\hat{\beta}_j^k - \hat{\beta}_j^I)^2$$

$$\hat{V}_j^I = \bar{W}_j + \frac{m+1}{m} B_j$$

Confidence intervals are based on the Student's t-distribution. When n is small a modified degrees of freedom formula is recommended.[15] The modified degrees of freedom ν_j^* is calculated, separately for each regression coefficient, as follows

$$\hat{\gamma}_j = \frac{(1+m^{-1})B_j}{\bar{W}_j + (1+m^{-1})B_j}, \quad \nu_j = (m-1)\left(1 + \frac{m}{m+1}\frac{\bar{W}_j}{B_j}\right)^2$$

$$\hat{\nu}_j^{obs} = (1-\hat{\gamma}_j)\left(\frac{\nu_j^{com}+1}{\nu_j^{com}+3}\right)\nu_j^{com}, \quad \nu_j^* = \{\nu_j^{-1} + (\hat{\nu}_j^{obs})^{-1}\}^{-1}$$

where ν_j^{com} is the degrees of freedom about β_j when there are no missing values.

We shall also consider a modification in which the within imputation variances, W_j^k, are calculated using the robust sandwich variance estimator.[16,17] This modification only affects the calculation of the variance of the imputation estimate; the imputation estimate remains the same as for Rubin's MI. We refer to this modified method as 'robust Rubin's MI'.

3.3 Robins and Wang's imputation inference

The Robins and Wang variance estimator for imputation does not require multiply imputed datasets, although it may be more efficient when a dataset is multiply imputed.[5] To our knowledge, there does not exist any commercial or freely available software that calculates this variance estimator. All derivatives involving scalars, vectors and matrices are defined as in,[18] e.g. for $n \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, define $\partial \boldsymbol{a}/\partial \boldsymbol{b}^T = [\partial a_r/\partial b_c]$, a square matrix of dimension n where c and r, respectively, refer to column and row numbers.

The m sets of imputations are drawn independently from the predictive distribution of the missing data given the observed data under the imputation model evaluated at $\hat{\boldsymbol{\theta}}$, the observed data maximum likelihood estimate of $\boldsymbol{\theta}$. Therefore, each set of imputations is drawn conditionally on the same parameter estimate. In our simple case, $\hat{\boldsymbol{\theta}}$ is the complete case estimate of $\boldsymbol{\theta}$ (i.e. estimation is based on the n–t subjects with observed Y_g). The m imputed datasets are then stacked into a $mn \times g$ dataset $(\boldsymbol{y}^1, \ldots, \boldsymbol{y}^m)^T$.

The analysis procedure is applied to the stack of imputed datasets $(\boldsymbol{y}^1, \ldots, \boldsymbol{y}^m)^T$, treating the mn rows as independent, and the estimate of $\boldsymbol{\beta}$ is the Robins and Wang imputation estimate $\hat{\boldsymbol{\beta}}^I$. The analysis procedure can be non-, semi- or fully parametric such that $\hat{\boldsymbol{\beta}}^I$ is the solution to the estimating equation

$$\sum_{i=1}^{n} m^{-1} \sum_{k=1}^{m} \boldsymbol{u}_i^k(\hat{\boldsymbol{\theta}}, \boldsymbol{\beta}) = 0$$

where $\boldsymbol{u}_i^k(\hat{\boldsymbol{\theta}}, \boldsymbol{\beta}) = \ddot{\boldsymbol{y}}_i^k(y_{ig}^k - \ddot{\boldsymbol{y}}_i^k \boldsymbol{\beta})$ for our example.

To calculate the Robins and Wang variance estimator, both the imputer and analyst must generate additional information. The imputer supplies two further datasets based on the score function of the imputation model. The analyst must generate a dataset and a matrix, which are both based on the estimating equation of the analysis procedure evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^I$. The analyst then inputs these pieces of information into a set of matrix formulae to generate $\hat{\boldsymbol{V}}^I$. Below we provide details on how to calculate these pieces of information and the matrix formulae.

The score function of the imputation model is the derivative, with respect to $\boldsymbol{\theta}$, of its log-likelihood function. The contribution from subject i to the log-likelihood function is $\log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta}) = 0.5(-\log 2\pi - \log\sigma - \sigma^{-1}(y_{ig} - \widetilde{\boldsymbol{y}}_i\boldsymbol{\mu})^2)$ and its derivative is column vector $\partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})/\partial\boldsymbol{\theta} = [\partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})/\partial\boldsymbol{\mu}, \partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})/\partial\sigma]^T$, where

$$\partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})/\partial\boldsymbol{\mu} = \left(\frac{y_{ig} - \widetilde{\boldsymbol{y}}_i\boldsymbol{\mu}}{\sigma}\right)\widetilde{\boldsymbol{y}}_i^T$$

$$\partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})/\partial\sigma = \frac{1}{2}\left(-\frac{1}{\sigma} + \frac{(y_{ig} - \widetilde{\boldsymbol{y}}_i\boldsymbol{\mu})^2}{\sigma^2}\right)$$

For subject i, when $\boldsymbol{y}_i$ is completely observed let $\boldsymbol{s}_i^{obs} = \boldsymbol{s}_i^{obs}(\hat{\boldsymbol{\theta}}) = (\partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})/\partial\boldsymbol{\theta})^T|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ and when $\boldsymbol{y}_i$ is incompletely observed let $\boldsymbol{s}_i^{obs}$ be a zero row vector of dimension p. Conversely, for $k = 1, \ldots, m$, when $\boldsymbol{y}_i$ is incompletely observed let $\boldsymbol{s}_i^{k,mis} = \boldsymbol{s}_i^{k,mis}(\hat{\boldsymbol{\theta}}) = (\partial \log f(\boldsymbol{y}_{ig}^k|\widetilde{\boldsymbol{y}}_i^k, \boldsymbol{\theta})/\partial\boldsymbol{\theta})^T|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ and when $\boldsymbol{y}_i$ is completely observed let $\boldsymbol{s}_i^{k,mis}$ be a zero row vector of dimension p. For each k we then have $n \times p$ dataset $\boldsymbol{S}^{mis,k} = (\boldsymbol{s}_1^{mis,k}, \ldots, \boldsymbol{s}_n^{mis,k})^T$ and stacking these m datasets we have $\boldsymbol{S}^{mis} = (\boldsymbol{S}^{mis,1}, \ldots, \boldsymbol{S}^{mis,m})^T$.

For the second dataset based on the score function of the imputation model, first calculate the derivative of column vector $\partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})/\partial \boldsymbol{\theta}$ with respect to row vector $\boldsymbol{\theta}^T$, which is the $p \times p$ matrix

$$\begin{bmatrix} \dfrac{\partial}{\partial \boldsymbol{\mu}^T}\left(\dfrac{\partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\mu}}\right) & \dfrac{\partial}{\partial \sigma}\left(\dfrac{\partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\mu}}\right) \\[2ex] \dfrac{\partial}{\partial \boldsymbol{\mu}^T}\left(\dfrac{\partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})}{\partial \sigma}\right) & \dfrac{\partial}{\partial \sigma}\left(\dfrac{\partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})}{\partial \sigma}\right) \end{bmatrix}$$

where

$$\frac{\partial}{\partial \boldsymbol{\mu}^T}\left(\frac{\partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\mu}}\right) = -\frac{1}{\sigma}\widetilde{\boldsymbol{y}}_i^T\widetilde{\boldsymbol{y}}_i$$

$$\begin{aligned} \frac{\partial}{\partial \sigma}\left(\frac{\partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\mu}}\right) &= -\frac{1}{\sigma^2}\widetilde{\boldsymbol{y}}_i^T(y_{ig} - \widetilde{\boldsymbol{y}}_i\boldsymbol{\mu}) \\ &= \left\{\frac{\partial}{\partial \boldsymbol{\mu}^T}\left(\frac{\partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})}{\partial \sigma}\right)\right\}^T \end{aligned}$$

$$\frac{\partial}{\partial \sigma}\left(\frac{\partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})}{\partial \sigma}\right) = \frac{1}{2\sigma^2} - \frac{(y_{ig} - \widetilde{\boldsymbol{y}}_i\boldsymbol{\mu})^2}{\sigma^3}$$

For subject i with observed y_{ig} define

$$\boldsymbol{d}_i^T = \boldsymbol{d}_i(\hat{\boldsymbol{\theta}})^T = -\left[n^{-1}\sum_{i=1}^{n}\left\{\frac{\partial}{\partial \boldsymbol{\theta}^T}\left(\frac{\partial \log f(y_{ig}|\widetilde{\boldsymbol{y}}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\right\}\Bigg|_{\theta=\hat{\theta}}\right]^{-1}\boldsymbol{s}_i^{obs}(\hat{\boldsymbol{\theta}})^T$$

When y_{ig} is incompletely observed $\boldsymbol{d}_i$ is a zero row vector of dimension p. These n row vectors form the second dataset $\boldsymbol{D} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n)^T$. The imputer's role is now completed and the stacked imputed datasets $(\boldsymbol{y}^1, \ldots, \boldsymbol{y}^m)^T$ and datasets $\boldsymbol{S}^{mis}$ and $\boldsymbol{D}$ are passed on to the analyst.

Evaluating $\boldsymbol{u}_i^k(\hat{\boldsymbol{\theta}}, \boldsymbol{\beta})$ at the imputation estimate $\hat{\boldsymbol{\beta}}^I$, the analyst generates m $(n \times q)$ datasets $\boldsymbol{U}^k = (\boldsymbol{u}_1^k(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}^I), \ldots, \boldsymbol{u}_n^k(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}^I))^T(k = 1, \ldots, m)$. These m datasets are stacked to form dataset $\boldsymbol{U} = (\boldsymbol{U}^1, \ldots, \boldsymbol{U}^m)^T$. The matrix generated by the analyst is $\boldsymbol{\tau} = \boldsymbol{\tau}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}^I)$, which is calculated by differentiating column vector $\boldsymbol{u}_i^k(\hat{\boldsymbol{\theta}}, \boldsymbol{\beta})^T$ with respect to row vector $\boldsymbol{\beta}^T$, to generate a square matrix of dimension q

$$\frac{\partial \boldsymbol{u}_i^k(\hat{\boldsymbol{\theta}}, \boldsymbol{\beta})^T}{\partial \boldsymbol{\beta}^T} = \left[\frac{\partial \boldsymbol{u}_i^k(\hat{\boldsymbol{\theta}}, \boldsymbol{\beta})_r}{\partial \beta_c}\right]; \quad r, c = 1, \ldots, q$$

where r and c, respectively, denote the row and column of the matrix and $\boldsymbol{u}_i^k(\hat{\boldsymbol{\theta}}, \boldsymbol{\beta})_r = y_{ir}^k(y_{ig}^k - \ddot{\boldsymbol{y}}_i^k\boldsymbol{\beta})$. We can then calculate

$$\begin{aligned}\boldsymbol{\tau} = \boldsymbol{\tau}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}^I) &= -(nm)^{-1}\sum_{i=1}^{n}\sum_{k=1}^{m}\left(\frac{\partial \boldsymbol{u}_i^k(\hat{\boldsymbol{\theta}}, \boldsymbol{\beta})^T}{\partial \boldsymbol{\beta}^T}\right)\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^I} \\ &= n^{-1}\sum_{i=1}^{n}\ddot{\boldsymbol{y}}_i^{k\,T}\ddot{\boldsymbol{y}}_i^k\end{aligned}$$

The analyst now inputs datasets $\boldsymbol{D}, \boldsymbol{S}^{mis}$ and $\boldsymbol{U}$ and matrix $\boldsymbol{\tau}$ into the following matrix formulae

$$\begin{aligned}&\overline{\boldsymbol{u}}_i = \overline{\boldsymbol{u}}_i(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}^I) = m^{-1}\sum_{k=1}^{m}\boldsymbol{u}_i^k(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}^I) \\ &\boldsymbol{\kappa} = \boldsymbol{\kappa}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}^I) = (nm)^{-1}\sum_{i=1}^{n}\sum_{k=1}^{m}(\boldsymbol{u}_i^k(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}^I))^T\boldsymbol{s}_{\boldsymbol{i}}^{mis,k} \\ &\boldsymbol{\Lambda} = \boldsymbol{\Lambda}(\hat{\boldsymbol{\theta}}) = n^{-1}\sum_{i=1}^{n}\boldsymbol{d}_i^T\boldsymbol{d}_i, \quad \boldsymbol{\Omega} = \boldsymbol{\Omega}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}^I) = n^{-1}\sum_{i=1}^{n}\overline{\boldsymbol{u}}_i^T\overline{\boldsymbol{u}}_i \\ &\boldsymbol{\Delta} = \boldsymbol{\Delta}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}^I) = \boldsymbol{\Omega} + \boldsymbol{\kappa}\boldsymbol{\Lambda}\boldsymbol{\kappa}^T + n^{-1}\sum_{i=1}^{n}\left\{\boldsymbol{\kappa}\boldsymbol{d}_i^T\overline{\boldsymbol{u}}_i + (\boldsymbol{\kappa}\boldsymbol{d}_i^T\overline{\boldsymbol{u}}_i)^T\right\} \\ &\boldsymbol{\Gamma} = n^{-1}\boldsymbol{\tau}^{-1}\boldsymbol{\Delta}(\boldsymbol{\tau}^{-1})^T\end{aligned}$$

Finally, for $j = 1, \ldots, g-1$, the jth diagonal entry of matrix $\boldsymbol{\Gamma}$ is the variance estimate corresponding to the coefficient imputation estimator β_j^I, i.e. $\hat{V}_j^I = \Gamma_{jj}$.

3.4 Full mechanism bootstrapping

FMB can be applied to parametric and non-parametric problems,[8] and under all three MDM assumptions.[19] Unless, the MDM is assumed to be missing completely at random (MCAR) then FMB requires modelling of the MDM. In Efron's worked example of the procedure a deterministic imputation model was used. Shao and Sitter[9] describe FMB with a random regression imputation method.

FMB is implemented as follows:[8]

(1) Impute the incomplete dataset $\boldsymbol{y}$ once to generate imputed dataset $\boldsymbol{y}^I$. Apply the analysis procedure to dataset $\boldsymbol{y}^I$. The estimate of $\boldsymbol{\beta}$ is the imputation estimate $\hat{\boldsymbol{\beta}}^I$.

(2) Sample with replacement n rows from $\boldsymbol{y}^I$ to obtain bootstrapped dataset $\boldsymbol{y}^{I*}$.
(3) Set observations in $\boldsymbol{y}^{I*}$ to missing by applying the same missing data pattern (MDP) as for $\boldsymbol{y}$ if MCAR is assumed. Otherwise, infer missingness under a model for the MDM.
(4) Singly impute the incomplete dataset from step 3, using the same imputation model as in step 1, to generate the bootstrapped imputed dataset $\boldsymbol{y}^{I\circledast}$.
(5) Apply the analysis procedure to dataset $\boldsymbol{y}^{I\circledast}$ and store the estimate of $\boldsymbol{\beta}$ as bootstrap replication $\hat{\boldsymbol{\beta}}^{I\circledast}$.
(6) Repeat steps 2–5 T times to obtain T bootstrap replicates.

Standard bootstrap formulae[20] can be applied to the bootstrap replicates $\hat{\boldsymbol{\beta}}^{I\circledast}$ to calculate the bootstrap variance and confidence intervals for each $\hat{\beta}_j^I$. Currently, there does not exist an algorithm or formula for the calculation of the acceleration constant, which is used to generate the bias-corrected and accelerated bootstrap confidence intervals. However, calculating the acceleration constant based on the formula used to construct non-parametric bias-corrected and accelerated confidence intervals can give a reasonable approximation.[8]

4 Return to motivating example

We applied the methods of Section 3 to data from the B-CAMHS99 study discussed in Section 2. The data (from the B-CAMHS99 study) consisted of fully observed measurements for sex and the SDQ subscales in 855 (433 boys and 422 girls) adolescents aged 15. The external data consisted of fully observed measurements for sex, the SDQ subscales and Rutter A subscales in 380 (203 boys and 177 girls) adolescents with median age 15 years (interquartile range 13–15 years).

For the purposes of this case study, each Rutter A subscale was imputed separately under an ordinal logistic regression model, with the three SDQ subscales (for conduct, hyperactive and emotional problems) and sex as covariates. The analysis of interest was a linear regression of the Rutter A subscale, with sex and the constant term as covariates, estimated in the B-CAMHS99 study only. We generated 50 imputed datasets for Rubin's MI, robust Rubin's MI and Robins and Wang's MI, and 2500 bootstrap replications for FMB.

Table 1 presents the results of the analyses of the B-CAMHS99 study. As expected, all point estimates were comparable. Robins and Wang's MI had the smallest standard errors and narrowest confidence intervals for all cases. The maximum percentage difference in the standard errors of Rubin's MI and Robins and Wang's MI was just under 23%, and the corresponding Rubin's confidence interval was 26.5% longer than the Robins and Wang's confidence interval. In almost all cases, FMB had smaller standard errors than Rubin's MI, and the maximum percentage difference in the standard errors was just under 14%. The larger standard errors of Rubin's MI (or robust Rubin's MI) indicate potential inflation due to superefficient imputations. Robust Rubin's MI showed little improvement over Rubin's MI.

5 Simulation study methods

We compared the methods described in Section 3 in four scenarios of incompatibility and misspecification of the imputation and analysis models. The simulation study was based on a hypothetical dataset of one binary variable, sex (0 denoting male and 1 denoting female), and four continuous variables, age, height, weight and natural log of insulin index (hereafter referred to as loginsindex). The data were generated under the following model

$$\begin{aligned} sex \ \sim Bernoulli(\pi),\ age, height|sex \ &\sim N(\alpha_0+\alpha_1 sex, \Sigma) \\ weight &= \iota_0+\iota_1 sex+\iota_2 age+\iota_3 height+\eta^{sex}\lambda\times error_W \\ loginsindex &= \beta_0+\beta_1 sex+\beta_2 age+\beta_3 weight+\eta^{sex}\omega\times error_L \end{aligned} \tag{1}$$

where $error_W$ and $error_L$ are error distributions and $\eta^{sex}=1$ when sex $=0$ and $\eta^{sex}=\eta$ when sex $=1$.

Model (equation (1)) was based on a dataset of standard anthropometric measurements of 951 young adults enrolled in the Barry Caerphilly Growth study.[21] The parameters of the data model

Table 1. Analysis results of the B-CAMHS99 study using Rubin's MI (Rubin), robust Rubin's MI (Rubin R), Robins and Wang's MI (RW) and full mechanism bootstrapping (FMB).

		β_{cons}				β_{sex}			
Scale		Estimate	SE	95% CI	CI width	Estimate	SE	95% CI	CI width
Conduct problems	Rubin	1.13	0.112	(0.910,1.35)	0.441	−0.155	0.158	(−0.467,0.157)	0.624
	Rubin R	1.13	0.110	(0.913,1.35)	0.434	−0.155	0.151	(−0.453,0.143)	0.596
	RW	1.13	0.0968	(0.944,1.32)	0.379	−0.158	0.131	(−0.415,0.0985)	0.514
	FMB								
	Normal	1.09	0.107	(0.883,1.30)	0.420	−0.133	0.147	(−0.421,0.156)	0.577
	Percentile	–	–	(0.911,1.35)	0.437	–	–	(−0.434,0.142)	0.576
	BC	–	–	(0.873,1.28)	0.407	–	–	(−0.393,0.202)	0.595
Emotional problems	Rubin	0.986	0.0887	(0.812,1.16)	0.350	0.291	0.146	(0.00215,0.579)	0.577
	Rubin R	0.986	0.0813	(0.826,1.15)	0.321	0.291	0.141	(0.0121,0.569)	0.557
	RW	0.984	0.0721	(0.842,1.12)	0.283	0.312	0.116	(0.0836,0.540)	0.456
	FMB								
	Normal	1.03	0.0832	(0.862,1.19)	0.326	0.242	0.127	(−0.00565,0.490)	0.496
	Percentile	–	–	(0.823,1.15)	0.326	–	–	(0.0606,0.553)	0.492
	BC	–	–	(0.913,1.24)	0.325	–	–	(−0.0777,0.423)	0.501
Hyperactive problems	Rubin	0.853	0.0863	(0.683,1.02)	0.340	−0.260	0.126	(−0.508,−0.0110)	0.497
	Rubin R	0.853	0.0883	(0.679,1.03)	0.348	−0.260	0.122	(−0.501,−0.0181)	0.483
	RW	0.853	0.0824	(0.686,1.01)	0.323	−0.253	0.107	(−0.462,−0.0433)	0.419
	FMB								
	Normal	0.871	0.0917	(0.691,1.05)	0.359	−0.288	0.119	(−0.520,−0.0551)	0.465
	Percentile	–	–	(0.680,1.04)	0.361	–	–	(−0.490,−0.0229)	0.467
	BC	–	–	(0.723,1.09)	0.369	–	–	(−0.543,−0.0812)	0.462

Imputation estimate with standard error (SE), 95% confidence interval (CI) and CI width for the constant (β_{cons}) and sex (β_{sex}) coefficients. Bootstrap confidence intervals: normal approximation (normal), percentile and bias-corrected (BC).

were set to the estimates from an analysis of this dataset. The values, to four significant figures, of these parameters were:

$$\pi = 0.4577, \quad \alpha_0 = (25.02, 1.774), \quad \alpha_1 = (-0.03616, -0.1336),$$
$$\Sigma = \begin{pmatrix} 0.5521 & 0.001574 \\ 0.001574 & 0.003705 \end{pmatrix}, \quad \iota_0 = -32.98$$

$\iota_1 = -2.314$, $\iota_2 = -0.01566$, $\iota_3 = 65.38$, $\lambda = 12.29$, $\beta_0 = 1.854$, $\beta_1 = 0.2908$, $\beta_2 = 0.08003$, $\beta_3 = 0.01119, \omega = 0.7887$ and $\eta = 0.5$. Different scenarios were created by setting parameters α_1, ι_1, η and β_1 to their null values; zero vector for α_1, 0 for ι_1 and β_1, and 1 for η. The values of the remaining parameters were fixed.

The analysis model was the normal linear regression of loginsindex on weight, with adjustment for other variables. The aim of the simulation study was to evaluate the methods with respect to the imputation variance estimator when the imputation estimator was unbiased. To avoid bias in the imputation estimator due to the MDM we set weight to be missing, for a subset of subjects, under an MCAR mechanism. The missing weight measurements were imputed under a normal linear regression model. Both imputation and analysis models included a constant term and assumed that the variance of the error terms was constant for all values of the outcome variable (i.e. homoscedasticity). Unless otherwise stated, error distributions $error_W$ and $error_L$ were normal, weight measurements were missing in men and women and the imputation and analysis models were fitted to the entire sample. The scenarios considered were as follows:

Scenario 1: Subgroup analysis. We set the true conditional distributions of age, height, weight and loginsindex to be the same in men and women; i.e. $\alpha_1 = (0, 0), \iota_1 = 0, \beta_1 = 0$ and $\eta = 1$. Weight measurements were missing (completely at random) in men only. The covariates of the imputation model were age, height and loginsindex. The covariates of the analysis model were age and weight, and the model was only fitted to the men's observations. There was incompatibility between the imputation and the analysis model since only the imputation model assumed that the continuous variables were identically distributed in men and women.

Scenario 2: Heteroscedastic errors. We set the true conditional distributions of age, height, weight and loginsindex to have different means in men and women; i.e. α_1, ι_1 and β_1 were set to their non-null values stated earlier. Additionally, among women the variance of weight and loginsindex was set to be 1/4 of the variance in men; i.e. $\eta = 1/2$. The covariates of the imputation model were sex, age, height and loginsindex, and the covariates of the analysis model were sex, age and weight. The imputation and analysis models were compatible but incorrectly specified because they assumed homoscedastic errors.

Scenario 3: Omitted interaction. We set the true conditional distributions of age, height, weight and loginsindex to have different means in men and women. The variances of weight and loginsindex were set to be the same in men and women; i.e. $\eta = 1$. The covariates of the imputation model were sex, age, height and loginsindex, and the covariates of the analysis model were sex, age, weight and the interaction between weight and sex. The imputation model was correctly specified but because the analysis model included an unnecessary interaction term the imputation and analysis models were incompatible.

Scenario 4: Moderate and severe non-normality. This scenario was motivated by a simulation study that investigated the performance of MI methods with non-normal distributions.[22] Unlike these authors, we were only interested in the effect of the shape of the error distribution, not the size of the error variances. The true conditional distributions of age, height, weight and loginsindex were

> set to be the same in men and women, i.e. $\alpha_1 = (0,0), \iota_1 = 0, \beta_1 = 0$ and $\eta = 1$. For moderate departures from non-normality, we investigated nine different parameter specifications by setting independently the distributions of $error_W$ and $error_L$ to be the uniform distribution, Student's t-distribution with six degrees of freedom or the log-normal distribution $\exp\{N(0, 1/4^2)\}$. For severe departures from non-normality, we investigated eight different parameter specifications by setting independently the distributions of $error_W$ and $error_L$ to be the uniform distribution, Student's t-distribution with three degrees of freedom or the log-normal distribution $\exp\{N(0, 1)\}$. All error distributions had mean zero and unit variance. The imputation and analysis models were the same as in the subgroup analysis scenario, although both models were fitted to the entire sample. The imputation and analysis models were compatible, but misspecified because they assumed a normal error distribution.

For all scenarios, we repeated the simulation study for sample sizes $n = 100$ and $n = 1000$ and probabilities 0.6 and 0.4 that weight was observed. For the subgroup analysis scenario only, the probability that weight was observed was one among women and 0.6 or 0.4 among men. For each combination of scenario, parameter specification, sample size and observation probability we generated 2500 independent simulated datasets. Based on 2500 simulations the Monte Carlo standard error for the true coverage probability of 0.95 is $\surd(0.95(1-0.95)/2500) = 0.0044$,[23] implying that the estimated coverage probability should lie within the range 0.941 and 0.960 (with 95% probability). For Rubin's MI and robust Rubin's MI, we imputed the data using the independent Jeffrey's prior. For FMB, the data were imputed using the same frequentist imputation method used for Robins and Wang's MI. Each incomplete dataset was imputed 50 times for Rubin's MI, robust Rubin's MI and Robins and Wang's MI, and 2500 bootstrap samples were generated for FMB.

6 Simulation study results

Table 2 presents the results of imputation inference according to the methods of Rubin, robust Rubin and Robins and Wang in the four scenarios of misspecified or incompatible imputation and analysis models, where the probability of observing weight was 0.4. For moderate departure from normality and severe departure from normality, we have reported the results corresponding to the error distributions $error_W \sim \exp\{N(0, 1/4^2)\}, error_L \sim \exp\{N(0, 1/4^2)\}$ and $error_W \sim$ Student's t with three degrees of freedom, $error_L \sim \exp\{N(0, 1)\}$. The results for other parameter specifications are summarized in the text below.

First consider the left-hand side of Table 2, corresponding to sample size 1000. The imputation estimates, $\hat{\beta}_3^I$, for the subgroup analysis, heteroscedastic errors and omitted interaction scenarios were unbiased; that is, the Monte Carlo 95% confidence interval $\hat{\beta}_3^I \pm 1.96 \times \sqrt{var(\hat{\beta}_3^I)/2500}$ contained the true value of β_3. For each of these three scenarios, the mean of the Robins and Wang variance estimate, $\hat{V}_3^I$, was close to the sampling variance of $\hat{\beta}_3^I$ and the confidence interval coverage probability was close to the nominal level. For the subgroup analysis and omitted interaction scenarios, Rubin's variance estimates were upwardly biased (i.e. the mean of $\hat{V}_3^I$ was larger than the sampling variance of $\hat{\beta}_3^I$), leading to conservative confidence intervals that were on average wider than those of Robins and Wang. Conversely, for the heterogeneous errors scenario, Rubin's variance estimate was downwardly biased and the coverage probability of the confidence intervals was more than 4% below the nominal level. The robust Rubin's MI method performed similar to Rubin's MI for the subgroup analysis and omitted interaction scenarios and showed a

Table 2. Summary of the simulation results for Rubin's MI (Rubin), robust Rubin's MI (Rubin R) and Robins and Wang's MI (RW) for the weight coefficient: mean bias and empirical variance (Emp. var) of the imputation estimate $\hat{\beta}_3^I$, mean of the estimated variance $\hat{V}_3^I$, empirical coverage probability (CP) and mean width of the 95% confidence interval for $\hat{\beta}_3^I$.

		$n=1000$					$n=100$				
Scenario		Mean $\hat{\beta}_3^I$	Emp. var $\hat{\beta}_3^I \times 1000$	Mean $\hat{V}_3^I \times 1000$	CP	Mean width	Mean $\hat{\beta}_3^I$	Emp. var $\hat{\beta}_3^I \times 1000$	Mean $\hat{V}_3^I \times 1000$	CP	Mean width
1. Subgroup analysis	Rubin	0.0112	0.0007	0.0123	0.990	0.014	0.0108	0.0811	0.1350	0.986	0.046
	Rubin R	0.0112	0.0007	0.0122	0.990	0.014	0.0108	0.0811	0.1320	0.984	0.046
	RW	0.0112	0.0007	0.0007	0.948	0.011	0.0108	0.0831	0.0721	0.913	0.033
2. Heterogeneous errors	Rubin	0.0112	0.0126	0.0095	0.909	0.012	0.0111	0.1553	0.1091	0.887	0.042
	Rubin R	0.0112	0.0126	0.0102	0.920	0.013	0.0111	0.1553	0.1146	0.895	0.043
	RW	0.0112	0.0127	0.0130	0.945	0.014	0.0118	0.1729	0.1310	0.863	0.043
3. Omitted interaction	Rubin	0.0112	0.0100	0.0135	0.977	0.015	0.0110	0.1159	0.1528	0.974	0.049
	Rubin R	0.0112	0.0100	0.0135	0.977	0.015	0.0110	0.1159	0.1504	0.974	0.049
	RW	0.0112	0.0100	0.0102	0.952	0.013	0.0117	0.1277	0.1111	0.904	0.040
4. Moderate non-normality	Rubin	0.0112	0.0085	0.0087	0.954	0.012	0.0111	0.0968	0.0955	0.946	0.039
	Rubin R	0.0112	0.0085	0.0087	0.953	0.012	0.0111	0.0968	0.0956	0.947	0.039
	RW	0.0112	0.0086	0.0087	0.947	0.012	0.0118	0.1074	0.0903	0.898	0.036
5. Severe non-normality	Rubin	0.0119	0.0181	0.0092	0.858	0.012	0.0125	0.1377	0.1076	0.880	0.041
	Rubin R	0.0119	0.0181	0.0138	0.918	0.014	0.0125	0.1377	0.1290	0.923	0.044
	RW	0.0119	0.0183	0.0159	0.919	0.014	0.0135	0.1594	0.1198	0.872	0.039

Bootstrap confidence intervals: normal approximation, percentile and bias-corrected (BC). Probability of observing weight was 0.4. Moderate non-normality using $error_W \sim \exp\{N(0,1/4^2)\}$ and $error_L \sim \exp\{N(0,1/4^2)\}$. Severe non-normality using $error_W \sim$ Student's t with three degrees of freedom and $error_L \sim \exp\{N(0,1)\}$.

slight improvement over Rubin's MI (i.e. higher coverage probability) for the heterogeneous errors scenario.

For the moderate non-normality scenario, the imputation estimate was unbiased and the confidence interval coverage probability was close to the nominal level for all three methods. However, for the severe non-normality scenario, the imputation estimates were upwardly biased for all three methods and for both sample sizes. Robins and Wang MI under-estimated the sampling variance of $\hat{\beta}_3^I$ the least and had the highest coverage probability, although this was still 3% below the nominal level. The robust Rubin's MI method was an improvement on Rubin's MI such that the confidence interval coverage probability for robust Rubin's MI was less than 1% below that of Robins and Wang's MI.

Now consider the right-hand side of Table 2, corresponding to sample size 100. The results for Rubin's MI and robust Rubin's MI followed the same patterns as noted for sample size 1000. There was a deterioration in the performance of Robins and Wang's MI when applied to a dataset of sample size 100. Firstly, the imputation estimates for all but the subgroup analysis scenario were (slightly) upwardly biased. Secondly, the Robins and Wang variance estimates were downwardly biased for all scenarios, resulting in confidence interval coverage probabilities that were at least 3% below the nominal level.

Table 3 reports the corresponding results for FMB. The FMB imputation estimates were unbiased for both sample sizes, and for all scenarios except severe non-normality. However, estimation of β_3 was less efficient than for the other methods. Of the three types of confidence intervals, in almost all cases the percentile confidence interval had the highest coverage probability for the same average confidence interval width.

For the subgroup analysis, heterogeneous errors, omitted interaction and moderate non-normality, when the sample size was 1000 Robins and Wang MI outperformed FMB; i.e. had the narrowest average confidence interval and at least nominal coverage. When the sample size was 100, for the subgroup analysis and omitted interaction scenarios FMB with the percentile confidence interval was marginally better than Rubin's MI because (for comparable coverage probabilities) the bootstrap confidence intervals were narrower on average than those of Rubin's MI. The relative inefficiency of FMB to Rubin's MI was outweighed by the upward bias of Rubin's variance estimates. For the heterogeneous errors scenario, sample size 100, FMB with the percentile confidence interval had the highest coverage probability, although still outside the expected range (0.941–0.960). For the moderate non-normality scenario, Rubin's MI was the best method for sample size 100, with close to nominal coverage and the narrowest mean confidence interval width. For the severe non-normality scenario, and for both sample sizes, FMB had the highest coverage probabilities, although still below 0.941.

When the probability of observing weight was 0.6, the pattern of results was the same as in Tables 2 and 3, although the differences between the methods were less marked. Across the nine different error distributions investigated for the moderate non-normality scenario, and for the severe non-normality specifications $error_W \sim \exp\{N(0,1)\}, error_L \sim uniform$ and $error_W \sim$ Student's t-distribution with three degrees of freedom, $error_L \sim uniform$ the results were virtually identical to those reported for moderate departures from normality in Tables 2 and 3. Across the remaining six error distributions investigated for severe non-normality, the pattern of the results was very similar to the severe non-normality scenario in Tables 2 and 3, so that the conclusions drawn with respect to the comparisons of the methods remained the same. For the other coefficients of the analysis model, either the pattern of results was the same as in Tables 2 and 3 but with smaller differences between the methods, or all methods had coverage probabilities of 94–95%, with Robins and Wang MI having the narrowest confidence interval on average and FMB having the widest.

Table 3. Summary of the full mechanism bootstrapping simulation results for the weight coefficient: mean and empirical variance (Emp. var) of the imputation estimate $\hat{\beta}_3^I$, mean of the estimated standard error $\hat{V}_3^I$ and empirical coverage probability (CP) and mean width of the 95% confidence interval for $\hat{\beta}_3^I$.

		$n = 1000$					$n = 100$				
Scenario		Mean $\hat{\beta}_3^I$	Emp. var $\hat{\beta}_3^I \times 1000$	Mean $\hat{V}_3^I \times 1000$	CP	Mean width	Mean $\hat{\beta}_3^I$	Emp. var $\hat{\beta}_3^I \times 1000$	Mean $\hat{V}_3^I \times 1000$	CP	Mean width
Subgroup analysis	Normal	0.0112	0.0108	0.0108	0.952	0.013	0.0111	0.1258	0.1238	0.937	0.043
	Percentile	–	–	–	0.988	0.013	–	–	–	0.981	0.043
	BC	–	–	–	0.858	0.013	–	–	–	0.856	0.043
Heterogeneous errors	Normal	0.0112	0.0146	0.0131	0.933	0.014	0.0112	0.1806	0.1468	0.891	0.046
	Percentile	–	–	–	0.933	0.014	–	–	–	0.910	0.047
	BC	–	–	–	0.934	0.014	–	–	–	0.906	0.047
Omitted interaction	Normal	0.0112	0.0135	0.0137	0.948	0.015	0.0109	0.1559	0.1591	0.940	0.048
	Percentile	–	–	–	0.972	0.015	–	–	–	0.966	0.049
	BC	–	–	–	0.910	0.015	–	–	–	0.906	0.049
Moderate non-normality	Normal	0.0112	0.0104	0.0106	0.950	0.013	0.0113	0.1205	0.1170	0.927	0.042
	Percentile	–	–	–	0.951	0.013	–	–	–	0.938	0.042
	BC	–	–	–	0.950	0.013	–	–	–	0.936	0.042
Severe non-normality	Normal	0.0118	0.0199	0.0176	0.944	0.016	0.0129	0.1648	0.1598	0.928	0.048
	Percentile	–	–	–	0.938	0.016	–	–	–	0.933	0.048
	BC	–	–	–	0.924	0.016	–	–	–	0.912	0.048

Bootstrap confidence intervals: normal approximation, percentile and bias-corrected (BC). Probability of observing weight was 0.4. Moderate non-normality using $error_W \sim \exp\{N(0,1/4^2)\}$ and $error_L \sim \exp\{N(0,1/4^2)\}$. Severe non-normality using $error_W \sim$ Student's t with three degrees of freedom and $error_L \sim \exp\{N(0,1)\}$.

For the subgroup analysis and interaction scenarios in several instances, the Robins and Wang variance estimate was smaller than the corresponding variance estimate that resulted from analysing the fully observed data, i.e. data without missing observations (data not shown). This is due to the fact that the imputations are superefficient with respect to the analysis procedure; i.e. the imputations contain extra information that is not contained in the true data.[2]

We conducted a second simulation study to assess the robustness of FMB when data were missing at random (MAR) and the MDM (for simulating missingness in a bootstrapped dataset) was not modelled; that is, missingness was simulated using the observed MDP as above. The design of this simulation study was based on the simulation study described above, with three modifications. There were three changes. First, data were simulated to be missing dependent upon the outcome variable of the analysis model (loginsindex), thus ensuring the complete case analysis produced biased estimates. Second, the imputation and analysis models were compatible and correctly specified. Third, we conducted two FMB methods: FMB with the MDM correctly modelled (which we shall call FMB correct MDM) and FMB with missingness simulated using the observed MDP (which we shall call FMB observed MDP). For both sample sizes, when data were MAR, omitting to model the MDM for FMB resulted in downwardly biased variance estimates, which led to under-coverage of the confidence intervals. Variance estimates were not downwardly biased when the MDM was correctly modelled (FMB correct MDM). For sample size 1000, any differences in the performances of Rubin's MI, robust Rubin's MI, Robins and Wang's MI and FMB correct MDM were very small. For sample size 100, the Robins and Wang method outperformed the other methods with respect to point estimation but its variance estimates were again downwardly biased. For interval estimation only FMB correct MDM with the percentile confidence interval provided coverage probabilities greater than 0.941 for all coefficients. (Results of this simulation study are available upon request from the authors.)

7 Discussion

We have conducted the first comparative evaluation of Rubin's MI, a modified version of Rubin's MI, Robins and Wang's MI and FMB. Our simulation study shows that Rubin's MI variance estimator failed in four common scenarios of misspecification of the imputation and analysis models, and of incompatibility between them. The variance estimates were substantially upwardly or downwardly biased, resulting in confidence intervals that over- and under-covered, respectively. For moderate sample size (n=1000) and all scenarios apart from severe non-normality, Robins and Wang's MI produced the narrowest confidence intervals on average, with close to nominal coverage. When the imputation and analysis models were both misspecified due to severe non-normality all methods had the same biased imputation estimate, but the larger imputation variance estimate of FMB resulted in confidence intervals with coverage probabilities closest to the nominal level.

A key feature of Rubin's MI is the separation of the imputation procedure from the substantive analysis. This has the advantage that the more technical process of imputation can be done by a specialist, following which multiple analyses can be done on the imputed datasets by non-specialists using standard software. However, this separation may also lead to incompatibility between the imputation and analysis models, when assumptions made during imputation are not carried forward to the analysis stage. For example, estimation for domains (i.e. subgroups or subpopulations) is common for the analysis of survey data and, in particular for large surveys analysed by many users, imputations can be generated ignoring the domain indicator.[24] In some situations, such as the subgroup analysis and omitted interaction scenarios of our simulation study, incompatibility can in principle be avoided, if the imputer provides sufficient documentation of the imputation model to future users. However, in some instances incompatibility may be unavoidable; e.g. for confidentiality reasons the provider of the imputed data may only disseminate to the users a subset of the observations (or records) used during the imputation stage.[11,25]

Misspecification of imputation and analysis models is a more general problem, which will arise whenever model assumptions are not justified. We investigated two scenarios, in which misspecification was due to heteroscedastic errors and non-normality. Our results suggest that when misspecification arises from heteroscedastic errors, there has to be a sizable difference between the subgroup variances in order for Rubin's imputation variance estimator to materially under-estimate the sampling variance of the imputation estimator. In this case, heteroscedastic errors in the imputation and analysis models could have been accommodated by conducting separate MI analyses in men and women.

A limitation of the Robins and Wang MI method is that it makes large sample assumptions, which led to downwardly biased variance estimates and confidence interval coverage when applied to small datasets. A major disadvantage of the Robins and Wang method is that calculation of the imputation variance estimate is considerably more complicated than for Rubin's MI and FMB, with a greater burden placed on both the imputer and the analyst. To our knowledge, there is no generally available software implementing the Robins and Wang method. The analyst must make available derivatives of the estimating equations for use in calculation of variance estimates, and these become harder to calculate as the complexity of the analysis procedure increases. Also, the complexity of the calculations conducted by the imputer increases when there is multiple incomplete variables with a general MDP. For this reason, our simulation scenarios were restricted to data missing in a single variable, as were the scenarios considered in the papers proposing the approach. The Robins and Wang method requires the data to be imputed under a single imputation model. Therefore, currently, it cannot be applied if imputation is conducted using chained equations imputation,[26] a flexible and commonly used method of imputation that imputes under two or more imputation models. In contrast, calculation of the variance of an imputation estimator by Rubin's MI method and FMB is straightforward for more complex MDPs and analysis procedures, and can be applied when data are imputed using chained equations imputation.

A limitation of FMB was its inefficiency relative to the other imputation inference methods. Furthermore, FMB requires modelling of the MDM when data are MAR, which is not required by the MI methods of Rubin or Robins and Wang. A further simulation study we conducted showed that FMB required modelling of the MDM when data were MAR. Further work is needed to compare Rubin's MI with FMB when data are MAR and the missing data model assumed by FMB is incorrectly specified.

In summary, accurate inference requires an unbiased estimator and variance estimator. Rubin's MI variance estimator may be biased in the presence of incompatibility between the imputation and analysis models and model misspecification. This can lead to over- or under-coverage of confidence intervals. These limitations should be noted in guidelines on the appropriate use of Rubin's MI,[27] which should emphasize how incompatibility can be avoided, and the pitfalls that can arise because of model misspecification. The simplicity and flexibility of Rubin's MI mean that it is likely to remain the method of choice to deal with data that are MAR. However, where these problems of incompatibility and misspecification cannot be avoided, Robins and Wang MI has the potential to provide more robust inferences, should the considerable challenges in provision of software implementing the procedure be overcome.

Acknowledgements

The authors acknowledge Professor Andrew Pickles from the Biostatistics Department of the Institute of Psychiatry, King's College London for granting access to data from the British Child and Adolescent Mental Health Study 1999 and the corresponding calibration study.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Medical Research Council grant [G0900724].

References

1. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc, 1987.
2. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc* 1996; **91**: 473–489.
3. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Stat Sci* 1994; **9**: 538–573.
4. Fay RE. Alternative paradigms for the analysis of imputed survey data. *J Am Stat Assoc* 1996; **91**: 490–498.
5. Robins JM and Wang N. Inference for imputation estimators. *Biometrika* 2000; **87**: 113–124.
6. Nielsen SF. Proper and improper multiple imputation. *Int Stat Rev* 2003; **71**: 593–627.
7. Schafer JL. Multiple imputation in multivariate problems when the imputation and analysis models differ. *Stat Neerlandica* 2003; **57**: 19–35.
8. Efron B. Missing data, imputation, and the bootstrap. *J Am Stat Assoc* 1994; **89**: 463–475.
9. Shao J and Sitter RR. Bootstrap for imputed survey data. *J Am Stat Assoc* 1996; **91**: 1278–1288.
10. Collishaw S, Maughan B, Goodman R, et al. Time trends in adolescent mental health. *J Child Psychol Psychiat* 2004; **45**: 1350–1362.
11. Reiter JP. Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika* 2008; **95**: 933–946.
12. Goodman R. The strengths and difficulties questionnaire: a research note. *J Child Psychol Psychiat* 1997; **38**: 581–586.
13. Goodman R. A modified version of the rutter parent questionnaire including extra items on children's strengths: a research note. *J Child Psychol Psychiat* 1994; **35**: 1483–1494.
14. Little RJA and Rubin DB. *Statistical analysis with missing data*. New York: John Wiley & Sons, Inc, 2002.
15. Barnard J and Rubin DB. Small-sample degrees of freedom with multiple imputation. *Biometrika* 1999; **86**: 948–955.
16. Huber PJ. The behavior of maximum likelihood estimates under non-standard conditions. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability,* Berkeley, 1967, 1: pp.221–233. Berkeley: University of California Press.
17. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometricia* 1980; **48**: 817–830.
18. Seber GAF. *A matrix handbook for statisticians*. Hoboken, New Jersey: John Wiley & Sons, Inc, 2008.
19. Rubin DB. Inference and missing data (with discussion). *Biometrika* 1976; **63**: 581–592.
20. Efron B and Tibshirani RJ. *An introduction to the bootstrap*. New York: Chapman and Hall/CRC, 1993.
21. McCarthy A, Hughes R, Tilling K, et al. Birth weight; postnatal, infant, and childhood growth; and obesity in young adulthood: evidence from the Barry Caerphilly growth study. *Am J Clin Nutr* 2007; **86**: 907–913.
22. He Y and Raghunathan TE. On the performance of sequential regression multiple imputation methods with non normal error distributions. *Commun Stat Simul Comput* 2009; **38**: 856–883.
23. White IR, Daniel R and Royston P. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Comput Stat Data Anal* 2010; **54**: 2267–2275.
24. Kim JK, Brick JM, Fuller WA, et al. On the bias of the multiple-imputation variance estimator in survey sampling. *J R Stat Soc Ser B* 2006; **68**: 509–521.
25. Reiter JP and Raghunathan TE. The multiple adaptations of multiple imputation. *J Am Statist Assoc* 2007; **102**: 1462–1471.
26. Van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, et al. Fully conditional specification in multivariate imputation. *J Stat Comput Simulat* 2006; **76**: 1049–1064.
27. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiology and clinical research: potential and pitfalls. *Br Med J* 2009; **339**: 157–160.