A method for increasing the robustness of multiple imputation

Rhian M. Daniel and Michael G. Kenward
Department of Medical Statistics,
London School of Hygiene and Tropical Medicine

Abstract

Missing data are common wherever statistical methods are applied in practice. They present a problem in that they require that additional assumptions be made about the mechanism leading to the incompleteness of the data. By incorporating two models for the missing data process, doubly robust (DR) weighting-based methods offer some protection against misspecification bias since inferences are valid when at least one of the two models is correctly specified. The balance between robustness, efficiency and analytical complexity is one which is difficult to strike, resulting in a split between the likelihood and multiple imputation (MI) school on one hand and the weighting and DR school on the other. A simple extension of MI is proposed that, in certain settings, can be shown to give rise to DR estimators. It is conjectured that this additional robustness holds more generally, as demonstrated using simulation studies. The method is applied to data from the RECORD study, a clinical trial comparing anti-glycaemic combination therapies in type II diabetes patients.
Keywords: Doubly robust estimation; Missing data; Multiple imputation

1 Introduction

Missing data are common wherever statistical methods are applied in practice. When some data are missing, additional assumptions must be made about the mechanism leading to the incompleteness of the data and/or the relationship between the observed and unobserved data. Depending on the method of analysis, these assumptions can take many forms.

One common assumption is that the data are missing at random (MAR) in the sense defined by Rubin [1976]; that is, conditional on the observed data, the probability of observing a partially-observed variable does not depend on the (potentially missing) value taken by that variable. It is impossible to justify MAR using the observed data; as shown by Molenberghs et al. [2008], every missing not at random (MNAR) model has a MAR counterpart fitting the observed data equally well. Thus the plausibility of MAR must be carefully considered using substantive knowledge external to the data. In most practical situations, the best we can hope for is that MAR approximately holds, and that any further dependence on the unobserved data has limited impact on our conclusions. When this is implausible, MNAR sensitivity analyses should be considered [Kenward, 1998, Little and Yau, 1996, Molenberghs et al., 1999, Robins et al., 1999].

Suppose we accept the MAR assumption, or wish to conduct a MAR analysis as a point of departure for sensitivity analyses. The *full data assumptions* are the assumptions we would have made when analysing the full data were they to have been completely observed. When analysing incomplete data under the MAR assumption, further assumptions are often necessary, in addition to the full data assumptions and MAR. Broadly speaking, these additional assumptions fall into two categories: (1) those regarding the form of the conditional distribution of the missing data given the observed data, which—in this paper—we call the *partially-observed data* (POD) model, and (2) those regarding the form of the conditional probability of observing the partially-observed variables given the observed data, which we call the *probability of missingness* (POM) model.

In this paper, we use *full data* (FD) model to refer to the model (fully-, semi- or non-parametric) which would have been specified if the full data had been completely observed. Often, the POD model is left unspecified in the FD model. For example, suppose that the full data analysis would have been a simple linear regression of the outcome Y on explanatory variables X_1 and X_2, but that X_1 is MAR conditional on X_2 and Y. In this case, a model for X_1 conditional on X_2 and Y is not implied by the FD model, and thus further assumptions will be necessary if, say, a maximum likelihood [Little and Rubin, 2002, p.145], expectation-maximisation (EM) [Dempster et al., 1977] or multiple imputation [Rubin, 1987] analysis of the incomplete data is to be conducted: namely, the POD model must be specified. Misspecification of the POD model, in general, induces bias in the resulting estimates.

If a POM model is needed, this is clearly not implied by the FD model. Thus, methods based on inverse probability weighting [dating back to Horvitz and Thompson, 1952], for example, require the additional specification of a POM model. Again, misspecification of the POM model, in general, induces bias in the resulting estimates.

An exception occurs when a maximum likelihood (or Bayesian) analysis is planned and the FD model implies a POD model. This occurs, for example, in a repeated measurements setting when the repeated outcomes are MAR and follow a monotone pattern (that is, if a particular outcome is missing, all subsequent outcomes are also missing) and a multivariate normal model is assumed (*e.g.* in a random intercepts and slopes analysis). In practice, however, the model for the full data is unlikely to be precisely correct, and the consequences of departures from this model are more serious when the data are incomplete. In addition, the fit of the FD model can only be assessed for the observed data. Even if the FD model appears to fit the observed data well, the assumption that it also fits the unobserved data well can rely on considerable extrapolation when the observed and unobserved data differ substantially on the values of some variables. Even apparently mild misspecification of the FD model can lead to bias in the resulting estimates.

So-called *doubly robust* (DR) methods [Bang and Robins, 2005, Creemers et al., 2011, Robins et al., 1994, Scharfstein et al., 1999, Tsiatis, 2006] specify (in addition to MAR and a FD model) *both* a POD and POM model, and the results from such an analysis have been shown to be valid when (in addition to the assumptions of MAR and the FD model) *at least one* of the POD and POM models is correctly-specified. In particular, they have mainly been proposed in situations in which the FD model is either semi- or even non-parametric, thus increasing the chances of correctly specifying the FD model. This (partial) protection against model misspecification makes DR methods attractive in many settings. Until recently, however, the absence of a general method for deriving DR estimators coupled with the complex mathematical underlying theory, has meant that, since their introduction in the 1990s, DR estimators have not been very widely used in practice. A paper by Bang and Robins [2005], in which a general method, implemented using existing software, is described both for cross-sectional univariate missing data and longitudinal data with monotone dropout, constitutes therefore a very important advance in the literature on this topic.

In this paper, we use the idea proposed by Bang and Robins [2005] within the multiple imputation [Rubin, 1987] framework to improve the robustness of MI estimators. In certain settings we show that the resulting estimators are DR. More generally, for example when the pattern of missingness is non-monotone and using chained equations [van Buuren et al., 1999], we conjecture that our approach leads to improved robustness, and demonstrate this using simulation studies. Our approach is motivated from two different perspectives. First, we aim to reduce the reliance of multiple imputation on the correct specification of the imputation model, and second, since multiple imputation is the only method that can easily deal with non-monotone patterns of missingness, it is a natural starting point when seeking doubly robust estimators for incomplete data with such a missingness pattern.

The outline of the remainder of this paper is as follows. We start, in Section 2, with an overview of some of the relevant theory from the missing data literature. In particular, we discuss randomised monotone missingness (RMM) processes, augmented inverse probability weighted (AIPW) estimators and multiple imputation with chained equations (MICE). In Section 3, we describe our proposed method, before demonstrating its properties in Section 4, using simulation

studies. In Section 5, we demonstrate how our method is implemented in practice using data from a clinical trial comparing three different anti-glycaemic drugs for type II diabetes patients. This dataset contains repeated measurements of HbA_{1c}, subject to non-monotone missingness.

2 Theoretical background

2.1 Full data estimating equation

Let $\mathbf{Z}_i = (Z_{1,i}, Z_{2,i}, \ldots, Z_{J,i})^T$ be the full data vector on subject i, that is, the data on J variables which would have been observed on subject i in an ideal setting with no missing data, where $i \in \{1, \ldots, n\}$. Let $\mathbf{R}_i = (R_{1,i}, R_{2,i}, \ldots, R_{J,i})^T$ be the corresponding vector of (non-)missingness indicators, where $R_{j,i} = I(Z_{j,i}$ is observed), and $I(\cdot)$ denotes the usual indicator function.

In the absence of missing data, we suppose that the following estimating equation

$$\sum_{i=1}^{n} \mathbf{U}_{\boldsymbol{\theta}}(\mathbf{Z}_i, \boldsymbol{\theta}^{\text{full}}) = \mathbf{0}$$

would be solved to estimate $\boldsymbol{\theta}$, a vector of parameters governing the distribution of $\mathbf{Z}_i$.

$\mathbf{U}_{\boldsymbol{\theta}}(\cdot)$ defines the FD model. If $\mathbf{U}_{\boldsymbol{\theta}}(\cdot)$ is a score function, $\boldsymbol{\theta}$ represents the parameters of a fully-parametric specification of the distribution of $\mathbf{Z}_i$. More often, the model defined by $\mathbf{U}_{\boldsymbol{\theta}}(\cdot)$ is semi- or even non-parametric. $\hat{\boldsymbol{\theta}}^{\text{full}}$ is a consistent estimator of $\boldsymbol{\theta}$ if and only if $E\{\mathbf{U}_{\boldsymbol{\theta}}(\mathbf{Z}_i, \boldsymbol{\theta}_0)\} = \mathbf{0}$ at the true value $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}$ [Cox and Hinkley, 1974].

2.2 Complete case estimating equation

When the data are incomplete, the complete case (CC) estimator $\hat{\boldsymbol{\theta}}^{\text{CC}}$ is the solution to

$$\sum_{i=1}^{n} I\left\{\mathbf{R}_i = \mathbf{1}^{(J\times 1)}\right\} \mathbf{U}_{\boldsymbol{\theta}}(\mathbf{Z}_i, \boldsymbol{\theta}^{\text{CC}}) = \mathbf{0}.$$

Here, $\mathbf{1}^{(J\times 1)}$ is used to denote a $(J \times 1)$ unit vector. Even when we assume that the full data model is correct (*i.e.* that $\hat{\boldsymbol{\theta}}^{\text{full}}$ is a consistent estimator of $\boldsymbol{\theta}$), it is easily shown that for the expectation of $I\left\{\mathbf{R}_i = \mathbf{1}^{(J\times 1)}\right\} \mathbf{U}_{\boldsymbol{\theta}}(\mathbf{Z}_i, \boldsymbol{\theta}_0)$ to be zero, $P\left\{\mathbf{R}_i = \mathbf{1}^{(J\times 1)} \middle| \mathbf{Z}_i\right\}$ must be independent of $\mathbf{Z}_i$, *i.e.* the data must be missing completely at random (MCAR) as defined by Rubin [1976], an assumption which is unlikely to hold in most practical settings.

2.3 Infeasible inverse probability weighted complete case estimating equation

To correct for the bias in the CC estimator whenever data are not MCAR, we can weight the contributions to the estimating equation according to the inverse probability of $\mathbf{R}_i = \mathbf{1}^{(J\times 1)}$ given $\mathbf{Z}_i$. The resulting inverse probability weighted complete case (IPW) estimating equation is

$$\sum_{i=1}^{n} \frac{I\left\{\mathbf{R}_i = \mathbf{1}^{(J\times 1)}\right\}}{P\left\{\mathbf{R}_i = \mathbf{1}^{(J\times 1)} | \mathbf{Z}_i\right\}} \mathbf{U}_{\boldsymbol{\theta}}(\mathbf{Z}_i, \boldsymbol{\theta}^{\text{i-IPW}}) = \mathbf{0}$$

and it is trivial to show that the Cox and Hinkley condition (zero expectation of the estimating function) holds whenever the FD model is correct. Intuitively, we assign a high weight to subjects with a low conditional probability of being fully-observed, so that they represent subjects with similar characteristics who were not fully-observed and thus are not included in the CC analysis.

We refer to this estimator as $\hat{\boldsymbol{\theta}}^{\text{i-IPW}}$, the *infeasible* [*cf.* Robins et al., 1992] IPW estimator, since, in most practical settings $P\left\{\mathbf{R}_i = \mathbf{1}^{(J\times 1)} \middle| \mathbf{Z}_i\right\}$ is unknown and must be estimated from the data, using the POM model under the MAR assumption.

2.4 The probability of missingness (POM) model

2.4.1 Univariate missing data

Consider first the simple setting in which only variable Z_j, say, is incomplete, where $Z_j = (Z_{j,1}, Z_{j,2}, \ldots, Z_{j,n})$. Given the MAR assumption, that $P(R_{j,i} = 1 | \mathbf{Z}_i)$ is independent of $Z_{j,i}$, we could estimate $P\left\{\mathbf{R}_i = \mathbf{1}^{(J\times 1)} \middle| \mathbf{Z}_i\right\}$ by fitting, for example, a regression model

$$E(R_{j,i} | Z_{1,i}, \ldots, Z_{j-1,i}, Z_{j+1,i}, \ldots, Z_{J,i}) = l(Z_{1,i}, \ldots, Z_{j-1,i}, Z_{j+1,i}, \ldots, Z_{J,i}; \boldsymbol{\alpha}).$$

The choice of $l(\cdot)$ (for example, inverse logit with each variable included in a linear predictor and no interactions) defines a POM model, giving rise to estimates

$$\hat{\pi}_i = \pi(Z_{1,i}, \ldots, Z_{j-1,i}, Z_{j+1,i}, \ldots, Z_{J,i}; \hat{\boldsymbol{\alpha}})$$

of $P\left\{\mathbf{R}_i = \mathbf{1}^{(J\times 1)} \middle| \mathbf{Z}_i\right\}$.

2.4.2 Multivariate monotone missing data

Write the full data $\mathbf{Z}_i$ for subject i as $\mathbf{Z}_i = \left(\mathbf{X}_i^T, \mathbf{Y}_i^T\right)^T$, where $\mathbf{X}_i$ is observed with probability 1 and $\mathbf{Y}_i = \left(Y_{1,i}, Y_{2,i}, \ldots, Y_{\tilde{J},i}\right)^T$ may be incompletely observed. $R_{j,i}$ now refers to $I(Y_{j,i}$ is observed). The missingness pattern is monotone if there exists a permutation $\psi(\cdot)$ of $\left(1, \ldots, \tilde{J}\right)$ for which $R_{\psi(j),i} = 0 \Rightarrow R_{\psi(k),i} = 0 \quad \forall \psi(k) > \psi(j)$; that is, that there exists an ordering of $\left\{Y_j : j = 1, \ldots \tilde{J}\right\}$ such that if $Y_{j,i}$ is missing, $Y_{k,i}$ is also missing for every k later than j in the permuted sequence. A common example of this occurs with dropout in repeated measurements studies.

The POM model can be defined for multivariate monotone missing data using a sequential extension of the POM model described in Section 2.4.1. Without loss of generality, we redefine $\left(1, \ldots, \tilde{J}\right)$ such that $\psi(\cdot)$ is the identity. First, $P(R_{1,i} = 1 | \mathbf{X}_i)$ is estimated from the model $E(R_{1,i}) = l_1(\mathbf{X}_i; \boldsymbol{\alpha}_1)$. Second, $P(R_{2,i} = 1 | \mathbf{X}_i, Y_{1,i}, R_{1,i} = 1)$ is estimated from the model $E(R_{2,i} | \mathbf{X}_i, Y_{1,i}) = l_2(\mathbf{X}_i, Y_{1,i}; \boldsymbol{\alpha}_2)$ fitted only to those who have Y_1 observed. At the jth step, $P(R_{j,i} = 1 | \mathbf{X}_i, Y_{1,i}, \ldots, Y_{j-1,i}, R_{j-1,i} = 1)$ is estimated from the model $E(R_{j,i} | \mathbf{X}_i, Y_{1,i}, \ldots, Y_{j-1,i}) = l_j(\mathbf{X}_i, Y_{1,i}, \ldots, Y_{j-1,i}; \boldsymbol{\alpha}_j)$ fitted only to those who have Y_{j-1} (and hence all previous Y's) observed. At each stage $j = 1, \ldots, \tilde{J}$, conditional probability estimates $\hat{\lambda}_{j,i} = l_j(\mathbf{X}_i, Y_{1,i}, \ldots, Y_{j-1,i}; \hat{\boldsymbol{\alpha}}_j)$ are obtained. Finally, $\hat{\pi}_i = \prod_{j=1}^{\tilde{J}} \hat{\lambda}_{j,i}$.

That $\hat{\pi}_i$ is a consistent estimator of $P\left\{\mathbf{R}_i = \mathbf{1}^{\left(\tilde{J}\times 1\right)} \middle| \mathbf{Y}_i\right\}$ under MAR and the assumption that the POM model defined by $l_2(\cdot), \ldots, l_{\tilde{J}}(\cdot)$ is correctly specified, is formally shown by Molenberghs et al. [1998].

2.4.3 Non-monotone missing data

Doubts have been cast over the appropriateness of the MAR assumption for non-monotone missing data [Robins and Gill, 1997]. These authors introduce a new mechanism, *randomised monotone missingness* (RMM)—a subset of MAR—and argue that this is the only plausible non-monotone MAR mechanism that is not MCAR. They show [Gill and Robins, 1996] that there exist mechanisms that are MAR but not RMM, but that for a computer to generate data under such a mechanism, it requires knowledge of the unobserved data which is then 'concealed' later in the process. They call this phenomenon 'MAR is more than it seems' and say:

> "We have been unable to conceive of a plausible social, economic, physical or biological process that would generate MAR processes that are not RMM representable, due to the subtle and precise manner in which the data must be 'hidden' to insure that the process is MAR. That is, we believe that natural missing data processes that are not representable as RMM processes will be [MNAR]."

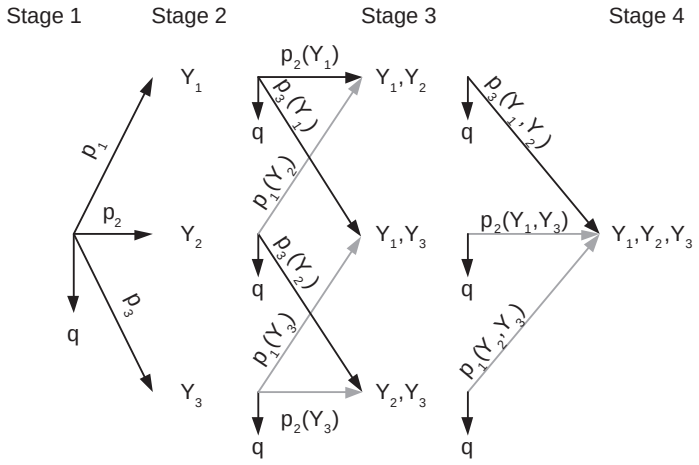

Figure 1: A Markov randomised monotone missingness process for $\tilde{J} = 3$. Dependence on $\mathbf{X}$ is implicit. The probabilities associated with the grey lines are all zero if the data are longitudinal.

Again we write $\mathbf{Z}_i = \left(\mathbf{X}_i^T, \mathbf{Y}_i^T\right)^T$, where $\mathbf{X}_i$ is observed with probability 1 and $\mathbf{Y}_i = \left(Y_{1,i}, Y_{2,i}, \ldots, Y_{\tilde{J},i}\right)^T$ may be incompletely observed. The RMM process for subject i is described by Robins and Gill as follows. We start by observing $\mathbf{X}_i$. Then, we either observe one of $Y_{j,i}$ $(j = 1, \ldots, \tilde{J})$ with probabilities $p_{j,i}(\mathbf{X}_i)$ $(j = 1, \ldots, \tilde{J})$, respectively, or we quit, having only observed $\mathbf{X}_i$ with probability $1 - \sum_{j=1}^{\tilde{J}} p_{j,i}(\mathbf{X}_i)$. Suppose we in fact observe $Y_{j_1,i}$. Now, at the next stage, we either observe one of $Y_{j,i}$ $(j = 1, \ldots, j_1 - 1, j_1 + 1, \ldots \tilde{J})$ with probabilities $p_{j,i}(\mathbf{X}_i, Y_{j_1,i})$ $(j = 1, \ldots, j_1 - 1, j_1 + 1, \ldots \tilde{J})$, respectively, or we quit, having only observed $\mathbf{X}_i$ and $Y_{j_1,i}$ with probability $1 - \sum_{j \neq j_1} p_{j,i}(\mathbf{X}_i, Y_{j_1,i})$. Suppose that after m stages of this algorithm, we have observed $\mathbf{X}_i, Y_{j_1,i}, Y_{j_2,i}, \ldots,$ and $Y_{j_{m-1},i}$. At the next stage, we either observe one of $Y_{j,i}$ $\left(j \in \left\{1, \ldots, \tilde{J}\right\} \setminus \{j_1, j_2, \ldots, j_{m-1}\}\right)$ with probabilities $p_{j,i}\left(\mathbf{X}_i, Y_{j_1,i}, \ldots, Y_{j_{m-1},i}\right)$ $\left(j \in \left\{1, \ldots, \tilde{J}\right\} \setminus \{j_1, j_2, \ldots, j_{m-1}\}\right)$, respectively, or we quit with probability $1 - \sum_{j \neq j_1, \ldots, j_{m-1}} p_{j,i}\left(\mathbf{X}_i, Y_{j_1,i}, \ldots, Y_{j_{m-1},i}\right)$.

Markov randomised monotone missingness (MRMM) mechanisms are a particular subset of RMM mechanisms in which the probability of observing a given variable conditional on the previous variables observed is independent of the order in which these variables were observed. Thus, for example, $p_{j,i}(\mathbf{X}_i, Y_{j_1,i}, Y_{j_2,i}) = p_{j,i}(\mathbf{X}_i, Y_{j_2,i}, Y_{j_1,i})$. Gill and Robins [1996] prove that any MAR mechanism representable as RMM is also representable as MRMM. The MRMM process when $\tilde{J} = 3$ is shown in Figure 1.

In Figure 1, p_1 is the probability (conditional on covariates $\mathbf{X}$) that, the first variable observed is Y_1. Similarly, p_2 and p_3 are the probabilities that, the first variable observed is Y_2 and Y_3, respectively. With probability $1 - p_1 - p_2 - p_3$, none of Y_1, Y_2 or Y_3 is observed. Then, for the second stage, $p_2(Y_1)$ is the probability that (conditional on covariates $\mathbf{X}$, the fact that Y_1 was observed at the first stage, and the value of Y_1 itself), the next variable to be observed is Y_2, and so on. Finally, $p_3(Y_1, Y_2)$ is the probability that (conditional on covariates $\mathbf{X}$, the fact that Y_1 and Y_2 have already been observed, in any order, and conditional on their values, Y_1 and Y_2), the third variable to be observed is Y_3, and so on.

Notice that (omitting the subscript i), for example,

$$P(Y_1, Y_2, Y_3 \text{ all observed}) = p_1 p_2(Y_1) p_3(Y_1, Y_2) + p_1 p_3(Y_1) p_2(Y_1, Y_3) + p_2 p_1(Y_2) p_3(Y_1, Y_2) + p_2 p_3(Y_2) p_1(Y_2, Y_3) + p_3 p_1(Y_3) p_2(Y_1, Y_3) + p_3 p_2(Y_3) p_1(Y_2, Y_3),$$

where these six terms are not constrained to be equal. Thus, the order in which the variables were observed is needed to estimate the probabilities $p_j(\cdot)$—even in an MRMM process—but this order is never observed.

At each stage in the process, $p_{j,i}$ are estimated from models (for example, multinomial logistic regressions) which together constitute a POM model. Robins and Gill [1997] describe a method for estimating $\pi_i = P\left\{\mathbf{R}_i = \mathbf{1}^{(\tilde{J}\times 1)} \middle| \mathbf{Z}_i\right\}$ from a MRMM mechanism, using an EM algorithm in which the unobserved orderings are treated as missing data. $\hat{\pi}_i$ is then a consistent estimator of π_i under the MAR assumption if the assumptions of the POM model are correct.

Note that, as the number of time-points increases, the computation involved in this procedure increases geometrically. When $\tilde{J}$ is large, Robins and Gill [1997] propose a simulated EM algorithm that greatly diminishes this computational burden.

In the case of non-monotone longitudinal repeated measurements, there exists only one plausible (*i.e.* temporal) ordering and MRMM reduces to a very special case (see Figure 1 with grey lines omitted) in which the probability of observing Y_3, say, is dependent on Y_2 if and only if Y_2 has been observed. As Vansteelandt et al. [2007] argue, it is implausible in most settings that the probability of observing Y_3 only depends on Y_2 if Y_2 happens to have been observed and that therefore, MAR is rarely a sensible assumption for non-monotone repeated measurements. However, as a point of departure for sensitivity analyses it is useful to be aware of the form of this MAR mechanism.

Obtaining $\hat{\pi}_i$ is then much more straightforward than in the general case, since the order in which the variables were observed is always known. More details are given in Appendix A.

2.5 Feasible inverse probability weighted complete case estimating equation

If both the FD and POM models are correctly specified, and the MAR assumption holds, the solution $\hat{\boldsymbol{\theta}}^{\text{IPW}}$ to

$$\sum_{i=1}^{n} \hat{\pi}_i^{-1} I\left\{\mathbf{R}_i = \mathbf{1}^{(J\times 1)}\right\} \mathbf{U}_{\boldsymbol{\theta}}\left(\mathbf{Z}_i, \boldsymbol{\theta}^{\text{IPW}}\right) = \mathbf{0}$$

is a consistent estimator of $\hat{\boldsymbol{\theta}}$.

2.6 Augmented inverse probability weighted estimating equation

Although $\hat{\boldsymbol{\theta}}^{\text{IPW}}$, under the assumptions mentioned above, is a consistent estimator of $\boldsymbol{\theta}$, it is, in general, inefficient [Robins et al., 1995]. Its efficiency can be particularly poor when the complete cases account for only a small proportion of the observed data.

Robins et al. [1994] show that, by considering estimating equations of the form

$$\sum_{i=1}^{n} \left(\hat{\pi}_i^{-1} I\left\{\mathbf{R}_i = \mathbf{1}^{(J\times 1)}\right\} \mathbf{U}_{\boldsymbol{\theta}}\left(\mathbf{Z}_i, \boldsymbol{\theta}^{\text{AIPW}}\right) + \left[1 - \hat{\pi}_i^{-1} I\left\{\mathbf{R}_i = \mathbf{1}^{(J\times 1)}\right\}\right] \phi\left(\mathbf{R}_i, \mathbf{Z}_i^{\text{obs}}, \boldsymbol{\theta}^{\text{AIPW}}\right)\right) = \mathbf{0}, \tag{1}$$

where $\mathbf{Z}_i^{\text{obs}}$ is the observed part of $\mathbf{Z}_i$, the efficiency can be increased.

Using the semiparametric theory based on influence functions and Hilbert spaces [Tsiatis, 2006], Robins *et al.* show that, for a particular $\mathbf{U}_{\boldsymbol{\theta}}(\mathbf{Z}_i, \boldsymbol{\theta})$, if there is only one variable subject to missingness, the most efficient choice for $\phi(\cdot)$ is

$$\phi\left(\mathbf{R}_i, \mathbf{Z}_i^{\text{obs}}, \boldsymbol{\theta}\right) = E\left\{\mathbf{U}_{\boldsymbol{\theta}}\left(\mathbf{Z}_i, \boldsymbol{\theta}\right) | \mathbf{Z}_i^{\text{obs}}\right\}.$$

When there are many variables subject to missingness, in a monotone pattern, then the same theory shows that the most efficient choice of $\phi\left(\mathbf{R}_i, \mathbf{Z}_i^{\text{obs}}, \boldsymbol{\theta}\right)$ is

$$\sum_{j=1}^{J} R_{j-1,i}\left(\frac{\hat{\pi}_{j,i}}{\hat{\pi}_{j-1,i}} - R_{j,i}\right) \frac{E\left\{\mathbf{U}_{\boldsymbol{\theta}}\left(\mathbf{Z}_i, \boldsymbol{\theta}\right) | \mathbf{Z}_i^{\text{obs}}\right\}}{\hat{\pi}_{j,i}},$$

where $\hat{\pi}_{j,i} = \prod_{k=1}^{j} \hat{\lambda}_{k,i}$ and $R_{0,i} = 1 \; \forall i$.

In both cases above, $\phi\left(\mathbf{R}_i, \mathbf{Z}_i^{\text{obs}}, \boldsymbol{\theta}\right)$ is a function of $E\left\{\mathbf{U}_{\boldsymbol{\theta}}\left(\mathbf{Z}_i, \boldsymbol{\theta}\right) | \mathbf{Z}_i^{\text{obs}}\right\}$. Since $\mathbf{U}_{\boldsymbol{\theta}}\left(\cdot\right)$ is a function of the missing and observed data, we can specify its conditional expectation given the observed data using a model for the missing data given the observed, *i.e.* a POD model. Thus we see that DR estimators involve *both* a POM and POD model. Scharfstein et al. [1999] were the first to show that, when MAR and the assumptions of the FD model hold, the DR estimator is consistent if, in addition, *at least one* of POM and POD are correctly specified. Thus DR estimators offer two advantages over their IPW counterparts: increased efficiency, and double robustness.

2.7 Multiple imputation with chained equations

An entirely different approach to the analysis of incomplete data is multiple imputation (MI) [Rubin, 1987]. Briefly, in MI, multiple stochastic imputations are drawn for each missing value in the dataset, using an *imputation model*. Each completed dataset is then analysed using the FD model and estimates and standard errors of the parameters of interest are obtained from each completed dataset. The standard errors from any particular completed dataset will be too small, since the imputed data are treated as if they had been observed. However, by comparing these FD estimates *between* imputed datasets, Rubin gives a formula for the standard errors, in which the between-dataset estimates inflate the within-dataset estimates. For this formula to lead to correct inferences, Rubin showed that the imputations need to be *proper*, in the sense that the stochastic draws must be drawn—not from the imputation model with each parameter replaced by a consistent estimate—but rather using Bayesian draws from the posterior distributions of these parameters.

In a fully-parametric setting, the imputation model is the POD model implied by the FD model. However, in settings when we cannot (or do not wish to) specify the joint distribution of the full data, *multiple imputation using chained equations* (MICE) [van Buuren et al., 1999] can be used. This approach works by specifying a univariate regression model for each partially observed variable, conditional on all the others. The method is practically very attractive as it can deal with missing data on a large number of variables with different univariate distributions and without requiring that the missing data pattern be monotone. Many software implementations exist: ice in Stata [Royston, 2004], mice in S-Plus and R [van Buuren and Oudshoorn, 2000], and IVEware in SAS [Raghunathan et al., 2007]. However, a full theoretical argument for the validity of this method has not been presented to date. Indeed, it is unlikely that such a proof exists except in the multivariate normal case, since a collection of univariate regression models—not all linear—in general does not correspond to a well-defined joint distribution. In other words, the univariate imputation models used in MICE together constitute the POD model, but there might not be a FD model which supports this POD model. The imputation model is then said to be *uncongenial*, *i.e.* the stationary distribution to which the Gibbs sampler attempts to converge does not exist. However, simulation studies suggest that the bias caused by the uncongeniality is likely to be small in practice [Gelman and Raghunathan, 2001, van Buuren et al., 2006].

MI uses only a POD model (in addition to MAR and the FD model), and when this model is correctly specified, MI is, in general, more efficient than DR estimation [Carpenter et al., 2006]. This occurs, however, at an increased risk of model misspecification, since a POM model is not involved, and thus it is not doubly robust.

3 Robust Multiple Imputation: the proposed method

3.1 Univariate MAR missing data

Let the full data $\mathbf{Z}_i = \left(\mathbf{X}_i^T, Y_i\right)^T$ for subject $i \in \{1, \ldots, n\}$ be a fully-observed vector of covariates $\mathbf{X}_i$ and a scalar outcome Y_i which could be missing ($R_i = 0$) or observed ($R_i = 1$). For simplicity,

we assume that interest lies is in estimating $\mu = E(Y_i)$, but the method could be applied more generally (*e.g.* for estimating the coefficients of a generalised linear model).

Following the same idea as proposed by Bang and Robins [2005], first a suitable regression model (such as logistic regression) is chosen for R conditional on $\mathbf{X}$—the POM model. Let $\hat{\boldsymbol{\alpha}}$ be the parameter estimates from this regression and let $\hat{\pi}_i = \pi(\mathbf{X}_i, \hat{\boldsymbol{\alpha}})$ be the predicted probabilities (that $R_i = 1$) from this model.

Next, we fit a suitable regression model for Y conditional on $\mathbf{X}$ and $\hat{\pi}^{-1}$ to those subjects who have complete data. We call the corresponding model *without* the inverse probability weights, *i.e.*

$$E(Y_i \,|\mathbf{X}_i, R_i = 1) = \Psi\{s(\mathbf{X}_i, \boldsymbol{\beta})\}, \tag{2}$$

the POD model, where $\Psi^{-1}(\cdot)$ is the canonical link function from an appropriate GLM and $s(\mathbf{X}, \boldsymbol{\beta})$ is a known function of $\boldsymbol{\beta}$ and $\mathbf{X}$. We call

$$E\left\{Y_i \left| \mathbf{X}_i, \hat{\pi}^{-1}(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}), R_i = 1 \right.\right\} = \Psi\left\{s(\mathbf{X}_i, \boldsymbol{\beta}) + \phi\hat{\pi}_i^{-1}\right\} \tag{3}$$

the *extended* POD model.

Let

$$\hat{e}\left(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\pi}_i^{-1}\right) = \Psi\left\{s\left(\mathbf{X}_i, \hat{\boldsymbol{\beta}}\right) + \hat{\phi}\hat{\pi}_i^{-1}\right\} \tag{4}$$

be the predictions from the extended POD model.

Now we draw $m > 1$ imputations, $Y_{i,j}^* : j = 1, \ldots, m$ for each of the missing Y_i's based on the extended POD model using the *proper* imputation procedure described by Rubin [1987]. The details are given in Appendix B. Then we set

$$\tilde{Y}_i^{(j)} = R_i Y_i + (1 - R_i) Y_{i,j}^*.$$

Finally, our proposed estimator is the solution $\hat{\mu}_{\text{RMI}}$ to

$$\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\tilde{Y}_i^{(j)} - \mu_{\text{RMI}}\right\} = 0.$$

Theorem 3.1 (Multiply imputed DR univariate estimator)**.** *The estimator $\hat{\mu}_{\text{RMI}}$ is doubly robust. That is, if at least one of the two models (the POM model and the POD model) is correctly specified (but not necessarily both), in addition to the MAR assumption and the assumptions of the FD model, $\hat{\mu}_{\text{RMI}}$ is a consistent estimator of μ.*

A sketch proof is given in Appendix C.

The variance of our estimator $\text{Var}(\hat{\mu}_{\text{RMI}})$ could be estimated using Rubin's variance formula as

$$\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{\left\{\tilde{Y}_i^{(j)} - \bar{\tilde{Y}}^{(j)}\right\}^2}{n-1} + \frac{m+1}{m}\sum_{j=1}^{m}\frac{\left\{\bar{\tilde{Y}}^{(j)} - \bar{\tilde{Y}}\right\}^2}{m-1} \tag{5}$$

where $\bar{\tilde{Y}}^{(j)} = \frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_i^{(j)}$ and $\bar{\tilde{Y}} = \frac{1}{m}\sum_{j=1}^{m}\bar{\tilde{Y}}^{(j)}$.

However, this variance estimator has two important drawbacks:

1. It treats the weights as just another covariate in the imputation model. Thus the variance estimator is conditional on $\hat{\pi}^{-1}(X_i, \hat{\alpha})$ and ignores the fact that these weights are estimated from the data.

2. Putting this problem to one side, when the POD model is correctly specified, the fact that the weights are treated as just another covariate justifies the use of Rubin's variance formula. In other words, if the weights were not estimated, the correct specification of the POD model would render (5) a consistent estimator of the variance of $\hat{\mu}_{\text{RMI}}$, by the standard argument for the consistency of Rubin's variance formula in (non-DR) ordinary multiple imputation. However, if the POD model is misspecified, but the POM model correctly specified, there is no reason to suppose that (5) remains consistent. Hence, our proposed variance formula is (ignoring the added problem noted in 1.) singly robust, but does not inherit the DR property of the estimator itself.

If the robust MI procedure is used as a sensitivity analysis, then the limitations above may be acceptable; otherwise, a bootstrap estimator of variance may be preferred.

Note that the consistency of $\hat{\mu_{\text{RMI}}}$ when the POD model is correctly specified (but not necessarily the POM model) relies on the fact that the true value of ϕ (as defined in (3)) in this situation is zero. In finite samples, however, the value of $\hat{\phi}$ in (4) will not be exactly zero. When the weights are very variable (which can happen, for example, if the conditional probability of having Y_i observed given $\mathbf{X}_i$ is close to zero for some i), $\hat{\phi}$ could be considerably different from zero leading to large finite sample bias and instability in the estimator $\hat{\mu_{\text{RMI}}}$. This problem, along with some proposed solutions, have received considerable attention in the recent literature. See the final paragraph of the Discussion for more details. This comment applies to all the robust MI estimators discussed in the remainder of the paper.

Relationship to other estimators Bang and Robins [2005] discuss what they call the outcome regression (OR) estimator, which is the solution to

$$\sum_{i=1}^{n} \left\{ \hat{e}\left(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}\right) - \mu_{\text{OR}} \right\} = 0,$$

where $\hat{e}\left(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}\right)$ are the predictions from the (non-extended) POD model (2). This is equivalent to a maximum likelihood analysis.

The doubly robust estimator proposed by Bang and Robins [2005] is the solution to

$$\sum_{i=1}^{n} \left\{ \hat{e}\left(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\pi}_i^{-1}\right) - \mu_{\text{DR}} \right\} = 0,$$

where $\hat{e}\left(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\pi}_i^{-1}\right)$ are as defined in (4).

3.2 Longitudinal MAR missing data

The same idea can be extended to the case of multivariate missing data, and—unlike the Bang and Robins [2005] approach—the pattern need not be monotone.

Let the full data $\mathbf{Z}_i = \left(\mathbf{X}_i^T, \mathbf{Y}_i^T\right)^T$ for subject $i \in \{1, \ldots, n\}$ consist of a fully-observed vector of covariates $\mathbf{X}_i$ and a vector of partially-observed outcome variables $\mathbf{Y}_i = (Y_{1,i}, \ldots, Y_{T,i})^T$ and that interest lies in estimating $\mu = E(Y_{i,T})$. Let $\mathbf{R}_i = (R_{1,i}, \ldots, R_{T,i})^T$ be the vector of missingness indicators with $R_{t,i} = I(Y_{t,i} \text{ is observed})$.

We first describe the RMI method for monotone longitudinal data before moving to the case of non-monotone longitudinal data in Section 3.2.2.

3.2.1 Monotone longitudinal data

When the missingness pattern is monotone, we can easily estimate $\hat{\pi}_{t,i} = P(R_{t,i} = 1 \,|\mathbf{Z}_i) = P\left(R_{t,i} = 1 \,\middle|\, \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}\right)$ at each time t, as described in Section 2.4.2.

We proceed by fitting the model using MI. The POD model is postulated sequentially by first specifying a model for Y_1 given $\mathbf{X}$, and then a model for Y_2 given Y_1 and $\mathbf{X}$ etc. To construct an extended POD model, for each $t \in \{1, \ldots, T\}$, $\hat{\pi}_{t,i}^{-1} = P(R_{t,i} = 1 \,|\mathbf{X}_i, Y_{1,i}, \ldots, Y_{t-1,i})^{-1}$ is included as an additional covariate, additional to $\mathbf{X}$ and $\bar{\mathbf{Y}}_{t-1}$, in the model for $Y_{t,i}$. Starting with Y_1, any missing values in Y_1 are multiply imputed, with the imputations drawn from the extended POD model for Y_1 conditional on $\mathbf{X}$ and $\hat{\pi}_1^{-1}$. Next, any missing values in Y_2 are multiply imputed, with the imputations drawn from the extended POD model for Y_2 conditional on Y_1, $\mathbf{X}$ and $\hat{\pi}_2^{-1}$; for subjects with Y_1 also missing, the imputed value of Y_1 from the jth imputed dataset is used to impute Y_2 in the jth imputed dataset, and so on.

By starting with Y_1 and working upwards in this way, we encounter a problem which does not arise in the method proposed by Bang and Robins [2005], which starts with Y_T and works

downwards. The problem is that $\hat{\pi}_{t,i}$ can only be calculated for subjects who have $Y_{t-1,i}$ observed, but (unlike Bang and Robins [2005]), we require that $\hat{\pi}_{t,i}$ be known for all subjects.

Suppose a particular subject, i_1, drops out after being observed at time $t-2$. At time $t-1$, in the jth imputed dataset, a value $\tilde{Y}^{(j)}_{t-1,i_1}$ of Y_{t-1,i_1} is imputed, based on $\mathbf{X}_{i_1}$, $\bar{\mathbf{Y}}_{t-2,i_1}$, and $\hat{\pi}_{t-1,i_1}$, which are all observed. But at the next timepoint, t, we would like to impute the missing Y_{t,i_1} using $\mathbf{X}_{i_1}$, $\bar{\mathbf{Y}}_{t-2,i_1}$, $\tilde{Y}^{(j)}_{t-1,i_1}$, and $\hat{\pi}_{t,i_1}$. The marginal probability $\hat{\pi}_{t,i_1}$ is the product of $\hat{\pi}_{t-1,i_1}$ and $\hat{\lambda}\left(t \left| \mathbf{X}_{i_1}, \bar{\mathbf{Y}}_{t-1,i_1} \right.\right)$, the estimate of the conditional probability that $R_{t,i_1} = 1$, conditional on $\mathbf{X}_{i_1}$, $\bar{\mathbf{Y}}_{t-1,i_1}$, and $R_{t-1,i_1} = 1$, as defined in section 2.4.2. It is this latter conditional probability which cannot be estimated directly for this subject. However, as a function of the missing Y_{t-1,i_1}, it is known. Thus our proposed method works by imputing a value for $\hat{\pi}_{t,i_1}$, based on $\hat{\pi}_{t-1,i_1}$, $\hat{\lambda}\left(t \left| \mathbf{X}_{i_1}, \bar{\mathbf{Y}}_{t-1,i_1} \right.\right)$ and $\tilde{Y}^{(j)}_{t-1,i_1}$ as follows:

$$\hat{\pi}^{(j)}_{t,i_1} = \hat{\pi}_{t-1,i_1} \hat{\lambda}\left(t \left| \mathbf{X}_{i_1}, \bar{\mathbf{Y}}_{t-2,i_1}, \tilde{Y}^{(j)}_{t-1,i_1} \right.\right).$$

In other words, no additional model is fitted to obtain the imputation $\hat{\pi}^{(j)}_{t,i_1}$, and no additional draws (for $\hat{\pi}^{(j)}_{t,i_1}$), nor additional draws from the Bayesian posterior distribution of any additional parameters are made. Rather, $\hat{\pi}^{(j)}_{t,i_1}$ is imputed as a deterministic function of $\hat{\pi}_{t-1,i_1}$ and $\hat{\lambda}\left(t \left| \mathbf{X}_{i_1}, \bar{\mathbf{Y}}_{t-2,i_1}, \tilde{Y}^{(j)}_{t-1,i_1} \right.\right)$, which, as function of $\mathbf{X}_i$ and $\bar{\mathbf{Y}}_{t-1,i}$, is estimated using subjects who have Y_{t-1} observed, as previously. This deterministic imputation is analogous to the way in which quadratic functions of covariates, say, are dealt with in ordinary multiple imputation. If X and X^2 are both covariates in the analysis model, multiple imputations $X^{(j)}_i$ of any missing X_i are obtained in the ordinary way, but then the imputed value of X^2_i can be simply $\left\{X^{(j)}_i\right\}^2$, the square of the imputation.

Similarly, for subject i_1 at time $t+1$, our proposed method works by first imputing a value for $\hat{\pi}_{t+1,i_1}$, based on $\hat{\pi}^{(j)}_{t,i_1}$, $\hat{\lambda}\left(t+1 \left| \mathbf{X}_{i_1}, \bar{\mathbf{Y}}_{t,i_1} \right.\right)$, $\tilde{Y}^{(j)}_{t-1,i_1}$ and $\tilde{Y}^{(j)}_{t,i_1}$ as follows:

$$\hat{\pi}^{(j)}_{t+1,i_1} = \hat{\pi}^{(j)}_{t,i_1} \hat{\lambda}\left(t+1 \left| \mathbf{X}_{i_1}, \bar{\mathbf{Y}}_{t-2,i_1}, \tilde{Y}^{(j)}_{t-1,i_1}, \tilde{Y}^{(j)}_{t,i_1} \right.\right),$$

and then Y_{t+1,i_1} is imputed using $\mathbf{X}_{i_1}$, $\bar{\mathbf{Y}}_{t-2,i_1}$, $\tilde{Y}^{(j)}_{t-1,i_1}$, $\tilde{Y}^{(j)}_{t,i_1}$ and $\hat{\pi}^{(j)}_{t+1,i_1}$.

Finally, $\hat{\mu}_{\mathrm{RMI}}$ can be calculated as the solution to

$$\sum_{j=1}^{m} \sum_{i=1}^{n} \left\{\tilde{Y}^{(j)}_{T,i} - \mu_{\mathrm{RMI}}\right\} = 0. \tag{6}$$

Theorem 3.2 (Multiply imputed DR monotone longitudinal estimator)**.** *The estimator $\hat{\mu}_{\mathrm{RMI}}$ is doubly robust. That is, if at least one of the two models (the POM model and the POD model) is correctly specified (but not necessarily both), $\hat{\mu}_{\mathrm{RMI}}$ is a consistent estimator of μ.*

A sketch proof is given in Appendix D.

A variance estimator analogous to (5) can be obtained using Rubin's variance formula. The same caveats that this variance estimator does not acknowledge the uncertainty due to the fact that the weights have been estimated, and (even ignoring this problem) is only singly robust, applies equally here as in the univariate case.

Relationship to other estimators The OR estimator is now the solution to

$$\sum_{i=1}^{n} \left\{H_0\left(\mathbf{X}^T_i, \hat{\boldsymbol{\beta}}_0\right) - \mu_{\mathrm{OR}}\right\} = 0,$$

where $H_0\left(\mathbf{X}^T_i, \hat{\boldsymbol{\beta}}_0\right)$ is as defined by Bang and Robins [2005]. Briefly, their sequential regression estimator is built as follows. They write $H_T = Y_T$ for those who have Y_T observed. Then, for

$t = T-1, \ldots, 0$, $H_t(\mathbf{X}, Y_1, \ldots, Y_t) = E(H_{t+1} | \mathbf{X}, Y_1, \ldots, Y_t)$ is defined for everyone for whom Y_t is observed. They estimate each H_t using regression models which together constitute the POD model and show that, if this model is correct and the MAR assumption holds, then $E\{H_0(\mathbf{X})\} = E(Y_T) = \mu$, leading to the estimating equation above.

The DR estimator proposed by Bang and Robins [2005] is the same as the OR estimator but with the POD model replaced with the extended POD model.

3.2.2 Non-monotone longitudinal data

For non-monotone missingness patterns, we recommend first testing the hypothesis that the missing data mechanism belongs to the randomised monotone missingness (RMM) sub-class described in Section 2.4.3 using the test described by Robins and Gill [1997]. If the data do not support this hypothesis, then MAR should be rejected as implausible; even in this case, however, an analysis which assumes ignorability might be required as a point of departure for subsequent sensitivity analyses.

Under the assumption that the data are RMM, the parameters shown in Fig. 1 (omitting the grey lines) can be estimated as described in 2.4.3.

From these estimated probabilities, we would like to estimate each of

$$P(R_{1,i} = 1 | \mathbf{X}_i) = p_1(\mathbf{X}_i), \tag{7}$$

$$P(R_{2,i} = 1 | \mathbf{X}_i, Y_{1,i}) = p_1(\mathbf{X}_i)\, p_2(\mathbf{X}_i, Y_{1,i}) + p_2(\mathbf{X}_i) \text{ and} \tag{8}$$

$$P(R_{3,i} = 1 | \mathbf{X}_i, Y_{1,i}, Y_{2,i}) = p_1(\mathbf{X}_i)\, p_2(\mathbf{X}_i, Y_{1,i})\, p_3(\mathbf{X}_i, Y_{1,i}, Y_{2,i}) + p_1(\mathbf{X}_i)\, p_3(\mathbf{X}_i, Y_{1,i}) + p_2(\mathbf{X})\, p_3(\mathbf{X}, Y_{2,i}) + p_3(\mathbf{X}_i). \tag{9}$$

Note that even in this non-monotone setting, since the data are longitudinal, it remains the case that $\hat{\pi}_{t,i} = P(R_{t,i} = 1 | \mathbf{Z}_i) = P\left(R_{t,i} = 1 \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}\right.\right)$, i.e. that the missingness probabilities at each timepoint depend only on past measurements of Y.

There is no problem with (7) but (8) and (9) are undefined for some subjects. For example, if subject i has only Y_2 observed then $p_2(\mathbf{X}_i, Y_{1,i})$ cannot be calculated. Up to a function of the unknown $Y_{1,i}$, it can, however, be specified and in such cases (8) and (9) are specified as known functions of the unknown $Y_{1,i}$ or $Y_{2,i}$. This completes the description of the POM model.

We proceed by fitting the model using MI, and to cope with the non-monotone pattern, MI using chained equations (MICE) as described in Section 2.7 is used. As with the monotone case, for each $t \in \{1, \ldots, T\}$, $\hat{\pi}_{t,i}^{-1}$ is included as an additional covariate (additional to the specified POD model) when imputing $Y_{t,i}$. As we noted above, $\hat{\pi}_{t,i}^{-1}$ itself, in general, is missing for some subjects, and is therefore imputed (deterministically) as a function of the (possibly imputed) $\tilde{Y}_{1,i}, \ldots, \tilde{Y}_{t-1,i}$.

Although when generating such data, we would only need to consider the distribution of each outcome variable Y_t conditional on the covariates and the previous $t-1$ outcome variables (since the future cannot determine the past), for the analysis model (the POD model), it will be necessary—in this non-monotone case—to postulate the implied models for Y_t given all future outcome variables as well, and the future outcome variables must be included in the imputation models, e.g. Y_2 must be included in the imputation model for Y_1. Thus, the extended POD model in the non-monotone case differs from that of the monotone case, since the imputation model for Y_t conditions on all past and future values of Y, as well as $\mathbf{X}$ and $\hat{\pi}_t^{-1}$.

Finally, $\hat{\mu}_{\text{RMI}}$ is again calculated as the solution to

$$\sum_{j=1}^{m} \sum_{i=1}^{n} \left\{ \tilde{Y}_{T,i}^{(j)} - \mu_{\text{RMI}} \right\} = 0 \tag{10}$$

and a variance estimate (subject to the same caveats as above) obtained using Rubin's variance formula.

Conjecture 3.1 (Multiply imputed DR non-monotone longitudinal estimator)**.** *When the full data are multivariate normal, the estimator $\hat{\mu}_{\text{RMI}}$ is doubly robust. That is, if at least one of the*

two models (the POM model and the POD model) is correctly specified (but not necessarily both), $\hat{\mu}_{\text{RMI}}$ *is a consistent estimator of* μ*.*

We note that, given that our proposed procedure relies on multiple imputation using chained equations, which itself has not been theoretically justified outside of the multivariate normal setting, our estimator inherits the theoretical limitations discussed in Section 2.7. However, MICE has been extensively and very successfully used in practice, supported by simulation studies, outside the multivariate normal setting, and similar simulation studies (not reported in this article) suggest that our estimator behaves similarly well for a mixture of continuous, binary and categorical variables.

3.3 Non-monotone non-longitudinal MAR missing data

The arguments above apply equally when the data are not constrained to be longitudinal, assuming that the weights can be estimated. Although Robins and Gill [1997] propose a method for calculating the complete case weights in the RMM setting using an EM algorithm with the path followed by a particular subject through Figure 1 treated as a missing value, the individual path probabilities are not estimated. Thus the variable-specific missingness probabilities cannot be estimated using the methods described by Robins and Gill [1997]. However, it is possible that such an EM procedure could be extended to estimate the individual path probabilities as well. Even if this were not the case, a particular path could be assumed for each subject, and the path probabilities computed on the basis of this assumption. Sensitivity analyses could then be conducted to assess the potential impact of varying the assumed path followed by each subject. Using either or these approaches, if it were possible to estimate the necessary weights, then RMI with chained equations could be used exactly as described for the longitudinal setting above.

Note that the settings discussed thus far, but with missingness affecting the covariates $\mathbf{X}$ rather than, or in addition to, the outcome(s) $\mathbf{Y}$, in general, fit under this heading. Thus, the problem of missing covariates could also potentially be addressed using robust MI if a method for estimating the weights in this setting were available.

4 Simulation studies

4.1 Univariate MAR missing data

First we repeat the first simulation study carried out by Bang and Robins [2005], adding our RMI estimator as a fourth estimator to be compared with the IPW estimator, the OR estimator and the DR estimator.

In this simulation study, $\mathbf{X} = (X_1, X_2, X_3)$ is always fully-observed and generated from a trivariate normal distribution with mean $(0, 0, 0)$ and variance-covariance matrix equal to the (3×3) identity matrix. Y is normally distributed with mean $s_{\text{true}}(\mathbf{X}, \boldsymbol{\beta})$ and unit variance, where $s_{\text{true}}(\mathbf{X}, \boldsymbol{\beta})$ is defined in part 1. of Table 1.

The POM model used to generate R is also described in part 1. of Table 1.

To investigate the double robustness property, an incorrect POM model and an incorrect POD model are also specified as defined in Table 1.

Stata code for implementing all the simulation studies discussed in this section is available as supplementary material in the electronic version of this paper.

4.2 Longitudinal monotone MAR missing data

Next, we repeat the longitudinal monotone simulation study carried out by Bang and Robins [2005], again adding our RMI estimator as a fourth estimator to be compared with the IPW estimator, the OR estimator and the DR estimator.

As before, $\mathbf{X} = (X_1, X_2, X_3)$ is always fully-observed with X_1, X_2, X_3 independent and identically distributed standard normal variables. Y_1 is normally distributed with mean $\tilde{s}_1^{\text{true}}(\mathbf{X}, \boldsymbol{\beta}_1)$

and unit variance, Y_2, conditional on Y_1, is normally distributed with mean $s_2^{\text{true}}(\mathbf{X}, Y_1, \boldsymbol{\beta}_2)$ and unit variance. Details are given in Table 1.

One feature not mentioned in Bang and Robins [2005] is that further calculation is needed to establish the implied form of the distribution of $Y_2 | \mathbf{X}$, which—using their notation—has mean $s_1^{\text{true}}(\mathbf{X}, \boldsymbol{\beta}_1)$. The conditional distribution of $Y_1 | \mathbf{X}$ is

$$N(3X_1 - 2X_1X_3, 1)$$

and the conditional distribution of $Y_2 | \mathbf{X}, Y_1$ is

$$N\left(-3X_1^2 + 3X_2 + Y_1^2 - 2X_2Y_1, 1\right).$$

The conditional expectation of $Y_2 | \mathbf{X}$ is therefore

$$\begin{aligned} &-3X_1^2 + 3X_2 + 1 + (3X_1 - 2X_1X_3)^2 - 2X_2(3X_1 - 2X_1X_3) \\ &= 1 + 3X_2 + 6X_1^2 - 6X_1X_2 - 12X_1^2X_3 + 4X_1X_2X_3 + 4X_1^2X_3^2. \end{aligned}$$

Thus, when carrying out the simulation study under the 'both models correct' scenario, the authors used $1, X_2, X_1^2, X_1X_2, X_1^2X_3, X_1X_2X_3, X_1^2X_3^2$ as the covariates for the second linear regression stage, as opposed to $1, X_1, X_1X_3$ as their paper suggests.

The implied $s_1^{\text{true}}(\mathbf{X}, \boldsymbol{\beta}_1)$ is given in Table 1, along with details of the correct POM model and the incorrect POM and POD models.

4.3 Longitudinal non-monotone MAR missing data

Next, we consider a longitudinal non-monotone simulation study. In this case, neither the OR nor the DR estimator can be used and thus we compare our RMI estimator with the IPW estimator and an ordinary multiple imputation (MI) estimator, *i.e.* an estimator identical to the RMI estimator but without the inverse probability weights as additional covariates in the imputation model.

In this simulation study, X (univariate) is always observed and generated from a standard normal distribution. Y_1 and Y_2 are normally distributed with means $\tilde{s}_1^{\text{true}}(X, \boldsymbol{\beta}_1)$ and $s_2^{\text{true}}(X, Y_1, \boldsymbol{\beta}_2)$, respectively (see Table 1), and unit variance. The implied $s_1^{\text{true}}(X, Y_2, \boldsymbol{\beta}_1)$ is also given in the table.

Note that $s_1(\cdot)$ is now a function of Y_2. This is essential, since some subjects have Y_2 but not Y_1 observed. If Y_2 is omitted from the imputation model for Y_1, the resulting estimator is, in general, biased since the stationary distribution to which the Gibbs sampler in the MICE procedure converges is not the correct full-data distribution, even under MAR.

The POM model is defined by the multinomial logit model described in Table 1, where the incorrect POM and POD models are also described.

4.4 Non-longitudinal non-monotone MAR missing data

Finally, we consider a non-longitudinal non-monotone simulation study. Again, neither the OR nor the DR estimator can be used and thus we compare our RMI estimator with the IPW estimator and an ordinary MI estimator.

As in the previous simulation study, X (univariate) is always observed and generated from a standard normal distribution. Y_1 and Y_2 are normally distributed with means $\tilde{s}_1^{\text{true}}(X, \boldsymbol{\beta}_1)$ and $s_2^{\text{true}}(X, Y_1, \boldsymbol{\beta}_2)$, respectively (see Table 1), and unit variance.

The implied $s_1^{\text{true}}(X, Y_2, \boldsymbol{\beta}_1)$, the correct POM model and the incorrect POD model are given in part 4. of Table 1.

Because of the difficulty associated with estimating the marginal weights (discussed in Section 3.3), we cannot obtain reliable estimates of $\hat{\pi}_1^{\text{true}}(X, \boldsymbol{\alpha}_{11}, \boldsymbol{\alpha}_{12}, \boldsymbol{\alpha}_{21})$ and $\hat{\pi}_2^{\text{true}}(X, Y_1\boldsymbol{\alpha}_{11}, \boldsymbol{\alpha}_{12}, \boldsymbol{\alpha}_{22})$ even for the complete cases. For the purposes of this simulation study, therefore, we will use the true (known) weights.

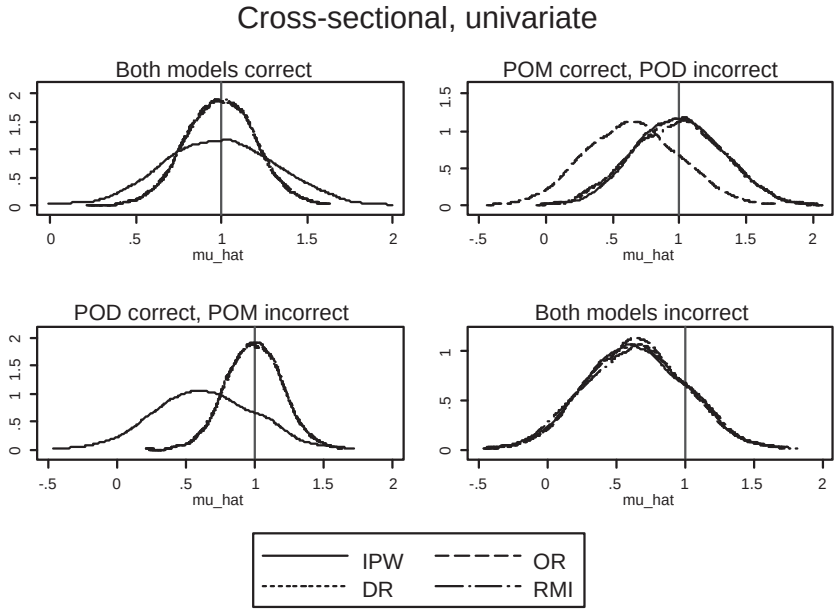

Figure 2: Kernel density plots showing the simulated values of $\hat{\mu}$ for the cross-sectional, univariate simulation study.

Since the true weights are being used, no 'POM model' exists. To investigate the double robustness property, we therefore define $\hat{\pi}_1^{\text{false}} = \sqrt{\hat{\pi}_1^{\text{true}}}$ and $\hat{\pi}_2^{\text{false}} = \sqrt{\hat{\pi}_2^{\text{true}}}$. There is no strong motivation for choosing this relationship between the correct and incorrect weights, except that it produced an appreciable, yet not too extreme, bias in the IPW estimator.

The simulation studies are all based on a sample size of 500 and 1,000 simulations, with the MI and robust MI procedures based on 10 imputations and 10 cycles of the chained equations procedure. The results are shown in Table 2, and a visual summaries using kernel density plots are given in Figures 2–5.

4.5 Conclusions

We see from parts 1. and 2. of Table 2, together with Figures 2 and 3, that in both the univariate cross-sectional and longitudinal monotone cases, where the Bang and Robins [2005] method can be applied, its performance and our estimator's performance are very similar. In addition, the variance estimates obtained using Rubin's variance formula perform well when both models are correctly specified, although, as expected, they do not share the double robustness property possessed by the estimators themselves. Even though our proposed variance estimator does not take into account the variability of the estimated weights, at least in our simulations, this effect is negligible.

When the missing data are longitudinal but non-monotone, the Bang and Robins [2005] method can no longer be used, but our estimator works very well: it appears to exhibit the desired double robustness property as well as improved efficiency compared with IPW (see part 3. of Table 2 and Figure 4). The loss of efficiency relative to OR and MI is negligible. We also see that RMI works well (see part 4. of Table 2 and Figure 5) for non-longitudinal non-monotone data, when the true weights are used.

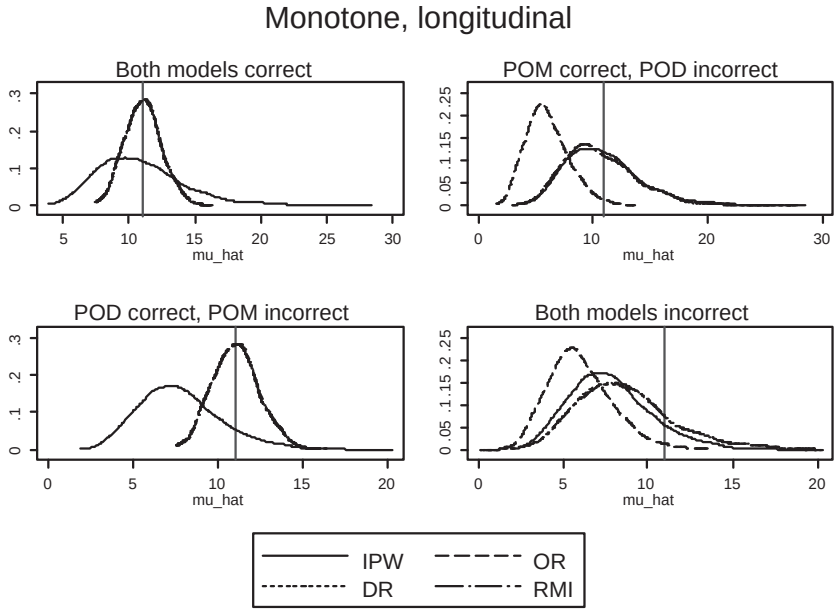

Figure 3: Kernel density plots showing the simulated values of $\hat{\mu}_2$ for the monotone, longitudinal simulation study.

5 The RECORD study

We now present a real data example to which our new method is applied.

The glycaemic data from the RECORD trial involves 1122 subjects, all of whom were taking either Metformin (Met) or Sulfonylurea (Su) prior to the start of the trial. The Met and Su arms were subsequently treated as two separate strata, with patients in the Met arm randomised to receive either additional Su or additional Rosiglitazone (Rosi), and patients in the Su arm randomised to receive either additional Met or additional Rosi. HbA_{1c} (a measure of long-term glucose control) was collected on patients at baseline, and at 8 further follow-up visits: at 2, 4, 6, 8, 10, 12, 15 and 18 months. 167 (14.9%) patients were lost to follow-up, and there were a further 161 intermittent (non-monotone) missing observations.

The aim was to investigate whether or not Rosi in combination with Met or Su is as good as Met+Su for achieving glycaemic control. A non-inferiority criterion (upper band 95% CI of difference) at 18 months was set at 0.4%.

The original analysis carried out by Home et al. [2007] assumed multivariate normality for the repeated HbA_{1c} measurements conditional on baseline HbA_{1c}, but there was concern that the residuals from this analysis exhibit some right skewness. We therefore use RMI to assess the sensitivity of the conclusions of Home et al. [2007] to the normality assumption. By using a non-parametric FD model, and confining normality assumptions to the POD model, our results will be robust to non-normality, as long as the assumptions of the POM model and the MAR assumption hold.

The missingness model corresponding to Figure 1 is appreciably more complex with 8 variables rather than 3. It is clear that some reduction in the dimensionality of the problem must be made if the weights are to be estimated efficiently. We will impose the restriction that, conditional on the most recently observed outcome, whether or not the next outcome is observed is independent of all other observed outcomes. Apart from this, the method is identical to the one described in the simulation study in Section 4.3. Such a restriction on the POM model is a potential source of bias in the robust MI analysis. However, when the next most recently-observed outcome was included in the POM model at each time-point, the multinomial logistic regression models did not converge, and thus the potential impact of this assumption could not be assessed.

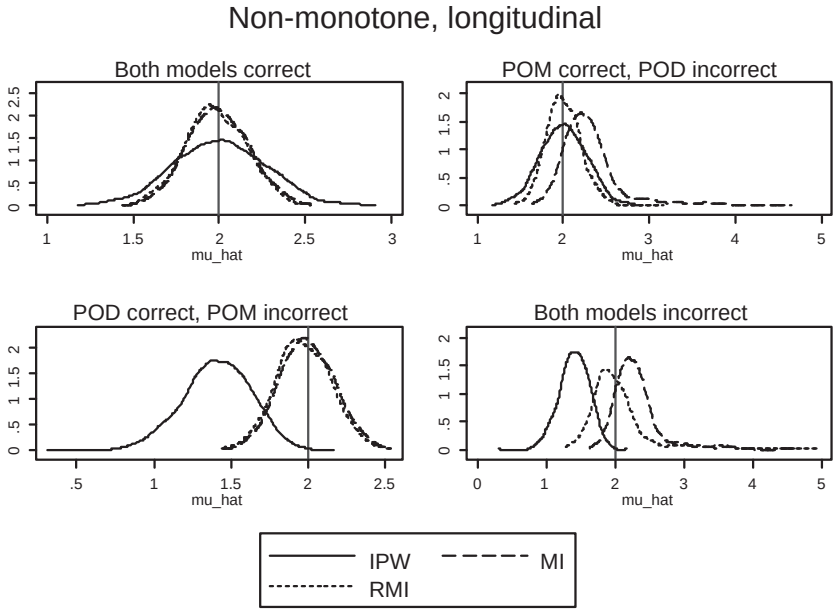

Figure 4: Kernel density plots showing the simulated values of $\hat{\mu}_2$ for the non-monotone, longitudinal simulation study.

The results are as follows. The direct likelihood analysis estimates a difference of 0.087% between the Met+Rosi and Met+Su arms in the change in HbA_{1c} from baseline to 18 months, with a standard error of 0.08%. The corresponding estimate from the RMI approach is 0.017% (SE 0.09%). The direct likelihood analysis estimates a difference of 0.066% between the Su+Rosi and Su+Met arms in the change in HbA_{1c} from baseline to 18 months, with a standard error of 0.08%. The corresponding estimate from the RMI approach is 0.033% (SE 0.07%). We see that the results from RMI are similar (but not identical) to those from the direct likelihood analysis. Certainly as regards the pre-specified non-inferiority criterion of 0.4%, neither method supports the rejection of non-inferiority.

Figure 6 shows the differences between these profiles for the two arms separately. Again we see that the profiles are similar but not identical. The differences are substantively very small and unlikely to be important in practice. If anything, the RMI approach suggests a lower HbA_{1c} for the Rosi groups compared with the corresponding estimates from the direct likelihood method, whereas the estimates for the standard groups show less of a difference between the two methods. As a result, Rosi compared with standard looks to be slightly better under the RMI analysis suggesting that the direct likelihood analysis is (in this particular case) slightly conservative in the sense that it is more likely to conclude that Rosi is inferior.

The reason for there being only a small difference between the two approaches is probably that the non-normality is not severe. We notice that what little difference there is increases over time. This is likely to be due to the increased dependency on modelling assumptions in the direct likelihood approach as the number of missing observations increases.

6 Discussion

By combining the regression-based doubly robust estimator of Bang and Robins with multiple imputation, we have shown how MI estimators with improved robustness can be constructed in settings (such as the non-monotone longitudinal pattern found in the RECORD study) where, hitherto, DR estimators have not been implemented.

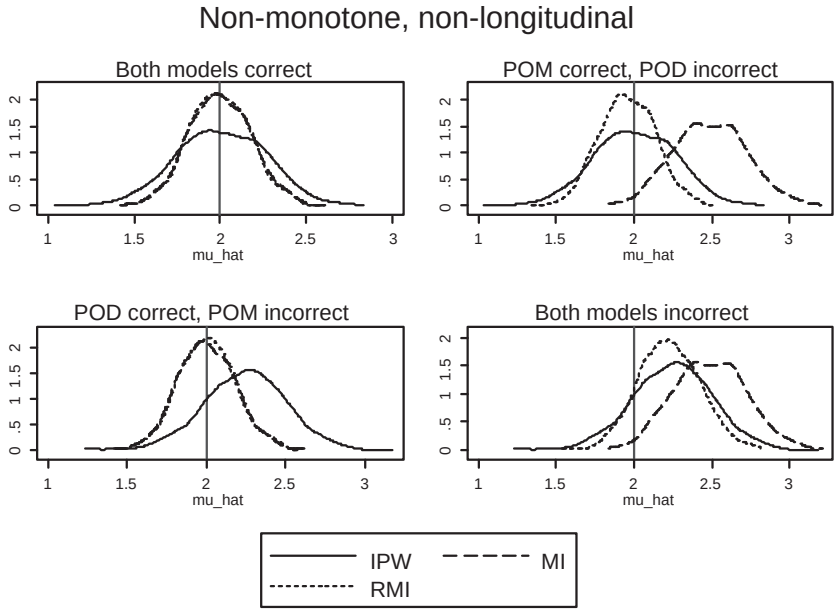

Figure 5: Kernel density plots showing the simulated values of $\hat{\mu}_2$ for the non-monotone, non-longitudinal simulation study.

We have seen from simulation studies that when the Bang and Robins [2005] method can be applied, its performance and our estimator's performance are very similar. When the missing data are longitudinal but non-monotone, the Bang and Robins [2005] method can no longer be used, but our estimator works very well: it appears to exhibit the desired double robustness property as well as improved efficiency compared with IPW. Furthermore, our method is easily implemented in standard software packages such as `ice` in Stata. Although the simulations presented here are for continuous variables, similar results were found for incomplete binary data.

We have also shown that RMI could in principle be applied to general (non-longitudinal) non-monotone data. However, the problem of estimating the variable-specific inverse probability weights needs first to be resolved. Unfortunately, the method proposed by Robins and Gill [1997] for estimating the complete case weights can not be used to identify the variable-specific weights. We have shown, by substituting the known true weights, that if a method were developed for estimating these probabilities, RMI could be used and would perform well.

Although our focus has been on examples where the aim is to estimate the marginal mean of one of the variables, RMI can be used much more generally (for example to estimate the parameters of a regression of one variable on another) and as easily to obtain estimators with improved robustness whenever ordinary MI is appropriate.

A criticism of DR methods has been their potential instability when the weights are too variable and their sub-optimal performance when both the POD and POM models are misspecified [Kang and Schafer, 2007]. These limitations apply equally to the RMI procedure proposed in this article. Recent work by Tan [2010], Tsiatis et al. [2011], Vansteelandt et al. [2011] proposes improvements to the standard DR estimators with respect to these issues, and the adoption of some of these strategies within RMI represents an exciting direction for future work.

Acknowledgements

The authors are grateful to GlaxoSmithKline and Drs. Adam Crisp and Paula Curtis for their permission to use the RECORD study data in this article.

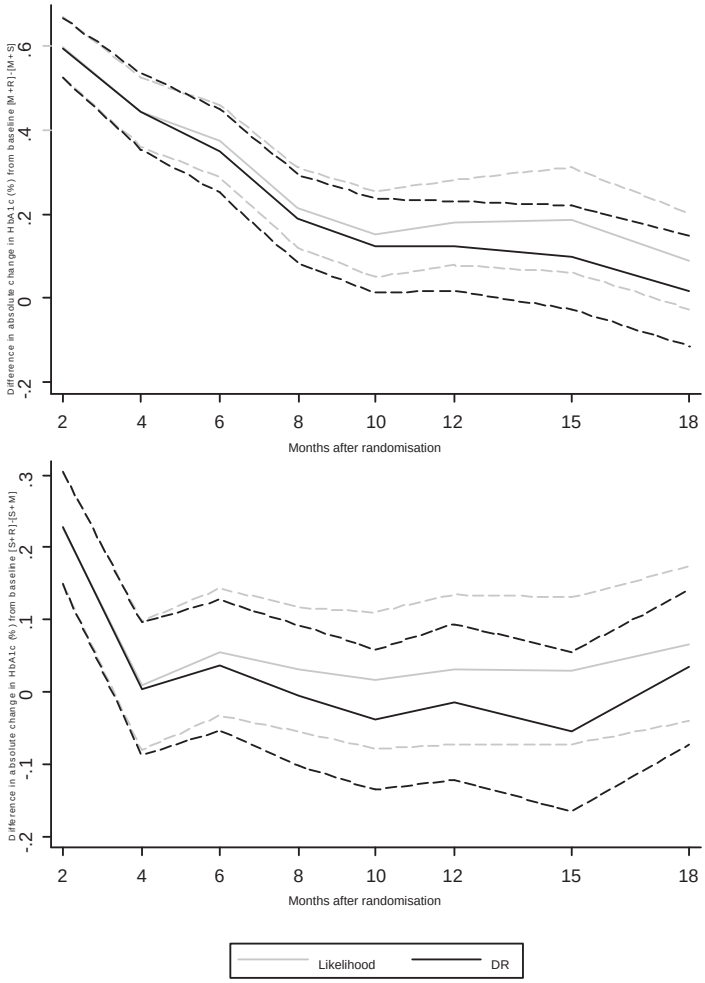

Figure 6: The differences between the HbA_{1c} profiles for the Met+Rosi and Met+Su arms (above) and for the Su+Met and Su+Rosi arms (below). The solid lines show the predicted differences, and the dotted lines show ± the pointwise standard errors for these differences.

References

Bang, H., Robins, J. M., 2005. Doubly robust estimation in missing data and causal inference models. Biometrics 61, 962–973.

Carpenter, J. C., Kenward, M. G., Vansteelandt, S., 2006. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. Journal of the Royal Statistical Society. Series A (Statistics in Society) 169, 571–584.

Cox, D. R., Hinkley, D. V., 1974. Theoretical Statistics. Chapman and Hall, London.

Creemers, A., Aerts, M., Hens, N., Molenberghs, G., 2011. A nonparametric approach to weighted estimating equations for regression analysis with missing covariates. Computational Statistics & Data Analysis (in press).

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39 (1), 1–38.

Gelman, A., Raghunathan, T. E., 2001. Using conditional distributions for missing-data imputation, in discussion of 'Conditionally specified distributions' by Arnold et al. Statistical Science 3, 268–269.

Gill, R., Robins, J. M., 1996. Missing at random from an algorithmic viewpoint. Proceedings of the First Seattle Symposium on Survival Analysis.

Home, P. D., Jones, N. P., Pocock, S. J., Beck-Nielsen, H., Gomis, R., Hanefeld, M., Komajda, M., Curtis, P., 2007. Rosiglitazone RECORD study: glucose control outcomes at 18 months. Diabetic Medicine 24, 626–634.

Horvitz, D. G., Thompson, D. J., dec 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47 (260), 663–685.

Kang, J. D. Y., Schafer, J. L., 2007. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). Statistical Science 22 (4), 523–580.

Kenward, M. G., 1998. Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. Statistics in Medicine 17 (23), 2723–2732.

Little, R., Yau, L., dec 1996. Intent-to-treat analysis for longitudinal studies with drop-outs. Biometrics 52 (4), 1324–1333.

Little, R. J. A., Rubin, D. B., 2002. Statistical Analysis with Missing Data. Wiley, New York.

Molenberghs, G., Beunckens, C., Sotto, C., Kenward, M. G., 2008. Every missingness not at random model has a missingness at random counterpart with equal fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (2), 371–388.

Molenberghs, G., Goetghebeur, E. J. T., Lipsitz, S. R., Kenward, M. G., 1999. Non-random missingness in categorical data: strengths and limitations. American Statistician 53, 110–118.

Molenberghs, G., Michiels, B., Kenward, M. G., Diggle, P. J., 1998. Monotone missing data and pattern-mixture models. Statistica Neerlandica 52, 153–161.

Raghunathan, T., Solenberger, P., Hoewyk, J., 2007. *IVEware: Imputation and Variance Estimation Software*. Available at `http://www.isr.umich.edu/src/smp/ive/.`

Robins, J. M., 1999. Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran and D. Berry (eds), Statistical Models in Epidemiology: The Environment and Clinical Trials, IMA **116**, 95–134. New York: Springer-Verlag.

Robins, J. M., Gill, R. D., 1997. Non-response models for the analysis of non-monotone ignorable missing data. Statistics in Medicine 16 (1), 39–56.

Robins, J. M., Mark, S. D., Newey, W. K., 1992. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. Biometrics 48, 479–495.

Robins, J. M., Rotnitzky, A., Scharfstein, D. O., 1999. Sensitivity Analysis for Selsction Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. Taken from *Statistical Models in Epidemiology: The Environment and Clinical Trials*. Vol. 116. Springer-Verlag, New York.

Robins, J. M., Rotnitzky, A., Zhao, L. P., sep 1994. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89 (427), 846–866.

Robins, J. M., Rotnitzky, A., Zhao, L. P., mar 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 90 (429), 106–121.

Royston, P., 2004. Multiple imputation of missing values. The Stata Journal 4, 227–241.

Rubin, D. B., dec 1976. Inference and missing data. Biometrika 63 (3), 581–592.

Rubin, D. B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

Scharfstein, D. O., Rotnitzky, A., Robins, J. M., dec 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association 94 (448), 1096–1120.

Tan, Z., 2010. Bounded, efficient and doubly robust estimation with inverse weighting. Biometrics 97, 661–682.

Tsiatis, A. A., 2006. Semiparametric Theory and Missing Data. Springer, New York.

Tsiatis, A. A., Davidian, M., Cao, W., 2011. Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. BiometricsIn press.

van Buuren, S., Boshuizen, H. C., Knook, D. L., 1999. Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine 18, 681–694.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., Rubin, D. B., 2006. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation 76, 1049–1064.

van Buuren, S., Oudshoorn, C. G. M., 2000. *Multiple Imputation by Chained Equations: MICE V1.0 User's manual*. Leiden: TNO Preventie en Gezondheid.

Vansteelandt, S., Bekaert, M., Claeskens, G., 2011. On model selection and model misspecification in causal inference. Statistical Methods in Medical ResearchIn press.

Vansteelandt, S., Rotnitzky, A., Robins, J., 2007. Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. Biometrika 94, 841–860.

APPENDIX

A Further details on estimating π_i in MRMM processes for longitudinal data

We consider the example discussed in Section 2.4.3 where there are three outcome variables, but the argument easily extends to any number of outcome variables.

We start by defining a 'stage 2' variable, $S_{2,i}$ taking the value $s_{2,i}$ where

$$s_{2,i} = \inf \{1, 2, 3 : Y_{s_2,i} \text{ is observed}\}$$

or the value 0 if none of $\{Y_{1,i}, Y_{2,i}, Y_{3,i}\}$ is observed. A multinomial logit model, for example, could be fitted to $S_{2,i}$, conditional on the covariates, and the probabilities $p_1(\mathbf{X}_i)$, $p_2(\mathbf{X}_i)$, and $p_3(\mathbf{X}_i)$ estimated. Then, a 'stage 3' variable, $S_{3,i}$, is defined to take the value $s_{3,i} - 1$ where

$$s_{3,i} = \inf \{2, 3 : Y_{s_3,i} \text{ and } Y_{k,i} \text{ are observed, where } k < s_3\}$$

or the value 1 if only one of $\{Y_{1,i}, Y_{2,i}, Y_{3,i}\}$ is observed. For each level $s_{2,i}$ of $S_{2,i}$, a multinomial logit model can be fitted to $S_{3,i}$ conditional on $Y_{s_{2,i},i}$ and the covariates. The probabilities $p_2(\mathbf{X}_i, Y_{1,i})$, $p_3(\mathbf{X}_i, Y_{1,i})$, and $p_3(\mathbf{X}_i, Y_{2,i})$ are estimated.

Finally, a 'stage 4' variable, $S_{4,i}$, taking the value $s_{4,i}$ where

$$s_{4,i} = \begin{cases} 1 & \text{if } Y_{1,i}, Y_{2,i}, Y_{3,i} \text{ are all observed} \\ 0 & \text{otherwise} \end{cases}$$

is defined and, for each pair $\{s_{2,i}, s_{3,i}\}$, a logistic regression can be fitted to $S_{4,i}$ conditional on $Y_{s_{2,i},i}$, $Y_{s_{3,i},i}$ and the covariates. The probabilities $p_3(\mathbf{X}_i, Y_{1,i}, Y_{2,i})$ are estimated.

Then,

$$\hat{\pi}_i = \hat{p}_1(\mathbf{X}_i)\,\hat{p}_2(\mathbf{X}_i, Y_{1,i})\,\hat{p}_3(\mathbf{X}_i, Y_{1,i}, Y_{2,i})$$

B An example of proper multiple imputation

Suppose that Y is continuous, and that the POD model is a linear regression. Let $\hat{V}_{(\hat{\boldsymbol{\beta}},\hat{\phi})}$ be the estimated variance-covariance matrix of $\left(\hat{\boldsymbol{\beta}}, \hat{\phi}\right)$ and $\hat{V}_{Y|\mathbf{X},\hat{\pi}^{-1}}$ be the estimator from the extended POD model of

$$\mathrm{Var}\left\{Y_i \left| \mathbf{X}_i, \hat{\pi}^{-1}(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}), R_i = 1\right.\right\}$$

We draw m times from the large-sample approximation to the posterior distribution of $\left(\hat{\boldsymbol{\beta}}, \hat{\phi}\right)$:

$$\left\{\boldsymbol{\beta}^{(j)}, \phi^{(j)}\right\} \overset{\text{i.i.d.}}{\sim} N\left\{\left(\hat{\boldsymbol{\beta}}, \hat{\phi}\right), \hat{V}_{(\hat{\boldsymbol{\beta}},\hat{\phi})}\right\}, \quad j = 1, \ldots, m$$

and m times from the large-sample approximation to the posterior distribution of $\hat{V}_{Y|\mathbf{X},\hat{\pi}^{-1}}$:

$$V^{(j)}_{Y|\mathbf{X},\hat{\pi}^{-1}} \overset{\text{i.i.d.}}{\sim} \hat{V}_{Y|\mathbf{X},\hat{\pi}^{-1}} \frac{1}{n_c - p} \chi^{2\,-1}_{n_c-p}, \quad j = 1, \ldots, m$$

where $n_c = \sum_{i=1}^{n} R_i$ and p is the number of parameters estimated in the extended POD model.

m imputed datasets are then generated with $\tilde{Y}_i^{(j)}$ replacing Y in the j^{th} dataset where

$$\tilde{Y}_i^{(j)} = R_i Y_i + (1 - R_i)\left[\hat{e}\left\{\mathbf{X}_i^T, \boldsymbol{\beta}^{(j)}, \phi^{(j)}\right\} + \varepsilon_i^{(j)}\right]$$

and $\varepsilon_i^{(j)} \overset{\text{i.i.d.}}{\sim} N\left\{0, V^{(j)}_{Y|\mathbf{X},\hat{\pi}^{-1}}\right\}$.

When the POD model is not a linear regression model, the imputations are drawn *properly* according to the appropriate imputation distribution.

C Sketch proof of Theorem 3.1

Sketch proof. The consistency of $\hat{\mu}_{\text{RMI}}$ when the POD model is correctly specified follows from the fact that the true value of ϕ is zero. If the POM model is correctly specified, but not the POD model, it is slightly less evident that $\hat{\mu}_{\text{RMI}}$ remains consistent.

We continue to write $\hat{\pi}_i$ for $\hat{\pi}(\mathbf{X}_i, \hat{\boldsymbol{\alpha}})$. The RMI estimating equation

$$\sum_{j=1}^{m}\sum_{i=1}^{n}\left[R_i Y_i + (1 - R_i)\tilde{Y}_i^{(j)} - \mu_{\text{RMI}}\right] = 0$$

can be rewritten as

$$\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{R_i\left[Y_i - \hat{e}\left(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\pi}_i^{-1}\right)\right] + R_i\hat{\pi}_i^{-1}\left[Y_i - \hat{e}\left(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\pi}_i^{-1}\right)\right]\right.$$
$$\left. + (1 - R_i)\left[\tilde{Y}_i^{(j)} - \hat{e}\left(\mathbf{X}_i^T, \boldsymbol{\beta}, \hat{\phi}, \hat{\pi}_i^{-1}\right)\right] + \hat{e}\left(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\pi}_i^{-1}\right) - \mu_{\text{RMI}}\right\} = 0 \quad (11)$$

This follows from the fact that $\sum_{j=1}^{m}\sum_{i=1}^{n} R_i\hat{\pi}_i^{-1}\left(Y_i - e_i\right)$ is numerically zero since we included $\hat{\pi}_i^{-1}$ in our extended POD model GLM.

$\sum_{j=1}^{m}\sum_{i=1}^{n} R_i\left[Y_i - \hat{e}\left(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\pi}_i^{-1}\right)\right]$ is also numerically zero, assuming that a constant term is included in our GLM. Furthermore, $(1-R_i)\left[\tilde{Y}_i^{(j)} - \hat{e}\left(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\pi}_i^{-1}\right)\right]$ has zero expectation, since the proper imputations have been drawn from a posterior predictive distribution with mean $\hat{e}\left(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\pi}_i^{-1}\right)$ (see Appendix B). Thus we can rewrite (11) as

$$\sum_{j=1}^{m}\sum_{i=1}^{n}\Big\{R_i\hat{\pi}_i^{-1}\left(Y_i - \mu_{\mathrm{RMI}}\right) + \left(1 - R_i\hat{\pi}_i^{-1}\right)\left[\hat{e}\left(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\pi}_i^{-1}\right) - \mu_{\mathrm{RMI}}\right] + (1-R_i)\left[\tilde{Y}_i^{(j)} - \hat{e}\left(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\pi}_i^{-1}\right)\right]\Big\} = 0$$

which we immediately recognise as being of the same form as (1) with the added term $(1-R_i)\left[\tilde{Y}_i^{(j)} - \hat{e}\left(\mathbf{X}_i^T, \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\pi}_i^{-1}\right)\right]$ which has zero expectation, even when the POD model is incorrect. Thus, $\hat{\mu}_{\mathrm{RMI}}$ is consistent whenever the POM model is correctly specified. □

D Sketch proof of Theorem 3.2

Let $H_t\left(\mathbf{X}_i^T, \bar{\mathbf{Y}}_{t,i}, \hat{\bar{\pi}}_{t,i}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}\right)$ be the predictions from the Bang and Robins procedure for longitudinal monotone data (as described in Section 4.2) after $T-t$ iterations. Let $\hat{E}\left(Y_{T,i}\left|\mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}, \hat{\bar{\pi}}_{t,i}\right.\right)$ be the mean of the distribution from which the RMI imputations for $Y_{T,i}$, for a subject who drops out after time t, are drawn.

Lemma D.1.

$$E\left\{H_t\left(\mathbf{X}_i^T, \bar{\mathbf{Y}}_{t,i}, \hat{\bar{\pi}}_{t,i}^{-1}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}\right)\right\} = E\left\{\hat{E}\left(Y_{T,i}\left|\mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}, \hat{\bar{\pi}}_{t,i}\right.\right)\right\}$$

where expectations are taken with respect to the true distribution of $\left(\mathbf{X}_i, \bar{\mathbf{Y}}_{T,i}\right)$.

Sketch proof of Lemma D.1. That Lemma D.1 is true is immediate if the POD model is correct, since both $H_t\left(\mathbf{X}_i^T, \bar{\mathbf{Y}}_{t,i}, \hat{\bar{\pi}}_{t,i}^{-1}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}\right)$ and $\hat{E}\left(Y_{T,i}\left|\mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}, \hat{\bar{\pi}}_{t,i}\right.\right)$ are consistent estimators of $E\left(Y_{T,i}\left|\mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}\right.\right)$. However, the argument [see Tsiatis, 2006, ch. 14] showing that multiple imputation recovers the full-data distribution when the imputation model is correctly specified can also be used to show that when it is incorrectly specified, the incorrect distribution it recovers is equivalent to the hypothetical full-data distribution implied by that incorrectly specified imputation distribution. □

Sketch proof of Theorem 3.2. As for the univariate case, that $\hat{\mu}_{\mathrm{RMI}}$ is consistent when only the POM model is misspecified is intuitively obvious. We therefore concentrate on the consistency of $\hat{\mu}_{\mathrm{RMI}}$ when only the POD model is misspecified.

Assuming that Y_1 is always observed, that $D_i = \{s : R_{1,i} = R_{2,i} = \cdots = R_{s-1,i} = 1, R_{s,i} = R_{s+1,i} = \cdots = R_{T,i} = 0\}$ is the dropout indicator, and that $\bar{\mathbf{Z}}_{t,i}$ denotes the history of $\mathbf{Z}_i$ up to and including t, the general form of the AIPW estimating equation [as described by Tsiatis, 2006, p.208] can be written as

$$\sum_{i=1}^{n}\left\{\frac{I\left(D_i = T+1\right)}{P\left(D_i = T+1\left|\mathbf{X}_i, \bar{\mathbf{Y}}_{T,i}\right.\right)}\left(Y_{T,i} - \mu_{\mathrm{AIPW}}\right) + \sum_{t=1}^{T} I\left(D_i \geq t\right)\left\{I\left(D_i = t\right) - P\left(D_i = t\left|D_i \geq t, \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}\right.\right)\right\} h_t\left(\mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \mu_{\mathrm{AIPW}}\right)\right\} = 0 \tag{12}$$

and the optimal choice of the functions $h_t(\cdot)$ is given by

$$h_t\left(\mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \mu_{\text{AIPW}}\right) = \frac{E\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}\right.\right) - \mu_{\text{AIPW}}}{P\left(R_{t,i} = 1 \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}\right.\right)}$$

This is not shown here but can be found both in Tsiatis [2006] and in Robins [1999]. In our notation, (12) can be rewritten as

$$\sum_{i=1}^{n}\left[\frac{R_{T,i}}{\hat{\pi}_{T,i}}\left(Y_{T,i} - \mu_{\text{AIPW}}\right) + \sum_{t=1}^{T} R_{t-1,i}\left(\frac{\hat{\pi}_{t,i}}{\hat{\pi}_{t-1,i}} - R_{t,i}\right)\frac{E\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}\right.\right) - \mu_{\text{AIPW}}}{\hat{\pi}_{t,i}}\right] = 0 \quad (13)$$

which is equivalent to

$$\sum_{i=1}^{n}\left\{E\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{1,i}\right.\right) - \mu_{\text{AIPW}} + \sum_{t=1}^{T}\frac{R_{t,i}}{\hat{\pi}_{t,i}}\left\{E\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}\right.\right) - E\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}\right.\right)\right\}\right\} = 0 \quad (14)$$

Our estimator (6) can be rewritten as

$$\begin{aligned}
\sum_{j=1}^{m}\sum_{i=1}^{n}\Bigg\{ & R_{T,i}\left(Y_{T,i} - \mu_{\text{RMI}}\right) + R_{T-1,i}\left(1 - R_{T,i}\right)\left\{\hat{E}\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \hat{\bar{\pi}}_{t-1,i}\right.\right) - \mu_{\text{RMI}}\right\} \\
& + \cdots + R_{1,i}\left(1 - R_{2,i}\right)\left\{\hat{E}\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{1,i}\right.\right) - \mu_{\text{RMI}}\right\} + \left(1 - R_{1,i}\right)\left\{\hat{E}\left(Y_{T,i} \left| \mathbf{X}_i\right.\right) - \mu_{\text{RMI}}\right\} \\
& + R_{T-1,i}\left(1 - R_{T,i}\right)\left\{\tilde{Y}_{T,i}^{(j)} - \hat{E}\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \hat{\bar{\pi}}_{t-1,i}\right.\right)\right\} + \cdots \\
& + R_{1,i}\left(1 - R_{2,i}\right)\left\{\tilde{Y}_{T,i}^{(j)} - \hat{E}\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{1,i}\right.\right)\right\}\left(1 - R_{1,i}\right)\left\{\tilde{Y}_{T,i}^{(j)} - \hat{E}\left(Y_{T,i} \left| \mathbf{X}_i\right.\right)\right\}\Bigg\} = 0
\end{aligned}$$

and this is equivalent to

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{m}\Bigg\{ & \hat{E}\left(Y_{T,i} \left| \mathbf{X}_i\right.\right) - \mu_{\text{RMI}} + \sum_{t=1}^{T} R_{t,i}\left\{\hat{E}\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}, \hat{\bar{\pi}}_{t,i}\right.\right) - \hat{E}\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \hat{\bar{\pi}}_{t-1,i}\right.\right)\right\} \\
& + \sum_{t=1}^{T} R_{t-1,i}\left(1 - R_{t,i}\right)\left\{\tilde{Y}_{T,i}^{(j)} - \hat{E}\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \hat{\bar{\pi}}_{t-1,i}\right.\right)\right\}\Bigg\} = 0
\end{aligned}$$

To show that $\hat{\mu}_{\text{RMI}}$ is a doubly-robust estimator of μ, we must show that

$$\begin{aligned}
E\Bigg(\sum_{j=1}^{m}\Bigg\{ & \hat{E}\left(Y_{T,i} \left| \mathbf{X}_i\right.\right) - \mu_{\text{RMI}} + \sum_{t=1}^{T} R_{t,i}\left\{\hat{E}\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}, \hat{\bar{\pi}}_{t,i}\right.\right) - \hat{E}\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \hat{\bar{\pi}}_{t-1,i}\right.\right)\right\} \\
& + \sum_{t=1}^{T} R_{t-1,i}\left(1 - R_{t,i}\right)\left\{\tilde{Y}_{T,i}^{(j)} - \hat{E}\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \hat{\bar{\pi}}_{t-1,i}\right.\right)\right\}\Bigg\}\Bigg) = 0
\end{aligned}$$

when at least one of the POD and POM models is correctly specified, where the outer expectation is with respect to the true distribution of $\mathbf{X}_i, \bar{\mathbf{Y}}_{T,i}$.

The final term is zero (by the definition of $\hat{E}\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}, \hat{\bar{\pi}}_{t,i}\right.\right)$ as the mean of the distribution from which $\tilde{Y}_{T,i}^{(j)}$ is drawn) and thus our requirement becomes that

$$E\left(\sum_{j=1}^{m}\left\{\hat{E}\left(Y_{T,i} \left| \mathbf{X}_i\right.\right) - \mu_{\text{RMI}} + \sum_{t=1}^{T} R_{t,i}\left\{\hat{E}\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t,i}, \hat{\bar{\pi}}_{t,i}\right.\right) - \hat{E}\left(Y_{T,i} \left| \mathbf{X}_i, \bar{\mathbf{Y}}_{t-1,i}, \hat{\bar{\pi}}_{t-1,i}\right.\right)\right\}\right\}\right) = 0$$

when at least one of the POD and POM models is correctly specified, or, equivalently:

$$E\left\{\hat{E}\left(Y_{T,i}\,|\mathbf{X}_i\right)-\mu_{\text{RMI}}+\sum_{t=1}^{T}R_{t,i}\left\{\hat{E}\left(Y_{T,i}\,\middle|\mathbf{X}_i,\bar{\mathbf{Y}}_{t,i},\hat{\bar{\pi}}_{t,i}\right)-\hat{E}\left(Y_{T,i}\,\middle|\mathbf{X}_i,\bar{\mathbf{Y}}_{t-1,i},\hat{\bar{\pi}}_{t-1,i}\right)\right\}\right\}=0$$

By Lemma D.1, this can be rewritten as

$$E\Bigg\{H_0\left(\mathbf{X}_i^T,\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\phi}}\right)-\mu_{\text{RMI}} + \sum_{t=1}^{T}R_{t,i}\left\{H_t\left(\mathbf{X}_i^T,\bar{\mathbf{Y}}_{t,i},\hat{\bar{\pi}}_{t,i},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\phi}}\right)-H_{t-1}\left(\mathbf{X}_i^T,\bar{\mathbf{Y}}_{t-1,i},\hat{\bar{\pi}}_{t-1,i},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\phi}}\right)\right\}\Bigg\}=0 \quad (15)$$

which is the same as

$$E\Bigg\{H_0\left(\mathbf{X}_i^T,\hat{\boldsymbol{\beta}}\right)-\mu_{\text{RMI}} + \sum_{t=1}^{T}\frac{R_{t,i}}{\hat{\bar{\pi}}_{t,i}}\left\{H_t\left(\mathbf{X}_i^T,\bar{\mathbf{Y}}_{t,i},\hat{\bar{\pi}}_{t,i},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\phi}}\right)-H_{t-1}\left(\mathbf{X}_i^T,\bar{\mathbf{Y}}_{t-1,i},\hat{\bar{\pi}}_{t-1,i},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\phi}}\right)\right\}\Bigg\}=0 \quad (16)$$

since both the second term in (15) and (16) are numerically zero (assuming that a constant term was included in the extended POD model).

Then we are done, since the expression inside the expectation in (16) is the same as the summand in (14). In other words, that the equality (16) holds whenever at least one of the POD and POM models is correctly specified follows from the double robustness of $\hat{\mu}_{\text{AIPW}}$. □

Table 1: Details of the simulation studies.

1. *Cross-sectional univariate*:	
True	$s_{\text{true}}(\mathbf{X},\boldsymbol{\beta}) = \boldsymbol{\beta}\left(1, X_1^2, X_2, X_2X_3\right)^T$, $\boldsymbol{\beta} = (0,1,2.5,3)$
	$\text{logit}\left\{\hat{\pi}_{\text{true}}(\mathbf{X},\boldsymbol{\alpha})\right\} = \boldsymbol{\alpha}\left(1, I_1, I_2, I_3, I_1I_2\right)^T$, $\boldsymbol{\alpha} = (-1,1,0,0,-1)$
False	$s_{\text{false}}(\mathbf{X},\boldsymbol{\beta}) = \boldsymbol{\beta}\left(1, X_1, X_2^2\right)^T$
	$\text{logit}\left\{\hat{\pi}_{\text{false}}(\mathbf{X},\boldsymbol{\alpha})\right\} = \boldsymbol{\alpha}\left(1, I_1, I_3\right)^T$
2. *Monotone longitudinal*:	
True	$\tilde{s}_1^{\text{true}}(\mathbf{X},\boldsymbol{\beta}_1) = \boldsymbol{\beta}_1\left(1, X_1, X_1X_3\right)^T$,$\boldsymbol{\beta}_1 = (0,3,2)$
	$s_2^{\text{true}}(\mathbf{X}, Y_1, \boldsymbol{\beta}_2) = \boldsymbol{\beta}_2\left(1, X_1^2, X_2, Y_1^2, X_2Y_1\right)^T$,$\boldsymbol{\beta}_2 = (0,-3,3,1,-2)$
	$s_1^{\text{true}}(\mathbf{X},\boldsymbol{\beta}_1) = \boldsymbol{\beta}_1\left(1, X_2, X_1^2, X_1X_2, X_1^2X_3, X_1X_2X_3, X_1^2X_3^2\right)$
	$\text{logit}\left\{\hat{\pi}_1^{\text{true}}(\mathbf{X},\boldsymbol{\alpha}_1)\right\} = \boldsymbol{\alpha}_1\left(1, I_1^X, I_2^X, I_3^X, I_1^XI_2^X\right)^T$,$\boldsymbol{\alpha}_1 = (1,-1,-1,1,1)$
	$\text{logit}\left\{\hat{\pi}_2^{\text{true}}(\mathbf{X}, Y_1, \boldsymbol{\alpha}_2)\right\} = \boldsymbol{\alpha}_2\left(1, I_1^X, I_2^X, I_3^X, I_1^XI_2^X, I_1^Y, I_3^XI_1^Y\right)^T$,$\boldsymbol{\alpha}_2 = (0,-1,-1,0,1,0,2)$
False	$s_1^{\text{false}}(\mathbf{X},\boldsymbol{\beta}) = \boldsymbol{\beta}\left(1, X_1, X_2\right)^T$
	$s_2^{\text{false}}(\mathbf{X}, Y_1, \boldsymbol{\beta}) = \boldsymbol{\beta}\left(1, X_1, X_2^2, X_3^2, Y_1\right)^T$
	$\text{logit}\left\{\hat{\pi}_1^{\text{false}}(\mathbf{X},\boldsymbol{\alpha})\right\} = \boldsymbol{\alpha}\left(1, I_2^X, I_3^X\right)^T$
	$\text{logit}\left\{\hat{\pi}_2^{\text{false}}(\mathbf{X},\boldsymbol{\alpha})\right\} = \boldsymbol{\alpha}\left(1, I_1^Y\right)^T$
3. *Non-monotone longitudinal*:	
True	$\tilde{s}_1^{\text{true}}(X,\boldsymbol{\beta}_1) = \boldsymbol{\beta}_1\left(1, X^2\right)^T$,$\boldsymbol{\beta}_1 = (0,1)$
	$s_2^{\text{true}}(X, Y_1, \boldsymbol{\beta}_2) = \boldsymbol{\beta}_2\left(1, X, Y_1\right)^T$,$\boldsymbol{\beta}_2 = (0,-1,2)$
	$s_1^{\text{true}}(X, Y_2, \boldsymbol{\beta}_1) = \boldsymbol{\beta}_1\left(1, X, X^2, Y_2\right)$
	$p_1^{\text{true}}(X,\boldsymbol{\alpha}_{11},\boldsymbol{\alpha}_{12}) = \frac{\exp\left\{\boldsymbol{\alpha}_{11}\left(1,\sqrt{\lvert X\rvert}\right)^T\right\}}{1+\exp\left\{\boldsymbol{\alpha}_{11}\left(1,\sqrt{\lvert X\rvert}\right)^T\right\}+\exp\left\{\boldsymbol{\alpha}_{12}\left(1,\sqrt{\lvert X\rvert}\right)^T\right\}}$
	$p_2^{\text{true}}(X,\boldsymbol{\alpha}_{11},\boldsymbol{\alpha}_{12}) = \frac{\exp\left\{\boldsymbol{\alpha}_{12}\left(1,\sqrt{\lvert X\rvert}\right)^T\right\}}{1+\exp\left\{\boldsymbol{\alpha}_{11}\left(1,\sqrt{\lvert X\rvert}\right)^T\right\}+\exp\left\{\boldsymbol{\alpha}_{12}\left(1,\sqrt{\lvert X\rvert}\right)^T\right\}}$
	$\boldsymbol{\alpha}_{11} = (2,-1)$,$\boldsymbol{\alpha}_{12} = (0,0.5)$
	$\text{logit}\left\{p_2^{\text{true}}(X, Y_1, \boldsymbol{\alpha}_2)\right\} = \boldsymbol{\alpha}_2\left(1, X, Y_1^2\right)^T$,$\boldsymbol{\alpha}_2 = (0,-2,0.5)$
False	$s_1^{\text{false}}(X, Y_2, \boldsymbol{\beta}_1) = \boldsymbol{\beta}_1\left(1, X, Y_2\right)^T$
	$s_2^{\text{false}}(X, Y_1, \boldsymbol{\beta}_2) = \boldsymbol{\beta}_2\left(1, Y_1^2\right)^T$
	$p_1^{\text{false}}(X,\alpha_{11},\alpha_{12}) = \frac{\exp(\alpha_{11})}{1+\exp(\alpha_{11})+\exp(\alpha_{12})}$
	$p_2^{\text{false}}(X,\alpha_{11},\alpha_{12}) = \frac{\exp(\alpha_{12})}{1+\exp(\alpha_{11})+\exp(\alpha_{12})}$
	$\text{logit}\left\{p_2^{\text{true}}(X, Y_1, \boldsymbol{\alpha}_2)\right\} = \boldsymbol{\alpha}_2\left(1, X, Y_1\right)^T$
4. *Non-monotone non-longitudinal*:	
True	$\tilde{s}_1^{\text{true}}(X,\boldsymbol{\beta}_1) = \boldsymbol{\beta}_1\left(1, X^2\right)^T$,$\boldsymbol{\beta}_1 = (0,1)$
	$s_2^{\text{true}}(X, Y_1, \boldsymbol{\beta}_2) = \boldsymbol{\beta}_2\left(1, X, Y_1\right)^T$,$\boldsymbol{\beta}_2 = (0,-1,2)$
	$s_1^{\text{true}}(X, Y_2, \boldsymbol{\beta}_1) = \boldsymbol{\beta}_1\left(1, X, X^2, Y_2\right)$
	$p_1^{\text{true}}(X,\boldsymbol{\alpha}_{11},\boldsymbol{\alpha}_{12}) = \frac{\exp\left\{\boldsymbol{\alpha}_{11}\left(1,X,X^2\right)^T\right\}}{1+\exp\left\{\boldsymbol{\alpha}_{11}\left(1,X,X^2\right)^T\right\}+\exp\left\{\boldsymbol{\alpha}_{12}\left(1,X,X^2\right)^T\right\}}$
	$p_2^{\text{true}}(X,\boldsymbol{\alpha}_{11},\boldsymbol{\alpha}_{12}) = \frac{\exp\left\{\boldsymbol{\alpha}_{12}\left(1,X,X^2\right)^T\right\}}{1+\exp\left\{\boldsymbol{\alpha}_{11}\left(1,X,X^2\right)^T\right\}+\exp\left\{\boldsymbol{\alpha}_{12}\left(1,X,X^2\right)^T\right\}}$
	$\boldsymbol{\alpha}_{11} = (1,-0.5,0.2)$,$\boldsymbol{\alpha}_{12} = (0,0.5,-0.3)$
	$\text{logit}\left\{p_2^{\text{true}}(X, Y_1, \boldsymbol{\alpha}_{22})\right\} = \boldsymbol{\alpha}_{22}\left(1, X, Y_1\right)^T$,$\boldsymbol{\alpha}_{22} = (0,-1,0.3)$
	$\text{logit}\left\{p_1^{\text{true}}(X, Y_2, \boldsymbol{\alpha}_{21})\right\} = \boldsymbol{\alpha}_{21}\left(1, X, Y_2\right)^T$,$\boldsymbol{\alpha}_{21} = (0,-1,0.3)$
False	$s_1^{\text{false}}(X, Y_2, \boldsymbol{\beta}_1) = \boldsymbol{\beta}_1\left(1, X^2, Y_2\right)^T$
	$s_2^{\text{false}}(X, Y_1, \boldsymbol{\beta}_2) = \boldsymbol{\beta}_2\left(1, Y_1\right)^T$
Key: I_l^Z stands for $I(Z_l > 0)$	

Table 2: The results of simulation studies comparing the doubly robust multiple imputation (RMI) estimator with the inverse probability weighted complete case (IPW), outcome regression (OR) and multiple imputation (MI) estimators; and the doubly robust (DR) estimator introduced by Bang and Robins. Results for the cross-sectional univariate, monotone longitudinal, non-monotone longitudinal and non-monotone non-longitudinal cases are given. No subscript indicates correct specification of the relevant model(s). $\pi-$false indicates that the estimator used an incorrectly-specified POM model, $y-$false indicates that the estimator used an incorrectly-specified POD model and $\pi \oplus y-$false indicates that both the POM and POD models were incorrectly specified.

Estimator	**Bias**	**True variance**	**Estimated variance**	**Coverage probability**
1. *Cross-sectional univariate*:				
$\hat{\mu}_{\text{IPW}}$	−0.01	0.11	–	–
$\hat{\mu}_{\text{OR}}$	−0.00	0.04	–	–
$\hat{\mu}_{\text{DR}}$	−0.00	0.04	–	–
$\hat{\mu}_{\text{RMI}}$	−0.00	0.04	0.04	0.95
$\hat{\mu}_{\text{IPW}\cdot\pi-\text{false}}$	−0.36	0.13	–	–
$\hat{\mu}_{\text{DR}\cdot\pi-\text{false}}$	−0.00	0.04	–	–
$\hat{\mu}_{\text{RMI}\cdot\pi-\text{false}}$	−0.01	0.04	0.04	0.95
$\hat{\mu}_{\text{OR}\cdot y-\text{false}}$	−0.35	0.12	–	–
$\hat{\mu}_{\text{DR}\cdot y-\text{false}}$	−0.01	0.11	–	–
$\hat{\mu}_{\text{RMI}\cdot y-\text{false}}$	−0.02	0.12	0.12	0.93
$\hat{\mu}_{\text{DR}\cdot\pi\oplus y-\text{false}}$	−0.35	0.13	–	–
$\hat{\mu}_{\text{RMI}\cdot\pi\oplus y-\text{false}}$	−0.35	0.14	0.12	0.79
2. *Monotone longitudinal*:				
$\hat{\mu}_{\text{IPW}}$	−0.11	10.98	–	–
$\hat{\mu}_{\text{OR}}$	0.06	1.92	–	–
$\hat{\mu}_{\text{DR}}$	0.06	1.92	–	–
$\hat{\mu}_{\text{RMI}}$	0.07	1.91	1.83	0.94
$\hat{\mu}_{\text{IPW}\cdot\pi-\text{false}}$	−3.21	5.87	–	–
$\hat{\mu}_{\text{DR}\cdot\pi-\text{false}}$	0.06	1.92	–	–
$\hat{\mu}_{\text{RMI}\cdot\pi-\text{false}}$	0.08	1.92	1.83	0.93
$\hat{\mu}_{\text{OR}\cdot y-\text{false}}$	−4.99	3.51	–	–
$\hat{\mu}_{\text{DR}\cdot y-\text{false}}$	−0.36	10.51	–	–
$\hat{\mu}_{\text{RMI}\cdot y-\text{false}}$	−0.37	10.63	4.28	0.74
$\hat{\mu}_{\text{DR}\cdot\pi\oplus y-\text{false}}$	−2.35	8.13	–	–
$\hat{\mu}_{\text{RMI}\cdot\pi\oplus y-\text{false}}$	−2.37	7.38	3.67	0.57
3. *Non-monotone longitudinal*:				
$\hat{\mu}_{\text{IPW}}$	0.00	0.07	–	–
$\hat{\mu}_{\text{MI}}$	−0.01	0.03	–	–
$\hat{\mu}_{\text{RMI}}$	−0.02	0.03	0.03	0.95
$\hat{\mu}_{\text{IPW}\cdot\pi-\text{false}}$	−0.59	0.05	–	–
$\hat{\mu}_{\text{RMI}\cdot\pi-\text{false}}$	−0.03	0.03	0.03	0.94
$\hat{\mu}_{\text{MI}\cdot y-\text{false}}$	3.07×10^{31}	2.16×10^{65}	–	–
$\hat{\mu}_{\text{RMI}\cdot y-\text{false}}$	0.00	0.04	0.06	0.97
$\hat{\mu}_{\text{RMI}\cdot\pi\oplus y-\text{false}}$	2.32	123.55	5.27×10^{8}	0.94
4. *Non-monotone non-longitudinal*:				
$\hat{\mu}_{\text{IPW}}$	0.01	0.07	–	–
$\hat{\mu}_{\text{MI}}$	0.00	0.03	–	–
$\hat{\mu}_{\text{RMI}}$	−0.00	0.03	0.03	0.95
$\hat{\mu}_{\text{IPW}\cdot\pi-\text{false}}$	0.25	0.06	–	–
$\hat{\mu}_{\text{RMI}\cdot\pi-\text{false}}$	0.00	0.03	0.03	0.95
$\hat{\mu}_{\text{MI}\cdot y-\text{false}}$	0.49	0.05	–	–
$\hat{\mu}_{\text{RMI}\cdot y-\text{false}}$	−0.04	0.03	0.03	0.95
$\hat{\mu}_{\text{RMI}\cdot\pi\oplus y-\text{false}}$	0.22	0.04	0.04	0.80