Inverse Probability Weighting with Missing Predictors of Treatment Assignment or Missingness

Abstract

Inverse probability weighting (IPW) can deal with confounding in non-randomised studies. The inverse weights are probabilities of treatment assignment (propensity scores), estimated by regressing assignment on predictors. Problems arise if predictors can be missing. Solutions previously proposed include assuming assignment depends only on observed predictors and multiple imputation (MI) of missing predictors. For the MI approach it was recommended that missingness indicators be used with the other predictors.

We determine when the two MI approaches (with/without missingness indicators) yield consistent estimators and compare their efficiencies. We find that, although including indicators can reduce bias when predictors are missing not at random, it can induce bias when they are missing at random. We propose a consistent variance estimator and investigate performance of the simpler Rubin's Rules variance estimator. In simulations we find both estimators perform well.

IPW is also used to correct bias when an analysis model is fitted to incomplete data by restricting to complete cases. Here weights are inverse probabilities of being a complete case. We explain how the same MI methods can be used in this situation to deal with missing predictors in the weight model, and illustrate this approach using data from the National Child Development Survey.

Running title: IPW with Missing Predictors

1 Introduction

In a randomised controlled trial individuals are randomly assigned to one of two or more treatments and an outcome is measured. The randomisation ensures that the measured effect of treatment on outcome is not confounded by other variables. In an observational study the assignment of individuals to a treatment is not random, and so the observed association between treatment and outcome may be confounded. If the confounding variables are observed, they can be adjusted for in the analysis. This may be done using regression models, in which confounders are included as covariates alongside treatment, or by using propensity scores (PS) (Rosenbaum and Rubin, 1983). Cepeda et al. (2003) and Stürmer et al. (2005) discuss the relation between, and relative advantages of, these two approaches. In the present article we are concerned with the PS approach in the situation of a binary treatment variable. We shall call the two treatments 'active' and 'control'. The term 'treatment' should be interpreted liberally: it could be any binary exposure.

In the PS approach a model is specified for the probability that an individual receives the active treatment. The covariates, $\boldsymbol{X}$, in this model are called treatment predictor variables, and the fitted probabilities from the model are called propensity scores. Rosenbaum and Rubin showed that if the set of treatment predictor variables include all confounders in the association between treatment and outcome, then adjustment for the PS is sufficient to obtain an unconfounded estimate of the treatment effect. Adjustment can be performed by stratifying or matching on the PS or by weighting by the reciprocal of the PS. The latter approach is known as inverse probability weighting (IPW). When the PS model is correctly specified, IPW yields a consistent estimator of treatment effect, unlike stratification, which is subject to residual confounding (Lunceford and Davidian, 2004). In the present article we are concerned with IPW.

IPW can also be used when estimating a population mean outcome from a sample in which the outcome variable is sometimes missing. In this situation one might estimate the population mean by the sample mean in individuals with observed outcome (the 'complete cases'). This 'complete-case' analysis yields consistent estimation when the probability that an individual's outcome is observed does not depend on that outcome, but it may be biased otherwise. The Horwitz-Thompson (IPW) estimator (Horwitz and Thompson, 1958) provides a straightforward way of correcting this bias. Again, only individuals with observed outcome are included, but weights are used to rebalance the set of complete cases so that it is representative of the whole sample. Each individual's weight is the inverse of their probability of being a complete case. Normally, this probability is unknown and needs to be estimated. This is done by specifying a model for the conditional probability of an individual being a complete case given a set of predictor variables. This application of IPW is also used when fitting a more general regression model (known as the 'analysis model'). In this more general situation, the complete cases are those individuals for whom all variables in the analysis model are observed.

There is a strong parallel between using IPW to deal with missing data and using it to deal with confounding in non-randomised studies. Estimating a population mean outcome when outcome can be missing in the sample is analogous to estimating the mean outcome that would result if everyone were assigned to active treatment using data from a sample in which some individuals are assigned to control treatment. In the latter case, the PS is the probability of being assigned to active treatment; in the former, it is the probability of being a complete case. In this paper, our real-data example (Section 7) concerns the use of IPW to deal with missing data. We shall study methods in the more complicated situation of confounding in non-randomised studies and then show how these methods transfer to the simpler situation of missing data.

When there are missing values in $\boldsymbol{X}$, estimation of the PS is not straightforward. A number of approaches have been suggested in the setting of estimating a treatment effect in a non-randomised study. D'Agostino and Rubin (2000) assume that the PS depends only on observed predictors. This implies that the PS model is different in different individuals: if all predictors are observed on an individual, his/her PS may depend on all predictors; if some are missing, his/her PS may not depend on these. Once this assumption has been made, the simplest approach is to stratify the individuals according to which predictors have been observed and then fit a separate PS model to each stratum. The number of individuals in some strata may, however, be small, which could cause problems when fitting the PS models in these strata. D'Agostino and Rubin proposed instead modelling the joint distribution of $\boldsymbol{X}$, T and R, where $T = 1$ if the individual is assigned to active treatment and $T = 0$ if assigned to control, and R denotes the missingness pattern of $\boldsymbol{X}$, i.e. it denotes which predictors of propensity are observed. The model is fitted using an Expectation Conditional Maximisation algorithm. One drawback with this approach is its unappealing assumption that the PS depends on a predictor of propensity only if it is observed, i.e. that a variable is not a confounder if it is unobserved. A second drawback is the difficulty of interpreting the parameter constraints needed to make the joint model for $(\boldsymbol{X}, T, R)$ estimable.

Qu and Lipkovich (2009) proposed multiply imputing missing values of $\boldsymbol{X}$ using the observed values of $\boldsymbol{X}$, T and the outcome, thus creating M multiple datasets in which $\boldsymbol{X}$ is complete. For each completed dataset, the PS model is fitted, PS's are estimated and the inverse PS's are used as weights in the estimator of treatment effect. The M treatment effect estimates are then averaged. In a refinement of this approach, Qu and Lipkovich (2009) propose including R as an additional covariate in the PS model. They explain that this may reduce bias when $\boldsymbol{X}$ is missing not at random. However, no formal justification for their methods is provided.

Mitra and Reiter (2011) also proposed multiply imputing missing $\boldsymbol{X}$. Their aim was to make inference more robust to misspecification of the imputation model. A drawback of their method is that the imputation model excludes the outcome data, and so missing $\boldsymbol{X}$ are imputed without using the observed outcome. This means that $\boldsymbol{X}$ is imputed using a model which assumes $\boldsymbol{X}$ is not a confounder.

In applied work, Mattei (2009) and Song et al. (2001) also deal with missing predictors of propensity by using MI, but without investigating the properties of their methods or providing theoretical justification for them. Their descriptions of the methods they used are somewhat limited, but these methods would appear to be the same as, or very similar to, that of Qu and Lipkovich (2009). Hayes and Groner (2008) multiply impute missing predictors and calculate propensity scores for each imputed dataset. However, they then choose one PS at random for each individual. Uncertainty in PS is ignored.

The purpose of the present article is fourfold. First, we investigate Qu and Lipkovich's (2009) two imputation methods, showing under what conditions each yields consistent parameter estimation and comparing their efficiencies. Second, as these estimators are not maximum likelihood estimators (MLE), it is not obvious that Rubin's Rules will apply in this case (Robins and Wang, 2000; Nielsen, 2003). Qu and Lipkovich (2009) proposed that variance estimates be obtained by bootstrapping, a computationally intensive procedure. We investigate how the simple Rubin's Rules variance estimator performs in this setting. Third, MI may be proper or improper. In proper MI, the uncertainty in the parameters of the imputation model is accounted for by including in the imputation procedure a random draw from the posterior distribution of these parameters. In improper MI, this step is omitted and the MLEs of the parameters are used instead. In most applications of MI, proper imputation is used, because it enables the variance to be estimated using Rubin's Rules. For improper MI, on the other hand, a closed-form variance estimator is available

(Robins and Wang, 2000). This latter estimator is complicated, but has the advantage of being valid even when the parameter estimator is not a MLE, as is the case here. Qu and Lipkovich (2009) use proper MI. We describe the analogous improper MI procedure and its closed-form variance estimator. Fourth, Qu and Lipkovich (2009) were concerned with estimating a simple treatment difference in a non-randomised study. We show how these methods can also be used to estimate the population mean of an outcome when this outcome can be missing, and extend them to more general analysis models.

The structure of the article is as follows. In Section 2 the PS approach with fully observed predictors of propensity is described. In Section 3 we describe Qu and Lipkovich's (2009) imputation method and prove consistency of their parameter estimator when R is not in the PS model. In Section 4 we examine the effect of including R. Section 5 contains a simulation study comparing various approaches for handling missing predictors. We look at asymptotic and finite-sample biases and at the coverage of confidence intervals constructed using both our explicit variance estimator and the Rubin's Rules variance estimator. In Section 6 we show how Qu and Lipkovich's (2009) methods transfer to the estimation of a population mean from a sample with missing outcomes and to more general analysis models. An application of these methods to data from the National Child Development Survey (NCDS) is described in Section 7. We end with a discussion.

2 PS Approach with Fully Observed Predictors

Let D_1 denote an individual's potential outcome if assigned to active treatment and D_0 denote the outcome if assigned to control. Only one of these can be observed. If $T = 1$, D_1 is observed and D_0 is missing; if $T = 0$, D_0 is observed and D_1 is missing. Let $\theta = E(D_1) - E(D_0)$ denote the average treatment effect and let θ_0 denote the true value of θ. Let $D_{\text{obs}} = TD_1 + (1 - T)D_0$ denote the observed outcome.

A model (e.g. a logistic regression model) $\pi(\boldsymbol{X}; \boldsymbol{\alpha})$ is specified for $\pi(\boldsymbol{X}) = P(T = 1 \mid \boldsymbol{X})$, where $\boldsymbol{\alpha}$ denotes unknown parameters. This is the PS model. Assume this is correctly specified and let $\boldsymbol{\alpha}_0$ denote the true value of $\boldsymbol{\alpha}$. So, $\pi(\boldsymbol{X}) = \pi(\boldsymbol{X}; \boldsymbol{\alpha}_0)$. The following additional assumptions are made:

$$(A1) \qquad T \perp\!\!\!\perp D_1, D_0 \mid \boldsymbol{X}$$

$$(A2) \qquad \exists c > 0 \text{ such that } P\{c < \pi(\boldsymbol{X}; \boldsymbol{\alpha}_0) < 1 - c\} = 1$$

$(A1)$ means that, given $\boldsymbol{X}$, treatment assignment is independent of the two potential outcomes. $(A2)$ means that, with probability one, a randomly chosen individuals will have positive probabilities of being assigned to each of the two treatments.

Suppose a sample of n individuals is drawn. Let subscript i denote individual i, and let $\hat{\boldsymbol{\alpha}}$ denote the solution to a set of consistent estimating equations $\sum_{i=1}^{n} \boldsymbol{S}_\alpha(\boldsymbol{\alpha}; \boldsymbol{X}_i, T_i) = \boldsymbol{0}$ for $\boldsymbol{\alpha}$. For example, if $\pi(\boldsymbol{X}; \boldsymbol{\alpha})$ is a logistic regression model fitted by maximum likelihood, then $\boldsymbol{S}_\alpha(\boldsymbol{\alpha}; \boldsymbol{X}_i, T_i)$ is the contribution of individual i to the score equations of, and $\hat{\boldsymbol{\alpha}}$ is the MLE from, the logistic regression of $T_1, \ldots, T_n$ on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

A consistent estimator of θ is (Lunceford and Davidian, 2004)

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_i D_{\text{obs},i}}{\pi(\boldsymbol{X}; \hat{\boldsymbol{\alpha}})} - \frac{(1 - T_i) D_{\text{obs},i}}{1 - \pi(\boldsymbol{X}; \hat{\boldsymbol{\alpha}})} \right\} \tag{1}$$

and a consistent estimator of the variance of $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\alpha}}^T)^T$ is

$$\left(\sum_{i=1}^{n} \left. \frac{\partial \boldsymbol{U}_i}{\partial \boldsymbol{\beta}^T} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{U}_i \boldsymbol{U}_i^T \right) \left(\sum_{i=1}^{n} \left. \frac{\partial \boldsymbol{U}_i}{\partial \boldsymbol{\beta}^T} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} \right)^{-1} \tag{2}$$

where $\boldsymbol{U}_i = (\boldsymbol{S}_\theta^T(\theta; \boldsymbol{X}_i, D_{\text{obs},i}, T_i), \boldsymbol{S}_\alpha^T(\boldsymbol{\beta}; \boldsymbol{X}_i, T_i))^T$ and

$$S_\theta(\theta, \boldsymbol{\alpha}; \boldsymbol{X}, D_{\text{obs}}, T) = \frac{T D_{\text{obs}}}{\pi(\boldsymbol{X}; \boldsymbol{\alpha})} - \frac{(1 - T) D_{\text{obs}}}{1 - \pi(\boldsymbol{X}; \boldsymbol{\alpha})} - \theta \tag{3}$$

3 PS Approach with Missing Predictors

We now describe two MI procedures for estimating $\boldsymbol{\theta}$ when $\boldsymbol{X}$ is not fully observed.

Let $\boldsymbol{X}_{\text{obs}}$ and $\boldsymbol{X}_{\text{mis}}$ denote the observed and missing parts of $\boldsymbol{X}$, respectively, and let $\boldsymbol{W} = (\boldsymbol{X}_{\text{obs}}, D_{\text{obs}}, T)$. We shall use M to denote the number of imputations. A model $f(\boldsymbol{X} \mid D_{\text{obs}}, T; \boldsymbol{\psi})$, with parameters $\boldsymbol{\psi}$, is specified for the distribution of $\boldsymbol{X}$ given D_{obs} and T. If a component, $\boldsymbol{X}_{\text{full}}$, of $\boldsymbol{X}$ is fully observed, a model may instead be specified for $f(\boldsymbol{X} \mid \boldsymbol{X}_{\text{full}}, D_{\text{obs}}, T; \boldsymbol{\psi})$, the distribution of $\boldsymbol{X}$ given $\boldsymbol{X}_{\text{full}}$, D_{obs} and T. Assume that this model and the PS model $\pi(\boldsymbol{X};\boldsymbol{\alpha})$ are correctly specified. Qu and Lipkovich (2009) propose the following proper MI procedure.

1. Calculate the posterior distribution of $\boldsymbol{\psi}$ implied by likelihood function $f(\boldsymbol{X} \mid D_{\text{obs}}, T; \boldsymbol{\psi})$, observed data $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$ and a non-informative prior.

2. Sample $\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(M)}$ from this posterior distribution.

3. For each $m = 1, \ldots, M$ and $i = 1, \ldots, n$, sample $\boldsymbol{X}^{*(m)}_{\text{mis},i}$ from the distribution $g(\boldsymbol{X}_{\text{mis},i} \mid \boldsymbol{W}_i; \boldsymbol{\psi}^{(m)})$ implied by model $f(\boldsymbol{X} \mid D_{\text{obs}}, T; \boldsymbol{\psi}^{(m)})$. Let $\boldsymbol{X}^{*(m)}_i = (\boldsymbol{X}_{\text{obs},i}, \boldsymbol{X}^{*(m)}_{\text{mis},i})$.

4. For each $m = 1, \ldots, M$, let $\hat{\boldsymbol{\alpha}}^{(m)}$ denote the solution to estimating equations $n^{-1}\sum_{i=1}^n \boldsymbol{S}_\alpha(\hat{\boldsymbol{\alpha}}^{(m)}; T_i, \boldsymbol{X}^{*(m)}_i) = \mathbf{0}$.

5. For each $m = 1, \ldots, M$, calculate
$$\hat{\theta}^{(m)} = \frac{1}{n}\sum_{i=1}^n \left\{ \frac{T_i D_{1i}}{\pi(\boldsymbol{X}^{*(m)}_i; \hat{\boldsymbol{\alpha}}^{(m)})} - \frac{(1-T_i)D_{0i}}{1-\pi(\boldsymbol{X}^{*(m)}_i; \hat{\boldsymbol{\alpha}}^{(m)})} \right\}. \tag{4}$$

6. Calculate $\hat{\theta}_A = M^{-1}\sum_{m=1}^M \hat{\theta}^{(m)}$.

An alternative, improper MI procedure is as follows.

1. Calculate the MLE, $\hat{\boldsymbol{\psi}}$, of $\boldsymbol{\psi}$ from likelihood function $f(\boldsymbol{X} \mid D_{\text{obs}}, T; \boldsymbol{\psi})$ and observed data $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$.

2. For each $m = 1, \ldots, M$ and $i = 1, \ldots, n$, generate $\boldsymbol{X}^{*(m)}_{\text{mis},i}$ from $g(\boldsymbol{X}_{\text{mis},i} \mid \boldsymbol{W}_i; \hat{\boldsymbol{\psi}})$. Let $\boldsymbol{X}^{*(m)}_i = (\boldsymbol{X}_{\text{obs},i}, \boldsymbol{X}^{*(m)}_{\text{mis}})$.

3. Calculate $\hat{\boldsymbol{\alpha}}$ as the solution to $(nM)^{-1}\sum_{i=1}^n \sum_{m=1}^M \boldsymbol{S}_\alpha(\boldsymbol{X}^{*(m)}_i, T; \hat{\boldsymbol{\alpha}}) = \mathbf{0}$.

4. Calculate
$$\hat{\theta}_B = \frac{1}{nM}\sum_{i=1}^n \sum_{m=1}^M \left\{ \frac{T_i D_{1i}}{\pi(\boldsymbol{X}^{*(m)}_i; \hat{\boldsymbol{\alpha}})} - \frac{(1-T_i)D_{0i}}{1-\pi(\boldsymbol{X}^{*(m)}_i; \hat{\boldsymbol{\alpha}})} \right\} \tag{5}$$

These two MI procedures differ in two ways. The first procedure estimates θ using proper imputation of $\boldsymbol{X}$ and Rubin's Rule for the mean, i.e. $\boldsymbol{\alpha}$ and θ are estimated separately for each of the M imputed datasets and then the estimates of θ are averaged. The second procedure uses improper imputation and calculates a single estimate of $(\boldsymbol{\alpha}, \theta)$ directly from the whole set of M imputations.

Assume $(A1)$, $(A2)$ and $(A3)$ are true, where $(A3)$ is

$$(A3) \qquad p(R \mid \boldsymbol{X}, D_{\text{obs}}, T) = p(R \mid \boldsymbol{X}_{\text{obs}}, D_{\text{obs}}, T)$$

(i.e. $\boldsymbol{X}$ is MAR given D_{obs} and T). In Appendix B we prove that when these conditions are satisfied and $M = \infty$, $\hat{\theta}_A$ and $\hat{\theta}_B$ are asymptotically equivalent, consistent estimators of θ. Moreover, assuming that $\hat{\theta}_A$ and $\hat{\theta}_B$ are consistent when $M < \infty$, $\hat{\theta}_B$ is asymptotically more efficient than $\hat{\theta}_A$ when $M < \infty$ and the variance of $\hat{\theta}_B$ is consistently estimated by the formula given in Appendix C. For the Rubin's Rules variance estimator of $\hat{\boldsymbol{\theta}}_A$ the complete-data variance estimator we use is that given by equation (2).

In Appendix A we present an alternative pair of estimators of treatment effect, in which equations (4) and (5) are modified by dividing by the sum of the weights. We also present estimators of the treatment ratio, $E(D_1)/E(D_0)$.

4 Including R in the PS Model

Qu and Lipkovich (2009) recommend additionally including R in the PS model, saying it may reduce the bias in $\hat{\theta}_A$ when Assumption $(A3)$ is violated, i.e. when $\boldsymbol{X}$ is not MAR given D_{obs} and T. We now explore the consistency of $\hat{\theta}_A$ and $\hat{\theta}_B$ when R is included in the PS model and the efficiency relative to when R is not included.

Including R in the PS model implies replacing Assumption $(A1)$ by $(A1')$:

$$(A1') \qquad T \perp\!\!\!\perp D_1, D_0 \mid \boldsymbol{X}, R$$

When $(A1)$ and $(A3)$ are true, $(A1')$ is not true in general. An example illustrates this. Suppose that $R = r$, $\boldsymbol{X} = \boldsymbol{x}$ and $P(R = r \mid \boldsymbol{X} = \boldsymbol{x}, D_{\text{obs}}, T) = P(R = r \mid \boldsymbol{X}_{\text{obs}} = \boldsymbol{x}_{\text{obs}}, D_{\text{obs}})$ is an increasing function of D_{obs}. Then the probability that $T = 1$ is greater if $D_1 > D_0$ than if $D_0 > D_1$. Therefore $(A1')$ is false.

If $(A1)$ and $(A4)$ are true, where $(A4)$ is

$$(A4) \qquad R \perp\!\!\!\perp D_0, D_1 \mid \boldsymbol{X}, T$$

then $(A1')$ is also true, and so including R will not induce bias. A stronger assumption than both $(A3)$ and $(A4)$ is $(A5)$:

$$(A5) \qquad p(R \mid \boldsymbol{X}, D_0, D_1, T) = p(R \mid \boldsymbol{X}_{\text{obs}}, T)$$

So, when Assumptions $(A1)$ and $(A5)$ are true, Assumptions $(A1')$ and $(A3)$ are also true. In this case, including R in the PS model should not induce bias, although there is no need to include R, because it is not a confounder in the relation between T and (D_0, D_1) given $\boldsymbol{X}$. Moreover, there may be some loss of efficiency if it is included. This is because including R will cause individuals with the same values of T and $\boldsymbol{X}$ but different values of R to receive different weights, and because (D_0, D_1) is distributed equally in such individuals, efficiency is lost by weighting them differently. Asymptotically, however, the efficiency loss tends to zero (Tsiatis, 2006).

Qu and Lipkovich (2009) describe a simulation study in which $p(R \mid \boldsymbol{X}, D_0, D_1, T) = p(R \mid \boldsymbol{X})$, so that $(A4)$ is true. They found that including R made no difference to bias (as expected) and that the efficiency loss was very small.

Qu and Lipkovich (2009) suggested that including R would reduce bias when $(A3)$ is false, i.e. when $\boldsymbol{X}$ is MNAR given T and D_{obs}. They imagined an extreme MNAR scenario in which $\boldsymbol{X} = (\boldsymbol{X}^a, X^b)$, where $\boldsymbol{X}^a$ is fully observed and X^b is binary. The variable X^b was assumed to be always observed ($R = 1$) if $X^b = 0$ and always missing ($R = 0$) if $X^b = 1$. In this extreme situation R is a one-to-one mapping of X^b and so R can replace X^b in the PS model. In realistic situations the MNAR mechanism will be weaker and the missing variables may not be binary, so R will not be a one-to-one mapping. Whether including R increases or reduces the bias resulting from $\boldsymbol{X}$ not being MAR given D_{obs} and T will depend on the strength of the association between $\boldsymbol{X}_{\text{mis}}$ and R given $\boldsymbol{X}_{\text{obs}}$ and on the extent of deviation from Assumption $(A4)$. Qu and Lipkovich (2009) describe a MNAR simulation in which R is independent of T, D_0 and D_1 given $\boldsymbol{X}$. They found that including R reduced bias in this situation. As including R does not introduce bias when $(A4)$ is true, this is as expected. In the next section we consider bias under a wider range of MNAR mechanisms.

5 Asymptotic and Simulation Study

We now describe a study of asymptotic bias and finite-sample bias and efficiency, comparing several methods for dealing with missing predictors of propensity when using the PS approach to estimate average treatment effect. Both MAR and MNAR predictors of propensity will be considered.

We consider scenarios in which the outcome, D, and the predictors of propensity, X_1 and X_2, are binary variables and X_1 is fully observed. We assume $P(X_1 = 1) = 0.5$, $P(X_2 = 1) = 0.2 + 0.6X_1$, $P(T = 1 \mid X_1, X_2) = \{1 + \exp(1.5 - X_1 - 2X_2)\}^{-1}$, and $P(D_t = 1 \mid X_1, X_2) = \{1 + \exp(1 - X_1 - X_2 - 2t)\}^{-1}$. So, X_1 and X_2 are positively correlated, X_1 and X_2 both increase the probability of assignment to active treatment, and X_1, X_2 and active treatment all independently increase the probability of outcome $D = 1$. With these choices, $P(T = 1) = 0.5$, $P(D = 1) = 0.64$, the treatment effects (i.e. treatment differences) are 0.46, 0.38, 0.38 and 0.22 in the four strata defined by $(X_1, X_2) = (0, 0)$, $(0, 1)$, $(1, 0)$ and $(1, 1)$, respectively, and the overall treatment effect is $\theta_0 = 0.35$. Let $R = 1$ if X_2 is observed; $R = 0$ otherwise. The probability that X_2 is missing (i.e. $R = 0$) was $\{1+\exp(-\gamma_0 - X_1 - \gamma_2 X_2 - \gamma_T T - \gamma_D D_{\text{obs}})\}^{-1}$. When $\gamma_2 = 0$, $\boldsymbol{X}$ is MAR given D_{obs} and T. A variety of values of γ_T and γ_D were considered; γ_0 was chosen to make $P(R = 1) = 0.5$.

We used the method of Rotnitzky and Wypij (1994) to calculate the asymptotic biases of the estimators of treatment effects from seven methods:

Complete Data (Comp): using X_1 and X_2 in the PS model (before deleting missing X_2 values).

No Adjust (NoAdj): no adjustment for confounding, i.e. the difference between the means of the observed outcomes in the two treatment groups.

Partly Adjusted (PartAdj): using only X_1 in the PS model.

Missing indicator method (MissI): Using X_1, RX_2 and R in PS model.

Separate PS models by R (SepPS): Using X_1, RX_2, R and RX_1 in PS model (so effectively using two different PS models: one for individuals with observed X_2 and one for those with missing X_2).

Improper MI (Imp): using X_1 and X_2 in the PS model, and imputing missing X_2 values using improper MI with $M = \infty$.

Improper MI with R (ImpR): Same as Imp, but also using R in the PS model.

As mentioned earlier, when $M = \infty$ the treatment effect estimators from proper MI are asymptotically equivalent to those from improper MI. For Imp and ImpR, a saturated imputation model was used, i.e. $P(X_2 = 1 \mid X_1, T, D_{\text{obs}})$ was allowed to be different for each of the eight combinations of X_1, T and D_{obs}.

The asymptotic biases of Comp, NoAdj and PartAdj do not depend on $\boldsymbol{\gamma}$; they are 0.000, 0.174 and 0.064, respectively. Table 2 shows asymptotic biases for the other four methods for a variety of values of γ_0, γ_2, γ_D and γ_T. We consider four MAR scenarios: one where neither outcome nor treatment assignment affects the probability that X_2 is observed ($\gamma_T = \gamma_D = 0$); one where only treatment assignment affects it ($\gamma_T = 1$, $\gamma_D = 0$); one where only outcome affects it ($\gamma_T = 0$, $\gamma_D = 1$); and one where both affect it ($\gamma_T = \gamma_D = 1$). As expected, we see that Imp is asymptotically unbiased in all four MAR scenarios, but ImpR is only unbiased when outcome does not affect the probability that X_2 is observed. MissI and SepPS are biased in all scenarios.

Table 2 also shows asymptotic biases when $\gamma_2 = 2$, and so $\boldsymbol{X}$ is MNAR. As expected, Imp is no longer asymptotically unbiased. Its bias may be more or less than the biases of the other three methods. We also examined what happens when γ_2 assumes larger values (data not shown), concentrating on the case where $\gamma_D = \gamma_T = 0$. As γ_2 increases, whether X_2 is observed increasingly predicts whether $X_2 = 1$ and, as this happens, ImpR is expected to become less asymptotically biased than Imp, for the reasons explained in Section 4. Indeed, when $\gamma_2 = 4$, the asymptotic bias is 0.024 for Imp but only -0.011 for ImpR; when

$\gamma_2 = 8$, the biases were, respectively, 0.056 and -0.003. Likewise, the asymptotic biases of MissI and SepPS tend to diminish: when $\gamma_2 = 4$ they are 0.026 and 0.017, respectively; when $\gamma_2 = 8$ they are 0.005 and 0.003.

By simulating 1000 datasets, each with sample size $n = 500$, for each scenario, we also estimated the finite sample biases, empirical SEs and coverages of 95% confidence intervals. The biases of Comp, NoAdj and PartAdj were estimated as 0.0008, 0.173 and 0.064, respectively, agreeing closely with the asymptotic biases. The corresponding empirical SEs were 0.062, 0.036 and 0.045. Coverages were 95%, 0% and 69%. The imputation methods that we applied were Imp1 and Imp10 (missing X_2 imputed using improper MI with $M = 1$ and $M = 10$, respectively), Pro10 (proper MI with $M = 10$), and ImpR10 and ProR10 (like Imp10 and Pro10 but with R included in the PS model). A saturated imputation model was used for each imputation method, and for proper MI independent Beta$(1, 1)$ (i.e. uniform) priors were used for each element of $\boldsymbol{\psi}$. Confidence intervals were based on the Robins' variance estimator (see Appendix C) for Imp1, Imp10 and ImpR10, and on the Rubin's Rules variance estimator for Pro10 and ProR10.

Table 1 shows empirical SEs and coverages of confidence intervals for these imputation methods and for MissI and SepPS. Finite-sample biases are not shown, as these are very close to the corresponding asymptotic biases reported in Table 2; biases for Pro10 and ProR10 are very similar to those for Imp10 and ImpR10, respectively. Empirical SEs are reduced by using $M = 10$ imputations rather than $M = 1$ (compare Imp10 and Imp1). As expected (see Section 4), including R in the PS model leads to an increase in the empirical SE when $\gamma_D = 0$ and $\gamma_T = 1$. This difference becomes increasingly marked as γ_T increases: the SEs of Imp10 and ImpR10 are 0.065 and 0.085, respectively, when $\gamma_T = 3$, $\gamma_D = \gamma_2 = 0$ (data not shown). Coverages of Imp10 and Pro10 were close to their nominal levels when $\boldsymbol{X}$ is MAR given D_{obs} and T, suggesting that Rubin's Rule for the variance is valid.

The empirical SEs of Pro10 are generally slightly smaller than those of Imp10. Asymptotically (as $n \to \infty$), the SE of Imp10 should be smaller than that of Pro10 when $M < \infty$, and asymptotically equal to it when $M = \infty$ (Robins and Wang, 2000). So, we investigated further the case of $\gamma_T = 3$ and $\gamma_D = \gamma_2 = 0$, which was the MAR scenario where the difference was greatest. With independent Beta$(1,1)$ priors for the parameters, $\boldsymbol{\psi}$, of the imputation model, the SEs of Imp10 and Pro10 were 0.065 and 0.062, respectively. When these priors were replaced by Beta$(0,0)$ priors, the SE for Pro10 was 0.067, greater than that of Imp10 (0.065). Under improper Beta$(0,0)$ priors, the posterior mean of $\boldsymbol{\psi}$ is equal to its MLE, whereas Beta$(1,1)$ priors cause the posterior mean of each element of $\boldsymbol{\psi}$ to be closer to 0.5 than its corresponding MLE. This will slightly reduce the variance of the distribution of weights and hence reduce the SE. As the sample size n increases, the prior should become less influential. Indeed, when $n = 5000$ and Beta$(1,1)$ priors were used, the SE of Imp10 (0.0192) was slightly less than that of Pro10 (0.0193).

6 IPW Complete Case Analysis

Now consider the second use of IPW described in Section 1, i.e. the estimation of a population mean outcome when this outcome can be missing. This problem is analogous to that of estimating an average treatment effect. Let D denote the outcome and θ denote the population mean of D. Let $T = 1$ if D is observed; $T = 0$ otherwise. Let $\boldsymbol{X}$ be a vector of predictors of T, and $D_{\text{obs}} = TD$. The earlier results for a treatment difference imply that if the proper imputation procedure is used to impute missing values of $\boldsymbol{X}$, then

$$\frac{1}{nM}\sum_{m=1}^{M}\sum_{i=1}^{n}\frac{T_i D_{1i}}{\pi(\boldsymbol{X}_i^{*(m)}; \hat{\boldsymbol{\alpha}}^{(m)})}$$

is a consistent estimator of θ when $M = \infty$, provided that $T \perp\!\!\!\perp D \mid \boldsymbol{X}$ and $(A2)$ and $(A3)$ are true. This estimator comes from ignoring the second half of equation (4). An analogous estimator for the improper imputation procedure comes from ignoring the second half of equation (5). Note that since $\boldsymbol{X}_{\text{mis}}$ is imputed using D_{obs} and D_{obs} is non-zero only in complete-cases ($T = 1$), it may be desirable to impute $\boldsymbol{X}_{\text{mis}}$ separately in complete-cases (using D_{obs}) and incomplete cases (not using D_{obs}).

Now consider the more general problem of using IPW when fitting a general analysis model to complete cases. Let $\boldsymbol{D}$ and $\boldsymbol{\theta}$ denote the variables and parameters, respectively, in an analysis model of interest. Let $T = 1$ if the individual is a complete case (i.e. $\boldsymbol{D}$ is observed) and $T = 0$ otherwise (i.e. at least one element of $\boldsymbol{D}$ is missing). Let $\boldsymbol{X}$

be a vector of predictors of T, and let $\boldsymbol{D}_{\text{obs}} = T\boldsymbol{D}$. Let $\boldsymbol{Q}_{\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{D})$ denote an individual's contribution to the complete-data estimating equations $\sum_{i=1}^{n} \boldsymbol{Q}_{\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{D}_i) = \boldsymbol{0}$. So, the true value of $\boldsymbol{\theta}$ is the solution of $E\{\boldsymbol{Q}_{\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{D})\} = \boldsymbol{0}$. Let $\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\alpha};\boldsymbol{D}_{\text{obs}},T,\boldsymbol{X}) = T\boldsymbol{Q}_{\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{D})/\pi(\boldsymbol{X};\boldsymbol{\alpha})$ denote an individual's contribution to the IPW estimating equations $\sum_{i=1}^{n} \boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\alpha};\boldsymbol{D}_{\text{obs},i},T_i,\boldsymbol{X}_i) = \boldsymbol{0}$. Assume that $T \perp\!\!\!\perp \boldsymbol{D} \mid \boldsymbol{X}$ and that (A.2) and (A.3) are true. Proper or improper imputation can be used for missing values of $\boldsymbol{X}$. First, consider proper imputation. Let $\hat{\boldsymbol{\theta}}^{(m)}$ denote the solution of estimating equations $\sum_{i=1}^{n} \boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\hat{\boldsymbol{\alpha}}^{(m)};\boldsymbol{D}_{\text{obs},i},T_i,\boldsymbol{X}_i^{*(m)}) = \boldsymbol{0}$. Then $\hat{\boldsymbol{\theta}}_A = M^{-1}\sum_{m=1}^{M} \hat{\boldsymbol{\theta}}^{(m)}$ is a consistent estimator of $\boldsymbol{\theta}$ when $M = \infty$. Now, consider improper imputation. The solution $\hat{\boldsymbol{\theta}}_B$ to estimating equations $(nM)^{-1}\sum_{m=1}^{M}\sum_{i=1}^{n} \boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\hat{\boldsymbol{\alpha}};\boldsymbol{D}_{\text{obs},i},T_i,\boldsymbol{X}_i^{*(m)}) = \boldsymbol{0}$ is a consistent estimator of $\boldsymbol{\theta}$ when $M = \infty$. The variance of $\hat{\boldsymbol{\theta}}_B$ can be estimated using the formula given in Appendix C. As in the special case of estimating a population mean outcome, it may be better to impute $\boldsymbol{X}_{\text{mis}}$ separately in complete cases (using $\boldsymbol{D}_{\text{obs}}$) and incomplete cases (not using $\boldsymbol{D}_{\text{obs}}$). This was done in the analysis described in Section 7.

7 Application to NCDS Data

In this section, we demonstrate the use of IPW to reduce bias in a complete-case analysis. Note that the analysis we present is intended to be illustrative rather than definitive. The NCDS consists of 17638 people born in Britain during one week in 1958. 920 immigrants with the same birth dates were added later. Data were collected at birth, ages 7, 11, 16, 23, 33 and 45. 16334 non-immigrants were still alive and free from type-1 diabetes at age 45, and 8953 (55%) of these participated in a biomedical survey. Data from this biomedical survey have been previously used to investigate the effects of characteristics measured at birth and of adult adiposity (BMI and waist circumference at age 45) on glucose metabolism at age 45 (Thomas et al., 2007). Following Thomas et al. (2007), we classified subjects as having high blood glucose if their glycosylated haemoglobin (A1C) was $> 6\%$ or they had type-2 diabetes. Immigrants and subjects with type 1 diabetes were excluded. After these exclusions, 5673 partipants ('complete cases') had complete data for variables in the analysis model. The complete-case analysis will be valid if the

5673 complete cases are representative of the 16334 non-immigrants still alive and free from type 1 diabetes. Otherwise it may be biased. We use IPW to allow for possible unrepresentativeness of the complete cases.

For the missingness model, i.e. the model for the probability that an individual is a complete case, we used potential predictors of missingness recorded at birth and age 7 that were identified by Atherton et al. (2008) as well as further predictors measured at age 11. All were categorical. They were sex, mother's husband's social class (non-manual / manual III or IV / manual V or no husband), mother leaving school at or before minimum statutory age, breast-feeding $<$ 1 month, short stature at age 7, overweight at age 7, hospitalisation prior to age 7, social care prior to age 7 (all yes/no) and housing tenure at age 7 (owned/rented). Maths and reading scores (normal/low) and internalising and externalising hehaviour (normal/intermediate/problem) at ages 7 and 11 were also included, as were verbal and non-verbal scores at age 11 (normal/low).

Some missingness predictors were themselves incomplete. Most of this missingness was due to some individuals failing to attend the age-7 or 11 visits: 77% of the cohort attended both visits; 13% just the age-7 visit; 4% just the age-11 visit; 6% attended neither visit. The proportion of missing values in each missingness predictor among those attending the visit at which the missingness predictor should have been measured ranged from 0 to 13%. All missing values in missingness predictors were multiply imputed using the ice function (Royston, 2005) in STATA. This implements the chained equations (or 'fully-conditional specification') MI method, which is a proper imputation procedure. Ten imputed datasets were created (i.e. $M = 10$). Imputation was carried out separately in complete and incomplete cases (i.e. complete and incomplete for the variables in the analysis model). For the complete cases, the variables in the analysis model were also used for the imputation.

Two missingness models were used: one with just the missingness predictors described above, and one with an additional categorical variable describing the pattern of missingness in the missingness predictors. The first of these corresponds to not including R in the model; the second, to including it. As the main cause of missingness in the predictors was the failure of some individuals to attend the age-7 and 11 visits, the additional categorical variable we used in the second model was visit attendance: equal to 1 if both visits were attended; 2 if only age-7 visit was attended; 3 if only age-11 visit; and 4 if neither visit was attended.

When fitting the analysis model to each multiply imputed dataset in turn, SEs were estimated using a sandwich estimator that accounts for the weights and the uncertainty in these weights (i.e. the uncertainty in the parameters $\boldsymbol{\alpha}$ of the missingness model). Rubin's Rules were used to combine point estimates and SEs.

For the first missingness model (the model not including visit attendance), the mean weight in the complete cases averaged over the ten imputed datasets was 2.88, the 95th centile was 4.27 and the maximum was 9.25. For the second missingness model (the model including visit attendance), the mean, 95th centile and maximum were 2.89, 4.75 and 33.68, respectively. The greater variability in the second set of weights indicates that visit attendance is a strong predictor of being a complete case for the variables in the analysis model. This is also evident from the estimated odds ratio of being a complete case associated with missing both age-7 and 11 visits relative to attending both visits: 0.21 (95% CI 0.17–0.25). In this application it seems plausible for the following reasons that including visit attendance in the weighting model may reduce bias in the analysis model. First, it seems quite possible that the missingness predictors may not be MAR: e.g. whether or not social care prior to age 7 is observed may depend on social care prior to age 7 even after adjusting for the missingness predictors that are observed. Second, the relation represented by the analysis model, i.e. that between high blood glucose and its predictors, may be different in individuals who attend both age-7 and 11 visits from that in individuals who attend neither, even after adjusting for the missingness predictors.

Table 3 shows results for the analysis model using IPW with the weights from both missingness models. (Unweighted) complete-case estimates are also shown. As can be seen, using IPW with either missingness model does not substantially change the results. The biggest differences are in the ORs for short gestation, pre-eclampsia and smoking during pregnancy. The effects of short gestation and pre-eclampsia have increased slightly when the second missingness model is used. On the other hand, the effect of smoking during pregnancy has increased slightly when the first missingness model is used. As expected, all SEs have increased slightly, especially when the second missingness model is used. No variable except pre-pregnancy BMI has changed from being non-significant to significant or vice versa; pre-pregnancy BMI is on the borderline of significance in all three cases.

These data were also analysed by Thomas et al. (2007) and Seaman and White (2011). Seaman and White (2011) used IPW but dealt with missing $\boldsymbol{X}$ using the missing indicator method. Thomas et al. used essentially a complete-case analysis, but increased the number of complete cases by imputing some of the missing variables in the analysis model. Both sets of authors reached similar conclusions to those reported here.

8 Discussion

We have shown that the MI procedure described by Qu and Lipkovich (2009) that does not use the missingness pattern of $\boldsymbol{X}$ in the PS model yields consistent estimation when $\boldsymbol{X}$ is MAR given observed outcome D_{obs} and treatment T. Including R may induce bias if it is associated with the outcome. However, when $\boldsymbol{X}$ is MNAR given D_{obs} and T, inclusion of R may reduce bias. The decision of whether to include R might reasonably depend on one's beliefs in a particular given application about whether $\boldsymbol{X}$ is approximately MAR given D_{obs} and T, about whether R is likely to be associated with the outcome, and about how useful R is as a predictor of missing $\boldsymbol{X}$.

Two MI procedures have been presented in the current article: proper and improper. The improper procedure has the advantage that an asymptotically unbiased estimator for sampling variance is available. It has the disadvantages that this estimator is quite complicated and has not been implemented in current software, and that a parametric imputation model is required, thus ruling out the chained equations MI approach. The proper imputation procedure is more flexible, but the properties of the Rubin's Rules variance estimator when used in this case are not fully understood. In our simulation, however, we found it gave good coverage. Seaman et al. (2011) also found good performance of the Rubin's Rules variance estimator when it was applied in another situation involving

IPW. Schafer (2003) comments that "although we may find it difficult to prove good performance for [the Rubin's Rules variance estimator when not using the MLE], that does not imply that good performance will not be seen in practice. Experience suggests that Bayesian MI does interact well with a variety of semi- and non-parametric estimation procedures." On this basis, we cautiously recommend that Rubin's Rules can be used with the proper imputation procedure. An alternative method of variance estimator for either MI procedure is bootstrap.

Finally, note that we have treated the situation where adjustment for confounding is done using IPW, but the proper imputation procedure could also be used when adjustment is by stratification or matching on the PS.

Acknowledgements: We thank Chris Power for valuable discussion and assistance in obtaining the NCDS data, Claudia Thomas for preparing the variables we used. The Centre for Longitudinal Studies provided the official NCDS data. S. Seaman and I. White were funded by MRC grants U105260558 and MC_US_A030_0015.

References

Atherton, K., Fuller, E., Shepherd, P., Strachan, D. P., and Power, C. (2008). Loss and representativeness in a biomedical survey at age 45 years: 1958 British birth cohort. *Journal of Epidemiology and Community Health* **62,** 216–223.

Cepeda, M. S., Boston, R., Farrar, J. T., and Strom, B. L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology* **158,** 280–287.

D'Agostino, R. B. and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association* **95,** 749–759.

Hayes, J. R. and Groner, J. I. (2008). Using multiple imputation and propensity scores to test the effect of car seats and seat belt usage on injury severity from trauma registry data. *Journal of Pediatric Surgery* **43,** 924–927.

Horwitz, D. G. and Thompson, D. J. (1958). A generalisation of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47,** 663–685.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* **23,** 2937–2960.

Mattei, A. (2009). Estimating and using propensity scores in presence of missing background data: an application to assess the impact of childbearing on wellbeing. *Statistical Methods and Applications* **18,** 257–273.

Mitra, R. and Reiter, J. P. (2011). Estimating propensity scores with missing covariate data using general location mixture models. *Statistics in Medicine* **30,** 627–641.

Nielsen, S. F. (2003). Proper and improper multiple imputation. *International Statistical Review* **71,** 593–627.

Qu, Y. and Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine* **28,** 1402–1414.

Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87,** 113–124.

Rosenbaum, P. R. and Rubin, R. J. A. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55.

Rotnitzky, A. and Wypij, D. (1994). A note on the bias of estimators with missing data. *Biometrics* **44,** 1163–1170.

Royston, J. P. (2005). Multiple imputation of missing values: Update of ice. *Stata Journal* **5,** 527–536.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica* **57,** 19–35.

Seaman, S. R. and White, I. R. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research* (in press).

Seaman, S. R., White, I. R., Copas, A. J., and Li, L. (2011). Combining multiple imputation and inverse-probability weighting. (submitted).

Song, J., Belin, T. R., Lee, M. B., Gao, X., and Rotheram-Borus, M. J. (2001). Handling baseline differences and missing items in a longitudinal study of HIV risk among runaway youths. *Health Services and Outcomes Research Methodology* **2,** 317–329.

Stürmer, T., Schneeweiss, S., Brookhart, M. A., Rothman, K. J., Avorn, J., and Glynn, R. J. (2005). Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: Nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *American Journal of Epidemiology* **161,** 891–898.

Thomas, C., Hypponen, E., and Power, C. (2007). Prenatal exposures and glucose metabolism in adulthood. *Diabetes Care* **30,** 918–924.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data.* (pp. 30, 206), Springer, New York.

Appendix A

An alternative estimator of treatment difference, $E(D_1) - E(D_0)$, is obtained by replacing equation (4) in the proper MI procedure by

$$\hat{\theta}^{(m)} = \left\{ \sum_{i=1}^{n} \frac{T_i D_{1i}}{\pi_i^{*(m)}} \Big/ \sum_{i=1}^{n} \frac{T_i}{\pi_i^{*(m)}} \right\} - \left\{ \sum_{i=1}^{n} \frac{(1-T_i) D_{0i}}{1-\pi_i^{*(m)}} \Big/ \sum_{i=1}^{n} \frac{1-T_i}{1-\pi_i^{*(m)}} \right\}, \tag{6}$$

where $\pi_i^{*(m)} = \pi(\boldsymbol{X}_i^{*(m)}; \hat{\boldsymbol{\alpha}}^{(m)})$. Another estimator of treatment difference can be obtained by replacing equation (5) in the improper MI procedure by

$$\hat{\theta}_B = \left\{ \sum_{i=1}^{n} \frac{T_i D_{1i}}{\pi_i^{*}} \Big/ \sum_{i=1}^{n} \frac{T_i}{\pi_i^{*}} \right\} - \left\{ \sum_{i=1}^{n} \frac{(1-T_i) D_{0i}}{1-\pi_i^{**}} \Big/ \sum_{i=1}^{n} \frac{1-T_i}{1-\pi_i^{**}} \right\}, \tag{7}$$

where $\pi_i^{*-1} = M^{-1} \sum_{m=1}^{M} \pi(\boldsymbol{X}_i^{*(m)}; \hat{\boldsymbol{\alpha}})^{-1}$ and $(1-\pi_i^{**})^{-1} = M^{-1} \sum_{m=1}^{M} \{1-\pi(\boldsymbol{X}_i^{*(m)}; \hat{\boldsymbol{\alpha}})\}^{-1}$.

To estimate treatment ratio, $E(D_1)/E(D_0)$, replace equations (6) and (7) by

$$\hat{\theta}^{(m)} = \left\{ \sum_{i=1}^{n} \frac{T_i D_{1i}}{\pi_i^{*(m)}} \Bigg/ \sum_{i=1}^{n} \frac{T_i}{\pi_i^{*(m)}} \right\} \Bigg/ \left\{ \sum_{i=1}^{n} \frac{(1-T_i)D_{0i}}{1-\pi_i^{*(m)}} \Bigg/ \sum_{i=1}^{n} \frac{1-T_i}{1-\pi_i^{*(m)}} \right\} \tag{8}$$

$$\text{and} \quad \hat{\theta}_B = \left\{ \sum_{i=1}^{n} \frac{T_i D_{1i}}{\pi_i^{*}} \Bigg/ \sum_{i=1}^{n} \frac{T_i}{\pi_i^{*}} \right\} \Bigg/ \left\{ \sum_{i=1}^{n} \frac{(1-T_i)D_{0i}}{1-\pi_i^{**}} \Bigg/ \sum_{i=1}^{n} \frac{1-T_i}{1-\pi_i^{**}} \right\}, \tag{9}$$

respectively. Appendix B contains a proof of the consistency of these estimators when $M = \infty$. The formula in Appendix C for a consistent variance estimator of $\hat{\theta}_B$ still applies.

An alternative to the estimator given in Section 6 of a population mean outcome when outcomes may be missing is

$$\frac{1}{M} \sum_{m=1}^{M} \left\{ \sum_{i=1}^{n} \frac{T_i D_{1i}}{\pi(\boldsymbol{X}_i^{*(m)}; \hat{\boldsymbol{\alpha}}^{(m)})} \Bigg/ \sum_{i=1}^{n} \frac{T_i}{\pi(\boldsymbol{X}_i^{*(m)}; \hat{\boldsymbol{\alpha}}^{(m)})} \right\}$$

Appendix B

Consider the improper MI procedure of Section 3. Let $\boldsymbol{S}_\psi(\boldsymbol{\psi}; \boldsymbol{W}, \boldsymbol{X}_{\text{mis}}) = \partial \log f(\boldsymbol{X} \mid D_{\text{obs}}, T; \boldsymbol{\psi})/\partial \boldsymbol{\psi}$ and $\boldsymbol{\psi}_0$ be the true value of $\boldsymbol{\psi}$. The MLE, $\hat{\boldsymbol{\psi}}$, of $\boldsymbol{\psi}$ is the solution to observed-data score equations $n^{-1} \sum_{i=1}^{n} \boldsymbol{S}_{\text{obs}\psi}(\boldsymbol{\psi}; \boldsymbol{W}_i) = \boldsymbol{0}$, where $\boldsymbol{S}_{\text{obs}\psi}(\boldsymbol{\psi}; \boldsymbol{W}) = E_{\boldsymbol{X}_{\text{mis}}}[\boldsymbol{S}_\psi(\boldsymbol{\psi}; \boldsymbol{W}, \boldsymbol{X}_{\text{mis}}) \mid \boldsymbol{W}]$. If $(A3)$ is true, $\hat{\boldsymbol{\psi}}$ is a consistent estimator of $\boldsymbol{\psi}$ and

$$E_{\boldsymbol{W},R}[\boldsymbol{S}_{\text{obs}\psi}(\boldsymbol{\psi}_0; \boldsymbol{W})] = \boldsymbol{0}. \tag{10}$$

Let $\bar{\boldsymbol{S}}_\alpha(\boldsymbol{\alpha}, \boldsymbol{\psi}; \boldsymbol{W}) = E_{\boldsymbol{X}_{\text{mis}}^*}[\boldsymbol{S}_\alpha(\boldsymbol{\alpha}; T, \boldsymbol{X}) \mid \boldsymbol{W}]$ and $\bar{S}_\theta(\theta, \boldsymbol{\alpha}, \boldsymbol{\psi}; \boldsymbol{W}) = E_{\boldsymbol{X}_{\text{mis}}^*}[S_\theta(\theta, \boldsymbol{\alpha}; \boldsymbol{W}, \boldsymbol{X}_{\text{mis}}^*) \mid \boldsymbol{W}]$, where $\boldsymbol{X}_{\text{mis}}^*$ is distributed $g(\boldsymbol{X}_{\text{mis}} \mid \boldsymbol{W}; \boldsymbol{\psi})$.

If $(A3)$ is true, the distribution of $\boldsymbol{X}_{\text{mis}}$ given $\boldsymbol{W}$ and R is $g(\boldsymbol{X}_{\text{mis}} \mid \boldsymbol{W}; \boldsymbol{\psi}_0)$. So, $\bar{\boldsymbol{S}}_\alpha(\boldsymbol{\alpha}, \boldsymbol{\psi}_0; \boldsymbol{W}) = E_{\boldsymbol{X}_{\text{mis}}}[\boldsymbol{S}_\alpha(\boldsymbol{\alpha}; T, \boldsymbol{X}) \mid \boldsymbol{W}] = E_{\boldsymbol{X}_{\text{mis}}}[\boldsymbol{S}_\alpha(\boldsymbol{\alpha}; T, \boldsymbol{X}) \mid \boldsymbol{W}, R]$ and $\bar{S}_\theta(\theta, \boldsymbol{\alpha}, \boldsymbol{\psi}_0; \boldsymbol{W}) = E_{\boldsymbol{X}_{\text{mis}}}[S_\theta(\theta, \boldsymbol{\alpha}; \boldsymbol{W}, \boldsymbol{X}_{\text{mis}}) \mid \boldsymbol{W}] = E_{\boldsymbol{X}_{\text{mis}}}[S_\theta(\theta, \boldsymbol{\alpha}; \boldsymbol{W}, \boldsymbol{X}_{\text{mis}}) \mid \boldsymbol{W}, R]$.

Let $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\theta}$ be the solutions to estimating equations $n^{-1} \sum_{i=1}^{n} \bar{\boldsymbol{S}}_\alpha(\tilde{\boldsymbol{\alpha}}, \hat{\boldsymbol{\psi}}; \boldsymbol{W}_i) = \boldsymbol{0}$ and

$n^{-1}\sum_{i=1}^{n}\bar{S}_\theta(\tilde{\theta},\tilde{\boldsymbol{\alpha}},\hat{\boldsymbol{\psi}};\boldsymbol{W}_i)=0$, and let $\boldsymbol{\alpha}_0$ denote the true value of $\boldsymbol{\alpha}$. Then

$$\begin{aligned} E_{\boldsymbol{W},R}[\bar{\boldsymbol{S}}_\alpha(\boldsymbol{\alpha}_0,\psi_0;\boldsymbol{W})] &= E_{\boldsymbol{W},R}\, E_{\boldsymbol{X}_{\text{mis}}}[\boldsymbol{S}_\alpha(\boldsymbol{\alpha}_0;T,\boldsymbol{X}) \mid \boldsymbol{W},R] \\ &= E_{\boldsymbol{W},R,\boldsymbol{X}_{\text{mis}}}[\boldsymbol{S}_\alpha(\boldsymbol{\alpha}_0;T,\boldsymbol{X})] \\ &= E_{\boldsymbol{X},T}[\boldsymbol{S}_\alpha(\boldsymbol{\alpha}_0;T,\boldsymbol{X})] \\ &= \mathbf{0} \qquad (11) \end{aligned}$$

$$\begin{aligned} \text{and} \quad E_{\boldsymbol{W},R}[\bar{S}_\theta(\theta_0,\boldsymbol{\alpha}_0,\psi_0;\boldsymbol{W})] &= E_{\boldsymbol{W},R}\, E_{\boldsymbol{X}_{\text{mis}}}[S_\theta(\theta_0,\boldsymbol{\alpha}_0;\boldsymbol{X},D_{\text{obs}},T) \mid \boldsymbol{W},R] \\ &= E_{\boldsymbol{W},R,\boldsymbol{X}_{\text{mis}}}[S_\theta(\theta_0,\boldsymbol{\alpha}_0;\boldsymbol{X},D_{\text{obs}},T)] \\ &= E_{\boldsymbol{X},D_{\text{obs}},T}[S_\theta(\theta_0,\boldsymbol{\alpha}_0;\boldsymbol{X},D_{\text{obs}},T)] \\ &= 0 \qquad (12) \end{aligned}$$

Lines (11) and (12), respectively, follow because the PS model is correctly specified and because

$$E_{\boldsymbol{X},D_{\text{obs}},T}\left[\frac{TD_1}{\pi(\boldsymbol{X};\boldsymbol{\alpha}_0)}-\frac{(1-T)D_0}{1-\pi(\boldsymbol{X};\boldsymbol{\alpha}_0)}\right]=\theta_0$$

It follows from equations (10), (11) and (12) that, subject to regularity conditions on the missingness and imputation models for $\boldsymbol{X}$ (Tsiatis, 2006), $(\hat{\boldsymbol{\psi}},\tilde{\boldsymbol{\alpha}},\tilde{\theta}) \to (\boldsymbol{\psi}_0,\boldsymbol{\alpha}_0,\theta_0)$ as $n\to\infty$. That is, $\tilde{\theta}$ is consistent.

$\bar{\boldsymbol{S}}_\alpha(\boldsymbol{\alpha},\boldsymbol{\psi};\boldsymbol{W})$ and $\bar{S}_\theta(\theta,\boldsymbol{\alpha},\boldsymbol{\psi};\boldsymbol{W})$ can be estimated by Monte Carlo integration, by sampling M values of $\boldsymbol{X}^*_{\text{mis}}$ from $g(\boldsymbol{X}_{\text{mis}} \mid \boldsymbol{W},\boldsymbol{\psi})$. When $M=\infty$, this Monte Carlo integration is exact and so $\tilde{\theta}=\hat{\theta}_B$ and $\tilde{\boldsymbol{\alpha}}=\hat{\boldsymbol{\alpha}}$.

This improper MI procedure is a special case of the improper MI discussed by Robins and Wang (2000). It follows that $\hat{\theta}_A$ and $\hat{\theta}_B$ are asymptotically ($n\to\infty$) equivalent when $M=\infty$. Moreover, assuming $\hat{\theta}_A$ and $\hat{\theta}_B$ are also consistent when $M<\infty$, $\hat{\theta}_B$ will be asymptotically more efficient than $\hat{\theta}_A$ when $M<\infty$.

After replacing θ by $\boldsymbol{\theta}=(\delta,\theta)$ and $S_\theta(\theta,\boldsymbol{\alpha};\boldsymbol{X},D_{\text{obs}},T)$ in equation (3) by

$$\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\alpha};\boldsymbol{X},D_{\text{obs}},T) \;=\; \left[\begin{array}{c} \frac{1-T}{1-\pi(\boldsymbol{X};\boldsymbol{\alpha})}(D_0-\delta) \\ \frac{1-T}{1-\pi(\boldsymbol{X};\boldsymbol{\alpha})}(D_0-\delta)+\frac{T}{\pi(\boldsymbol{X};\boldsymbol{\alpha})}(D_1-\delta-\theta) \end{array}\right], \qquad (13)$$

the preceding proof shows that equations (6) and (7) yield consistent estimators of treatment difference when $M = \infty$. If $S_\theta(\theta, \boldsymbol{\alpha})$ in equation (3) is replaced by

$$\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\alpha}; \boldsymbol{X}, D_{\text{obs}}, T) \;=\; \begin{bmatrix} \frac{1-T}{1-\pi(\boldsymbol{X};\boldsymbol{\alpha})}(D_0 - \delta) + \frac{T}{\pi(\boldsymbol{X};\boldsymbol{\alpha})}(D_1 - \delta\theta) \\ \frac{T}{\pi(\boldsymbol{X};\boldsymbol{\alpha})}(D_1 - \delta\theta) \end{bmatrix}, \tag{14}$$

then the proof shows that equations (8) and (9) yield consistent estimators of the treatment ratio when $M = \infty$.

Appendix C

Let $\boldsymbol{\beta} = (\boldsymbol{\theta}^T, \boldsymbol{\alpha}^T)^T$ and $\boldsymbol{\beta}_0 = (\boldsymbol{\theta}_0^T, \boldsymbol{\alpha}_0^T)^T$. Assuming $\hat{\boldsymbol{\beta}}$ is consistent and the regularity conditions for Corollary 1 of Robins and Wang (2000), a consistent estimator of the asymptotic variance of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is $\boldsymbol{\tau}^{-1}\boldsymbol{\Omega}(\boldsymbol{\tau}^T)^{-1}$, where $\boldsymbol{\tau}$ and $\boldsymbol{\Omega}$ are given below.

Let $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{W}, \boldsymbol{X}_{\text{mis}}) = (\boldsymbol{S}_\theta^T(\boldsymbol{\beta}; \boldsymbol{W}, \boldsymbol{X}_{\text{mis}}), \boldsymbol{S}_\alpha^T(\boldsymbol{\beta}; \boldsymbol{W}, \boldsymbol{X}_{\text{mis}}))^T$ and let $\bar{\boldsymbol{U}} = \bar{\boldsymbol{U}}(\boldsymbol{\beta}; \boldsymbol{W}) = M^{-1}\sum_{m=1}^{M} \boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{W}, \boldsymbol{X}_{\text{mis}}^{*(m)})$. Let $\boldsymbol{U}_i^{(m)} = \boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{W}_i, \boldsymbol{X}_{\text{mis},i}^{*(m)})$, let $\bar{\boldsymbol{U}}_i = \bar{\boldsymbol{U}}(\boldsymbol{\beta}; \boldsymbol{W}_i)$ and let $\boldsymbol{S}_{\text{obs}\psi i}(\boldsymbol{\psi}) = \boldsymbol{S}_{\text{obs}\psi}(\boldsymbol{\psi}; \boldsymbol{W}_i)$. Let

$$\begin{aligned}
\boldsymbol{\Omega} &= \boldsymbol{\Omega}_C + \boldsymbol{\kappa}\boldsymbol{\Delta}\boldsymbol{\kappa}^T + \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{\kappa}\boldsymbol{D}_i\bar{\boldsymbol{U}}_i^T + \left(\boldsymbol{\kappa}\boldsymbol{D}_i\bar{\boldsymbol{U}}_i^T\right)^T\right], \\
\boldsymbol{\tau} &= -\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial\bar{\boldsymbol{U}}_i}{\partial\boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}, \qquad \boldsymbol{\Omega}_C = \frac{1}{n}\sum_{i=1}^{n}\bar{\boldsymbol{U}}_i\bar{\boldsymbol{U}}_i^T \\
\boldsymbol{\kappa} &= \frac{1}{nM}\sum_{i=1}^{n}\sum_{m=1}^{M}\boldsymbol{U}_i^{(m)}\boldsymbol{S}_{\text{mis}\psi i}^{(m)T}, \qquad \boldsymbol{\Delta} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{D}_i\boldsymbol{D}_i^T, \\
\boldsymbol{D}_i &= -\left\{\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial\boldsymbol{S}_{\text{obs}\psi i}(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}^T}\right|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}\right\}^{-1}\boldsymbol{S}_{\text{obs}\psi i}(\boldsymbol{\psi}), \\
\boldsymbol{S}_{\text{mis}\psi i}^{(m)} &= \partial f(\boldsymbol{X}_i, \boldsymbol{X}_{\text{mis},i}^{*(m)} \mid \boldsymbol{W}_i; \boldsymbol{\psi})/\partial\boldsymbol{\psi}\mid_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} = \boldsymbol{S}_\psi(\hat{\boldsymbol{\psi}}; \boldsymbol{W}_i, \boldsymbol{X}_{\text{mis},i}^{*(m)}) - \boldsymbol{S}_{\text{obs}\psi i}(\hat{\boldsymbol{\psi}}).
\end{aligned}$$

If they are unavailable analytically, $\boldsymbol{S}_{\text{obs}\psi i}(\boldsymbol{\psi})$ and $\partial\boldsymbol{S}_{\text{obs}\psi i}(\boldsymbol{\psi})/\partial\boldsymbol{\psi}^T\big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}$ can be estimated by Monte Carlo integration. Note that $\boldsymbol{S}_{\text{mis}\psi i}^{(m)} = \boldsymbol{0}$ if $\boldsymbol{X}_i$ is observed.

Table 1: Empirical standard errors and 95% coverages of estimators of average treatment effect, θ, for seven methods of handling missing values of X_2 in PS model. Monte Carlo SEs are 0.0013 for SEs when true SE is 0.06, and 0.7% for coverage when true coverage is 95%.

γ_T	γ_D	γ_2	MissI	SepPS	Imp1	Imp10	ImpR10	Pro10	ProR10
			Empirical SE						
MAR									
1	1	0	0.057	0.054	0.061	0.058	0.061	0.058	0.061
1	0	0	0.053	0.050	0.060	0.058	0.061	0.056	0.059
0	1	0	0.051	0.049	0.062	0.060	0.059	0.059	0.059
0	0	0	0.051	0.048	0.062	0.058	0.058	0.057	0.057
MNAR									
1	1	2	0.055	0.052	0.061	0.059	0.067	0.056	0.063
1	0	2	0.052	0.050	0.060	0.057	0.063	0.056	0.061
0	1	2	0.051	0.049	0.072	0.069	0.073	0.062	0.064
0	0	2	0.051	0.048	0.071	0.068	0.069	0.062	0.062
			95% Coverage						
MAR									
1	1	0	93	92	96	96	83	96	87
1	0	0	82	89	95	94	94	95	96
0	1	0	92	93	95	95	93	95	95
0	0	0	92	90	95	93	93	94	94
MNAR									
1	1	2	93	90	87	87	97	88	94
1	0	2	87	92	92	93	97	93	97
0	1	2	93	94	89	89	96	91	96
0	0	2	87	91	92	92	95	93	95

Table 2: Asymptotic biases of estimators of average treatment effect, θ, for four methods of handling missing values of X_2 in PS model

$P(X_2$ miss$)$			Asymptotic Bias			
γ_T	γ_D	γ_2	MissI	SepPS	Imp	ImpR
MAR						
1	1	0	-0.003	-0.024	0.000	-0.062
1	0	0	0.055	0.035	0.000	0.000
0	1	0	0.021	0.014	0.000	-0.017
0	0	0	0.029	0.032	0.000	0.000
MNAR						
1	1	2	-0.016	-0.028	0.039	-0.047
1	0	2	0.039	0.022	0.016	-0.009
0	1	2	0.014	0.003	0.033	-0.019
0	0	2	0.039	0.028	0.004	-0.008

Table 3: log ORs and SEs for predictors of high blood glucose, using CC and IPW. Binary predictors are gestational age < 38 weeks, pre-eclampsia, smoking during pregnancy, pre-pregnancy BMI ≥ 25 Kg/m^2, and manual socio-economic position (SEP) at birth. Ordinal and continuous predictors are birth weight for gestational age (per tertile), BMI at age 45 (per Kg/m^2) and waist circumference at age 45 (per cm). Adjustment was also made for sex and family history of diabetes.

	CC		IPW without visit		IPW with visit	
	log OR	SE	log OR	SE	log OR	SE
Short gestation	0.562	0.244	0.559	0.265	0.637	0.282
Pre-eclampsia	0.645	0.283	0.651	0.290	0.823	0.314
Smoking	0.103	0.160	0.157	0.168	0.106	0.175
Pre-preg BMI	0.332	0.163	0.328	0.172	0.364	0.180
Manual SEP	0.281	0.199	0.310	0.202	0.315	0.208
Birth weight	-0.303	0.097	-0.313	0.100	-0.292	0.106
BMI	0.066	0.028	0.060	0.029	0.059	0.030
Waist size	0.061	0.012	0.061	0.013	0.062	0.013