SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
0(0) 1–18
© The Author(s) 2011
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280210395740
smm.sagepub.com
SAGE

Review of inverse probability weighting for dealing with missing data

Shaun R Seaman and Ian R White

Abstract

The simplest approach to dealing with missing data is to restrict the analysis to complete cases, i.e. individuals with no missing values. This can induce bias, however. Inverse probability weighting (IPW) is a commonly used method to correct this bias. It is also used to adjust for unequal sampling fractions in sample surveys. This article is a review of the use of IPW in epidemiological research. We describe how the bias in the complete-case analysis arises and how IPW can remove it. IPW is compared with multiple imputation (MI) and we explain why, despite MI generally being more efficient, IPW may sometimes be preferred. We discuss the choice of missingness model and methods such as weight truncation, weight stabilisation and augmented IPW. The use of IPW is illustrated on data from the 1958 British Birth Cohort.

Keywords

asymptotic efficiency, doubly robust, model misspecification, propensity score

1 Introduction

Almost all datasets collected for medical or social research are missing some information that was intended to be collected. This complicates their analysis. A commonly used approach is to exclude individuals with missing data. However, estimates obtained from this 'complete-case' (CC) analysis may be biased if the excluded individuals are systematically different from those included. Inverse probability weighting (IPW) is one of several methods that can reduce this bias. In this method, complete cases are weighted by the inverse of their probability of being a complete case.

Another use of IPW is to correct for unequal sampling fractions. When a survey is conducted, if the sample is representative, i.e. everyone is equally likely to be sampled, few or no individuals with rare characteristics will be chosen. Such individuals may be of particular interest. To ensure that an adequate number of them are sampled, sampling weights are used. Each individual in the population is given a sampling weight and the probability that he or she is chosen is proportional to this weight. Sample estimates of population quantities may now, however, be biased, as the sample is

MRC Biostatistics Unit, Institute of Public Health, Forvie Site, Robinson Way, Cambridge, CB2 0SR, UK.

Corresponding author:
Shaun Seaman, MRC Biostatistics Unit, Institute of Public Health, Forvie Site, Robinson Way, Cambridge, CB2 0SR, UK.
Email: shaun.seaman@mrc-bsu.cam.ac.uk

systematically different from the population. Again, IPW can remove this bias. IPW may also be used to deal simultaneously with missing data and unequal sampling fractions.

This article is a review of the implementation of, and advantages and disadvantages of using, IPW to handle missing data in epidemiological research. There are several earlier reviews.[1–4] This article's purpose is to bring together points made in these and other articles and discuss and clarify issues not adequately covered elsewhere. In Section 2, we explain why bias may arise in the CC analysis and how IPW can remove this. Bias does not always arise in the CC analysis; we identify such situations. As stated above, a major use of IPW is with missing data. It is not the only method: multiple imputation (MI) and full-likelihood methods are also commonly used. In Section 3, we compare IPW with these, and in Section 4, we identify reasons why, and situations in which, IPW may be preferred. IPW requires a model for the probability that data are missing. The choice of this model is discussed in Section 5. We also describe, in Section 6, two techniques used to improve efficiency of IPW: weight stabilisation and augmented IPW (AIPW). Section 7 describes an illustration of the application of IPW. There are several other recent applications in the literature.[5–20]

2 Bias in CC analysis and IPW

2.1 Notation

Consider a generalised linear model for regression of scalar outcome Y on covariates $\boldsymbol{X}$. We call this the 'analysis model'. Let Y_i and $\boldsymbol{X}_i$ denote the values of Y and $\boldsymbol{X}$ for individual i ($i = 1, \ldots, n$) and $\boldsymbol{\theta}$ denote the model parameters.

$\boldsymbol{\theta}$ is estimated as the value $\hat{\boldsymbol{\theta}}$ that solves the score equations:

$$\sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\theta}) = \boldsymbol{0} \qquad (1)$$

where $\boldsymbol{U}_i(\boldsymbol{\theta})$ is the first derivative with respect to $\boldsymbol{\theta}$ of the log likelihood function. For example, for (multiple) linear regression $\boldsymbol{U}_i(\boldsymbol{\theta}) = \boldsymbol{X}_i(Y_i - \boldsymbol{\theta}^T \boldsymbol{X}_i)$.

When there are missing data, we say that a variable, or set of variables, is 'fully observed' if its values are observed on all individuals in the sample. An individual is a 'complete case' (or 'complete') if his or her $\boldsymbol{X}$ and Y are observed. We use the term 'observed' to mean completely observed; i.e. a set of variables is 'observed' on an individual only if none of his or her values of those variables is missing.

The analysis model implies a model, called the 'mean model', $E[Y \mid \boldsymbol{X}] = g^{-1}(\boldsymbol{\theta}^T \boldsymbol{X})$, where g is the link function of the generalised linear model. We say the mean model is correctly specified (or 'correct') if there exists a value $\boldsymbol{\theta}_0$ for which $E[Y \mid \boldsymbol{X}] = g^{-1}(\boldsymbol{\theta}_0^T \boldsymbol{X})$ for all values of $\boldsymbol{X}$. Otherwise, it is misspecified.

If the mean model is correct, $E[\boldsymbol{U}(\boldsymbol{\theta}_0) \mid \boldsymbol{X}] = \boldsymbol{0}$ for all $\boldsymbol{X}$. It is this that guarantees that $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ as $n \to \infty$[21]. If the mean model is misspecified, however, there is no $\boldsymbol{\theta}_0$ for which $E[\boldsymbol{U}(\boldsymbol{\theta}_0) \mid \boldsymbol{X}] = \boldsymbol{0}$ for all $\boldsymbol{X}$. In this situation, $\hat{\boldsymbol{\theta}}$ still converges to a particular value, which we call the 'least false' value of $\boldsymbol{\theta}$ and still denote as $\boldsymbol{\theta}_0$, but $\boldsymbol{\theta}_0$ depends on the distribution of $\boldsymbol{X}$ in the population; it is the solution of $E[\boldsymbol{U}(\boldsymbol{\theta}_0)] = \boldsymbol{0}$.

Although we focus on generalised linear analysis models, this article applies more generally to other analysis models fitted by maximum likelihood or quasi-likelihood. In general, $U_i(\boldsymbol{\theta})$ is individual i's score or quasi-score contribution and we say 'the mean model is correct' if there exists a value $\boldsymbol{\theta}_0$ for which $E[U(\boldsymbol{\theta}_0) \mid \boldsymbol{X}] = \boldsymbol{0}$ for all $\boldsymbol{X}$.

2.2 Bias in CC and IPW analyses

One common method of estimating $\boldsymbol{\theta}$ when data are missing is to include in the analysis only complete cases. This is known as a CC analysis. Let $R_i = 1$ if Y_i and $\boldsymbol{X}_i$ are observed, and $R_i = 0$ otherwise. The CC analysis, then, involves solving the CC score equations: $\sum_{i=1}^{n} R_i \, \boldsymbol{U}_i(\boldsymbol{\theta}) = \boldsymbol{0}$.

We now describe two situations where the CC estimator is consistent. First, when completeness of a case is independent of Y and $\boldsymbol{X}$, i.e. $P(R=1 \mid \boldsymbol{X}, Y) = P(R=1)$, the CC estimator is consistent, because the complete cases are a representative subsample of the original sample. Second, when the mean model is correct and completeness of a case is independent of Y given $\boldsymbol{X}$, i.e. $P(R=1 \mid \boldsymbol{X}, Y) = P(R=1 \mid \boldsymbol{X})$, the CC estimator is consistent.[3] This is because if completeness is independent of Y given $\boldsymbol{X}$, each complete case is a random sample of size one from the population of individuals with the same $\boldsymbol{X}$ value. As the mean model is correct, $E[\boldsymbol{U}(\boldsymbol{\theta}_0) \mid \boldsymbol{X}] = \boldsymbol{0}$ for all values of $\boldsymbol{X}$, and so $E[\boldsymbol{U}(\boldsymbol{\theta}_0) \mid \boldsymbol{X}] = \boldsymbol{0}$ for each complete case. More mathematically, $E[\boldsymbol{U}(\boldsymbol{\theta}_0) \mid \boldsymbol{X}, R=1] = E[\boldsymbol{U}(\boldsymbol{\theta}_0) \mid \boldsymbol{X}] = \boldsymbol{0}$, with the first equality due to the independence of Y and R given $\boldsymbol{X}$ and the second due to the mean model being correct. In situations other than these two, the CC estimator will generally not be consistent. Note that $P(R=1 \mid \boldsymbol{X}, Y) = P(R=1 \mid \boldsymbol{X})$ corresponds to missing at random (MAR)[22] if $\boldsymbol{X}$ is fully observed, but not if Y is fully observed.

Figure 1 illustrates some of these points for linear regression of Y on univariate X. In Figure 1(a), completeness is independent of Y given X and the mean model is correct. When X is less than a threshold (shown by dotted line), X and Y are both observed. When X is greater than the threshold, at least one of X and Y is nearly always missing. Notice that regressing Y on X in the complete cases gives consistent estimation of the intercept and slope: the CC analysis is valid. In Figure 1(b), Y is fully observed and whether X is observed is independent of X given Y. When Y is less than a threshold, X is observed. When Y is greater than the threshold, X is nearly always missing. So, completeness is independent of X given Y. The mean model is correct. Regressing Y on X in the complete cases gives a negatively biased estimate of the slope parameter: the CC analysis is invalid. Figure 1(c) shows an example where X is fully observed and whether Y is observed is independent of Y given X. So, completeness is independent of Y given X. However, the CC analysis is biased because the mean model is misspecified. In this case, the true relation between Y and X is quadratic; so, the least false slope for the linear regression depends on the distribution of X in the population. As the average value of X in the complete cases is less than the average value in the whole sample, using just complete cases underestimates the slope. In Figure 1(d), there is an additional unmodelled covariate Z independent of X, and the relation between X and Y varies with Z. X is fully observed and the mean model is correct (because $E[\boldsymbol{Y} \mid \boldsymbol{X}]$ is linear in X, with slope equal to an average of the slopes when $Z=0$ and $Z=1$), but completeness is not independent of Y given X. When $Z=0$, Y is always observed, but when $Z=1$, Y is nearly always missing. As most complete cases have $Z=0$, the slope estimated from them is biased towards the slope in the $Z=0$ set.

In the IPW approach, the analysis model is also fitted only to complete cases, but more weight is given to some complete cases than others. That is, the estimator $\hat{\boldsymbol{\theta}}$ is the solution of the IPW score equations:

$$\sum_{i=1}^{n} R_i \, w_i \, \boldsymbol{U}_i(\boldsymbol{\theta}) = \boldsymbol{0} \tag{2}$$

where w_i is the weight given to individual i. Note that if $w_i = 1$ for all i, the IPW and CC score equations are the same. The weight w_i used is the inverse of the probability that individual i is a complete case.

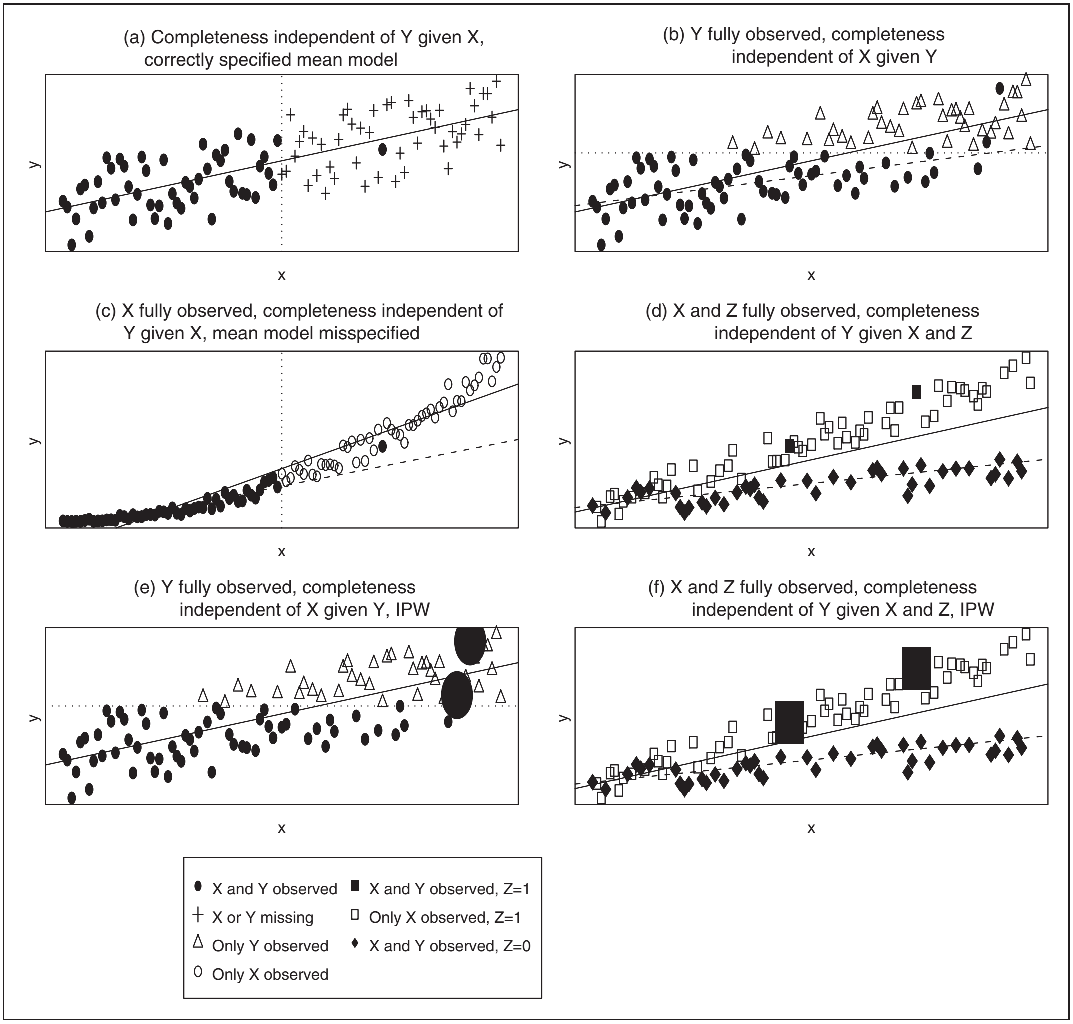

Figure 1. Linear regressions on *Y* on *X*. Solid line represents true population regression slope; broken line represents estimated regression slope

Figure 1(e) illustrates how the bias in the CC analysis is corrected by weighting. The data are the same as those in Figure 1(b). We saw earlier that the unweighted (i.e. CC) analysis is inconsistent in this case. The probability that an individual with Y value below the threshold will be complete is one; so, these individuals' weights equal 1. Individuals with Y above the threshold have a low probability of being complete, and so receive large weights. Two complete cases receive these large weights (represented by large circles in the figure). Using weighting gives more influence to these two individuals and so increases the slope estimate. In fact, these two individuals are effectively representing both themselves and all the individuals with missing data, with the result that the bias in the unweighted analysis is corrected.

When, as is usual, w_i is unknown, it can be estimated. Data available for this are the observed values of $\boldsymbol{X}$, Y and $\boldsymbol{Z}$, where $\boldsymbol{Z}$ represents any further variables measured but not used in the analysis model. Commonly, a logistic regression model is fitted with outcome R and predictors taken from the set $\{\boldsymbol{X}, Y, \boldsymbol{Z}\}$. This is the 'missingness model'. Weights w_i are then the inverse of the fitted probabilities of being complete.

Let $\boldsymbol{H}$ denote the predictors in the missingness model. In Section 5, we discuss in detail the choice of $\boldsymbol{H}$. In brief, the aim is to include sufficient variables such that

$$P(R = 1 \mid \boldsymbol{X}, Y, \boldsymbol{H}) = P(R = 1 \mid \boldsymbol{H}) \tag{3}$$

is a credible assumption. Figure 1(d) and (f) illustrates this. Although completeness is not independent of Y given X, it is independent of Y and X given Z, i.e. $P(R=1 \mid X, Y, Z)=P(R=1 \mid Z)$. So, provided Z is fully observed, IPW can be used with $H=Z$. Figure 1(f) shows the two individuals with $Z=1$ receiving large weights (represented by large circles). These two individuals then have more influence in the linear regression, and so the slope estimate is increased.

Note that in Figure 1(d), the mean model is correct despite ignoring Z because the analysis model is linear. In other analysis models, e.g. logistic regression, ignoring a covariate in a correctly specified model would make the mean model misspecified. However, IPW could still be needed to unbiasedly estimate the least false value of $\boldsymbol{\theta}$.

2.3 Non-fully observed predictors of missingness

The simplest missingness model uses only fully observed predictors. This ensures that all the data needed to fit the missingness model and to estimate individuals' weights are observed. However, when data are 'monotone missing', it is straightforward to include non-fully observed predictors too. A vector $\boldsymbol{V}=(\boldsymbol{V}^1, \ldots, \boldsymbol{V}^K)$ of variables measured on each of a set of individuals is monotone missing if, for each $k=2, \ldots, K$, $\boldsymbol{V}^k$ is only observed on an individual if $\boldsymbol{V}^{k-1}$ is also observed.

Data from a survey can be monotone missing if individuals are only asked certain questions if they reply to all earlier ones. More often, monotone missingness can arise in a cohort study with dropout. Let $\boldsymbol{V}^k$ denote a set of variables measured at visit k ($k=2, \ldots, K$). If missing values are due only to missed visits and individuals who miss a visit miss all subsequent visits, then $\boldsymbol{V}^k$ will only be observed if $\boldsymbol{V}^1, \ldots, \boldsymbol{V}^{k-1}$ are also observed. Suppose that everyone attends visit 1 (so $\boldsymbol{V}^1$ is fully observed) and that $\boldsymbol{V}^K$ includes Y or at least part of $\boldsymbol{X}$ (so complete cases are those individuals who attend all K visits). The conditional probability of being a complete case given $\boldsymbol{V}$ is

$$P(R = 1 \mid \boldsymbol{V}) = P(\{\boldsymbol{V}^K \text{obs}'\text{d}\} \mid \boldsymbol{V}) = \prod_{k=2}^{K} P(\{\boldsymbol{V}^k \text{obs}'\text{d}\} \mid \{\boldsymbol{V}^{k-1} \text{obs}'\text{d}\}, \boldsymbol{V}) \tag{4}$$

('obs'd' means 'observed'). If we assume that the probability that an individual who attends visit $k-1$ also attends visit k does not depend on variables measured at visit k or later given the variables measured earlier, i.e.

$$P(\{\boldsymbol{V}^k \text{obs}'\text{d}\} \mid \{\boldsymbol{V}^{k-1} \text{obs}'\text{d}\}, \boldsymbol{V}) = P(\{\boldsymbol{V}^k \text{obs}'\text{d}\} \mid \{\boldsymbol{V}^{k-1} \text{obs}'\text{d}\}, \boldsymbol{V}_1, \ldots, \boldsymbol{V}^{k-1}), \tag{5}$$

then $P(\{\boldsymbol{V}^k \text{obs'd}\} \mid \{\boldsymbol{V}^{k-1} \text{obs'd}\}, \boldsymbol{V})$ can be estimated by fitting a model (e.g. logistic regression) involving predictors $\boldsymbol{V}^1, \ldots, \boldsymbol{V}^{k-1}$ to only those individuals who attend visit $k-1$. $P(R=1 \mid \boldsymbol{V})$ can then be calculated for complete cases using equation (4). Note that none of these involves missing values.

Robins and Gill[23] describe the Markov randomised monotone missingness (RMM) model, which allows non-fully observed predictors of missingness to be used even when they are not monotone missing. However, the number of parameters in such a missingness model is typically of order $Q2^Q$, where Q is the number of non-fully observed predictors, limiting its practical use to situations where Q is small. The Markov RMM is also more complicated to fit.

2.4 Sampling fractions

Hitherto, we have concentrated on bias resulting from missing data and correction of this using IPW. However, the theory expounded in this section also applies to the situation of unequal sampling fractions in surveys if we redefine R_i and n: $R_i = 1$ if individual i is in the sample and $R_i = 0$ otherwise; n is the number of individuals in the population. The CC estimator will be consistent if the probability that an individual is included in the sample is independent of his or her values of $\boldsymbol{X}$ and Y (which is so, e.g., for a simple random sample) or if both the inclusion probability depends only on $\boldsymbol{X}$ and the mean model is correct. IPW means weighting each individual in the sample by the inverse of his or her known probability of being sampled.

Missing data can, of course, arise in surveys employing unequal sampling fractions. In this case, $R_i = 1$ if individual i is sampled and a complete case. Weight w_i for individual i is then the product of two weights: the inverse of the probability of being sampled and the inverse of the probability of being complete.

3 Comparing IPW and MI

In the IPW approach, a missingness model is specified, i.e. a model for the probability that an individual is a complete case. In the MI approach, an imputation model is specified instead, i.e. a model for the distribution of the missing values given the observed data. The missing values are then replaced by values randomly generated from this model to create a complete set of data. This is done repeatedly, so that several imputed datasets are generated. The analysis model is fitted to each of these in turn and the estimated parameters averaged over the datasets. One can perform MI just on the data ($\boldsymbol{X}$, Y) used in the analysis model, but typically auxiliary variables $\boldsymbol{Z}$ are also used.

A third approach is full likelihood. We shall not discuss this here, except to note that MI can be viewed as a computationally easier way to achieve what the full-likelihood approach does. When the imputation and analysis models are the same, full likelihood and MI with a very large number of imputations (i.e. completed datasets) yield the same estimates[24] (with a smaller number, MI is subject to stochastic noise). When the imputation model uses auxiliary variables, MI is generally easier to implement than full likelihood.[24] The similarity between the two methods means that comments in this article about properties of MI generally apply also to full-likelihood methods.

The fundamental difference between IPW and MI is that IPW needs a model for the probability that an individual is a complete case, whereas MI needs a model for the distribution of the missing data given the observed data. A more subtle difference (Appendix 2) is the assumptions made about the conditional independence of the missingness pattern and missing values, given observed values.

MI has two main advantages over IPW. First, unless ($\boldsymbol{X}$, Y, $\boldsymbol{Z}$) is monotone missing or the more complicated Markov RMM is used, the missingness model of IPW can only use fully observed variables. Second, as discussed below, MI is generally more efficient.

Asymptotic equivalence of IPW and MI (as sample size and number of imputations tend to infinity) has been shown in the special case where ($\boldsymbol{X}$, Y) are MAR, the fully observed variables are categorical and the imputation and missingness models are both saturated. If either or both the missingness and imputation models are simplified by removing any interactions or main effects, and assuming that they remain correctly specified when this is done, MI remains at least as (and typically more) asymptotically efficient than IPW. See Appendix 1 for technical details.

Simulation studies have tended to show that the above observations about the asymptotic relative efficiency of IPW and MI extend to finite samples and to continuous covariates and outcomes.[4,25–27] That is, IPW with a correctly specified missingness model is generally less efficient than MI with a correctly specified imputation model. There are two reasons for this. First, IPW uses only complete cases, whereas MI can use information from individuals with partially missing data. This use of information is most obvious for individuals with observed Y, but even those with missing Y can contribute to an analysis using MI: they provide information about the joint distribution of $\boldsymbol{X}$, which reduces uncertainty when imputing missing $\boldsymbol{X}$ values in individuals with observed Y. Furthermore, an imputation model can use auxiliary variables $\boldsymbol{Z}$ to reduce the uncertainty in missing values of Y (and $\boldsymbol{X}$). Second, unless $\boldsymbol{X}$ is fully observed, an imputation model must (explicitly or implicitly) assume something about the distribution of $\boldsymbol{X}$, an assumption which is not made in the IPW approach and which brings greater efficiency, provided it is correct.

4 Why use IPW instead of MI?

If done carefully, MI is a versatile, powerful and reliable technique for handling missing data. We stress that none of the arguments below are intended to suggest the contrary. Given MI is generally more efficient than IPW and IPW is inhibited when predictors of missingness are not fully observed, what reasons could there be to use IPW?

One reason possibly relevant to some researchers is that IPW arguably requires less technical sophistication than MI. IPW is easier to understand and explain to collaborators, and it is easier to carry out an analysis with IPW than with MI, although the availability of statistical software[28] has made MI more accessible than it was. Furthermore, there is a danger that a less experienced user may apply MI incorrectly. In particular, the imputation model should accommodate structure present in the analysis model.[29] For example, the latter may contain interactions, quadratic terms or random effects. These should be included in the imputation model. This implies that imputation be done separately for each analysis or that all intended analyses be specified before doing the imputation. Of the two most common MI methods, one (the multivariate normal approach[22]) does not allow inclusion of such interactions and non-linear terms without defying distributional assumptions[30] and one (chained equations[31]) requires the user to specify that they be included. Random effects can be handled in the multivariate normal approach;[32] it is not clear how to do this with chained equations. Deficits in the imputation model are unlikely to cause substantial bias if few data are missing, but bias may increase as the amount of missing data increases. By not using imputation, IPW avoids these problems. Just like an imputation model, an IPW missingness model could be misspecified, but, arguably, specifying the latter correctly is an easier task.

A situation where MI may be done incorrectly is that of fully observed covariates, MAR outcome and misspecified mean model. Here, less obvious interactions may be needed in the imputation. If the mean model is correct, the CC analysis is valid (Figure 1(a)) – although MI can be more efficient if auxiliary variables are used. However, if the mean model is misspecified, the CC analysis will be inconsistent (Figure 1(d)) and IPW or MI is needed. A typical default imputation model will be inadequate here, and the user must specify that the imputation be different in individuals with different probabilities of being complete, e.g. by including extra interaction terms (in Figure 1(d) an X–Z interaction). In Figure 1(d), the model for R requires no interactions.

Another argument advanced[4] in favour of IPW concerns the situation where the distribution of missingness predictors in complete cases is quite different from in incomplete cases (as in Figure 1(e)). The weights will then be highly variable, as complete cases whose missingness predictor values are closer to the centre of the distribution of values in the incomplete cases can receive very large weights. This leads to large standard errors (SEs). MI can produce smaller SEs, but MI implicitly involves extrapolation from complete to incomplete cases: the joint distribution of variables in the complete cases is used to impute missing values in the incomplete cases. Crucially, there is little scope for assessing whether this implicit extrapolation is justified, as the data on incomplete cases necessary for such an assessment have, by definition, not been observed. In this situation, it is argued that the large SEs produced by IPW reflect genuine uncertainty that MI eliminates by making an untestable assumption. Unlike the imputation model, the adequacy of the missingness model can be assessed, as the latter is a model for R, which is always observed. However, although one can assess the fit of the missingness model, Equation (3) is an untestable assumption, like the MAR assumption of MI.

A situation where one may prefer IPW is when individuals with missing data tend to have missing values on many, rather than just one or two, variables. This could arise in a longitudinal study where several variables are measured at each visit. If an individual attends a visit, all variables to be measured at this visit are recorded; if not, all are missing. It could also arise in a survey if subjects tend to answer all of a block of related questions or none of them. In the extreme, missing data may be due only to non-participation: participants have complete data; non-participants have none. If many variables are imputed on some individuals and most other individuals are complete, one may feel uneasy using MI. Imputation in this situation relies on a model for the joint distribution of many variables, a model which if misspecified could cause bias in estimates of interest, and individuals with many missing variables may not provide much information anyway. In particular, it can be difficult to correctly specify an imputation model for a large number of categorical variables. IPW, on the other hand, needs only a model for R. In such situations, it may be worth combining IPW and MI, imputing missing values in individuals with almost complete data and using IPW to adjust for the exclusion of individuals with more missing data.[7,16,18,19,33,34] This approach offers some of the efficiency advantages of MI while minimising its dangers.

5 Building the missingness model

In this section, we discuss the choice of missingness model in IPW. For IPW to remove bias caused by missing data, it is necessary that, first, the predictors $\boldsymbol{H}$ be chosen so that Equation (3) is true and, second, that the relation between these and the probability of being a complete case is correctly modelled. To begin, we consider which predictors to include, assuming (unrealistically) that we know how to model the relation between predictors and the probability of being a complete case.

5.1 Choice of variables

One might expect that a variable should be included if and only if it influences the probability that $R=1$. This is not necessarily the case, however. First, some variables that independently predict the probability that $R=1$ are best not included. If Equation (3) holds and $\boldsymbol{H}=(\boldsymbol{H}^a, \boldsymbol{H}^b)$, where $\boldsymbol{H}^b$ is independent of $\boldsymbol{X}$ and Y given $\boldsymbol{H}^a$, then Equation (3) will still hold if $\boldsymbol{H}$ is replaced by $\boldsymbol{H}^a$, i.e. $\boldsymbol{H}^b$ is not needed. Including $\boldsymbol{H}^b$ will cause two individuals with the same $\boldsymbol{H}^a$ but different $\boldsymbol{H}^b$ values to be assigned different weights. As $\boldsymbol{H}^b$ is independent of $\boldsymbol{X}$ and Y given $\boldsymbol{H}^a$, the two individuals have the same distribution of ($\boldsymbol{X}$, Y), and hence there is no need to weight them differently. Doing so will only increase the variability of the weights and hence the SEs of the parameter estimates for the analysis model.

Second, in addition to including variables that predict missingness, efficiency can be increased by including variables that do not predict missingness but are associated with $\boldsymbol{X}$ and Y.[35] This is because, although such variables are not associated with missingness over a large number of samples, the random nature of the missingness means they will tend to be weakly associated with missingness in the single sample collected. The simultaneous association in the sample of these variables with both missingness and the values of the variables in the analysis model will be exploited if they are included in the missingness model. This is related to post-stratification in survey sampling.[36]

Whether the addition of a variable, H^c say, to a missingness model already satisfying Equation (3) will increase or reduce efficiency depends on how strongly H^c is associated with R given the variables $\boldsymbol{H}$ already in the missingness model, how strongly H^c is associated with Y given ($\boldsymbol{X}$, $\boldsymbol{H}$) (and/or associated with $\boldsymbol{X}$ given $\boldsymbol{H}$ if the mean model is misspecified), and how large the sample is. For a variable H^c only weakly associated (or unassociated) with R given $\boldsymbol{H}$ but strongly associated with Y given ($\boldsymbol{X}$, $\boldsymbol{H}$) (or strongly associated with $\boldsymbol{X}$ given $\boldsymbol{H}$ if the mean model is misspecified), efficiency is likely to be increased. However, if the former association is strong and the latter weak (or non-existent), efficiency will probably decrease. In fact, asymptotically (as sample size $\rightarrow \infty$) the addition of variables beyond that necessary to satisfy equation (3) never reduces efficiency.[35] So, the larger the sample, the stronger must be the association with R and the weaker must be the association with Y (or $\boldsymbol{X}$ if the mean model is misspecified) before adding H^c reduces efficiency.

There is a further consideration: with a finite sample, the addition of more and more variables will ultimately lead to a fitted probability of zero for at least one incomplete case. This zero indicates that $\boldsymbol{H}$ is so informative about missingness that there are no complete cases with $\boldsymbol{H}$ value similar to that of the incomplete case, and so the incomplete case is not being represented in the IPW equations by the upweighting of data on complete cases. The IPW equations (2) are now undefined, because $R_i w_i = 0/0$ for this individual, and the IPW method has failed. Categorical variables with rare categories can be particularly problematic, as it is quite likely that, just by chance, no complete cases will be observed in a rare category. Again, the larger the sample size, the more variables can typically be included in the missingness model before a zero fitted probability is obtained.

In conclusion, to minimise bias, the missingness model should contain enough covariates to make the assumption of Equation (3) plausible, but not necessarily all variables associated with missingness, and using further covariates associated with the variables in the analysis model may increase efficiency. Collins et al.[24] reach a similar conclusion when discussing which auxiliary variables to include in MI. This is also in agreement with findings about using propensity scores to adjust for confounding.[37,38]

5.2 Model specification

The arguments in this section so far have been based on the assumption that one knows how to model the relation between predictors $\boldsymbol{H}$ and probability of missingness. In reality, unless the predictors are all categorical and the sample is large enough to allow a saturated model, this will not be the case. A logistic regression model is often used, but this is for reasons of mathematical convenience: there is usually no particular reason to believe it is correctly specified. Even if a logistic regression form is correct, interaction terms and/or transformations of continuous predictors may be needed to make it so. Apart from the problem of IPW not being guaranteed to remove bias when the missingness model is misspecified, a further problem can arise: that of unstable weights. A misspecified model can yield very small fitted probabilities (and hence very large weights) for some individuals, not because those individuals genuinely have very small probabilities of being complete but because these probabilities are incorrectly estimated. This could be because linearity assumptions about continuous variables are violated and/or true interactions are missing from the model. The estimation of the analysis model is then dominated by a few very large weights, meaning a huge reduction in effective sample size. When large weights arise, one should question whether it is plausible that the variables available really are so strongly predictive of missingness or whether it is more likely that the model is misspecified. Kang and Schafer[39] suggest the latter is more common. One way to detect poor fit of the missingness model is the Hosmer–Lemeshow[40] test. Another, suggested by Kang and Schafer[39] is Hinkley's[41] method for testing the fit of a logistic regression. In this approach, the logit of the fitted probabilities are first calculated from the original missingness model and then the model is refitted including as an additional covariate the square of the logit fitted probability. If this extra term is significant, it indicates that the missingness model does not fit the data in the tails and hence that the large weights may be poorly estimated.

5.3 Handling large weights

Methods that have been proposed to deal with large weights are weight truncation, semi-parametric modelling with logistic regression and models other than logistic regression. In weight truncation, a maximum weight is chosen and all weights greater than this are set equal to it. If the missingness model is correctly specified and the large weights arise because the predictors of missingness are highly informative, truncation may re-introduce some of the bias IPW was used to eliminate. However, when large weights are likely due to model misspecification, truncating them is a reasonable measure. As the choice of maximum value is arbitrary, it is important to vary it in order to verify that parameter estimates for the analysis model are not overly sensitive to this value, i.e. that the substantive conclusions of the analysis do not change.

Wang et al.[42] propose semi-parametric modelling of the effect of the predictors on probability of missingness. They conclude, however, that this approach is limited to situations where the number of predictors is small to moderate. Kang and Schafer[39] suggest using robit regression[43] in place of logistic regression. Robit regression replaces the logistic link by the distribution function of a Student-t distribution with ν degrees of freedom (Kang and Schafer recommend $\nu = 4$). This distribution has heavier tails than the logit link, making robit regression more robust to outliers and less prone to producing very small weights when the model is misspecified. Similarly, Ridgeway and McCaffrey[44] propose using a generalised boosted model, Folsom and Witt[45] propose using constrained logistic regression, and Cao et al.[46] propose an enhanced logistic regression model that contains ordinary logistic regression as a special case. A cross-validation approach originally suggested in the context of propensity scores[47,48] might also be of use. Unfortunately, robit regression, generalised boosted models and constrained logistic regression are not routinely available in most statistical software, and the cross-validation approach is computationally intensive and may select a simpler model than is optimal.

5.4 A proposal

We propose the following as a possible strategy for developing a missingness model.

First, identify which variables available are *a priori* good candidates to predict missingness. Consider removing from this set any that are independent of X and Y and adding any strongly associated with Y. If the weights are to be used for different analyses involving different $\boldsymbol{X}$ and Y, this last step may not be possible.

Second, examine the distribution of any continuous predictors. Use transformations for any with long tails, as extreme values in these untransformed variables may be influential in the model fit and more likely to yield large weights.

Third, fit a missingness model using the full set of identified potential predictors. If robit regression software is available, use it. If there are too many predictors to make including all of them feasible, use forward selection or another variable selection method, e.g. Lasso.[49] If using forward selection, force into the model variables thought highly likely to influence missingness and outcome (e.g. sex, age and social class). Use forward selection to add significant interactions between variables.

Fourth, use Hosmer–Lemeshow and/or Hinkley's method to check model fit.

Fifth, examine the distribution of weights in both complete and incomplete cases. Check there are no zero fitted probabilities in the incomplete cases. If there are zeroes, either one should remove a predictor or predictors or merge categories of categorical predictors, or one should accept that IPW is unsuitable for this application because some individuals genuinely have zero or near-zero probability of being complete. Decide whether the weights are unacceptably unstable. One might do this by comparing SEs of the weighted and unweighted analyses and by looking at whether the sum of, say, the largest 10% of weights in complete cases is greater than half the total sum of weights in complete cases. If weights are unstable, the model may be misspecified. Examine the individuals (complete and incomplete cases) with large weights to try to identify which variables are causing the large weights. Explore whether adding further interaction terms (to allow the joint effect of two or more variables to be less than multiplicative) or transforming continuous variables improves the situation. Consider an alternative to logistic regression, e.g. robit regression. If using truncation, assess sensitivity of results to the choice of maximum weight. Be aware that unstable weights may indicate a major systematic difference between complete and incomplete cases. In this situation, one may choose to accept the large SEs of IPW as representing true uncertainty or may switch to MI but being aware that extrapolation of the data is then being made.[4]

6 Improving efficiency of IPW

In this section, we discuss weight stabilisation and augmentation of the IPW equations, two methods that can improve IPW's efficiency. Weight stabilisation is easy to apply. AIPW is more complicated to apply in standard software and we describe it here only briefly; Vansteelandt et al.[4] provide an accessible introduction.

Weight stabilisation can reduce instability of IPW estimators,[50,51] but only when the mean model is correct. Ordinary IPW uses weights w_i, where w^{-1} equals $P(R=1 \mid \boldsymbol{H})$ or an estimate thereof. Write $\boldsymbol{H}=(\boldsymbol{H}^a, \boldsymbol{H}^b)$, where $\boldsymbol{H}^b$ is independent of Y given $\boldsymbol{X}$. Obviously, $\boldsymbol{X}$ is independent of Y given $\boldsymbol{X}$; so $\boldsymbol{H}^b$ can include any components of $\boldsymbol{X}$ in $\boldsymbol{H}$. In fact, the main use of this method is with $\boldsymbol{H}^b$ taken to be fully observed components of $\boldsymbol{X}$. A regression model for R given $\boldsymbol{H}^b$ is specified and fitted, yielding fitted values $1/w_i^*$. There is no need for this model to be correctly specified. The stabilised IPW score equations are equations (2) with w_i replaced by w_i/w_i^*. If the mean model for Y given $\boldsymbol{X}$ is correct, the stabilised equations yield a consistent estimator of $\boldsymbol{\theta}$. The purpose of doing this is that the distribution of w_i/w_i^* should be less variable than that of w_i. Weight w equals $1/P(R=1 \mid \boldsymbol{H}^a, \boldsymbol{H}^b)$ or an estimate thereof, and w^* is an estimate of $1/P(R=1 \mid \boldsymbol{H}^b)$. Here, w^* can be regarded as a best guess of w when only $\boldsymbol{H}^b$ is available. So, to the extent that the probability that $R=1$ depends on $\boldsymbol{H}^b$, w^* 'tracks' w. Consequently, w/w^* values should tend to be closer to one than are their corresponding

w values. In the extreme case where the probability that $R = 1$ does not depend on $\boldsymbol{H}^a$ given $\boldsymbol{H}^b$, the stabilised weights $w/w^* = 1$ and the stabilised IPW estimator reduces to the CC estimator (the special case of $\boldsymbol{H}^b = \boldsymbol{X}$ and $\boldsymbol{H}^a = Y$ was mentioned in Section 2.2 as a situation where the CC estimator is consistent). In the special situation where Y and some elements of $\boldsymbol{X}$, $\boldsymbol{X}^b$ say, are fully observed and weight $w = w(Y, \boldsymbol{X}^b)$ is a function of $\boldsymbol{H}^a = Y$ and $\boldsymbol{H}^b = \boldsymbol{X}^b$ only, Paik and Wang[52] propose using $w^* = \text{minimum } \{w(y_1, \boldsymbol{X}^b), \ldots, w(y_n, \boldsymbol{X}^b)\}$.

A generalisation of IPW is AIPW. This approach has been developed for the situations where all variables in the analysis model are fully observed except for either Y or one element of $\boldsymbol{X}$. The AIPW score equations are

$$\sum_{i=1}^{n} R_i\, w_i\, \boldsymbol{U}_i(\boldsymbol{\theta}) + (1 - R_i\, w_i)\, \boldsymbol{\phi}_i(\boldsymbol{\theta}) = \boldsymbol{0} \tag{6}$$

where $\boldsymbol{\phi}_i(\boldsymbol{\theta})$ is an estimate of the expectation of $\boldsymbol{U}_i(\boldsymbol{\theta})$ given observed data on individual i. $\boldsymbol{\phi}_i(\boldsymbol{\theta})$ is obtained by fitting an imputation model to the observed data. Thus, AIPW is a hybrid of IPW and imputation. Whereas IPW score equations (2) make use only of complete cases (apart from using incomplete cases to estimate weights), equations (6) use data on incomplete cases and hence offer the prospect of increased efficiency. AIPW possesses the ‘doubly robust’ property: provided that either the missingness or imputation model is correctly specified, the AIPW estimator is consistent. If both models are correctly specified, AIPW will be more efficient than IPW. AIPW is an area of current research. Software for AIPW is currently lacking, but Bang and Robins[53] describe a doubly robust method asymptotically equivalent to AIPW that can be applied using standard software. Estimation of SEs for this method is, however, problematic; Bang and Robins use bootstrap.[54]

7 Application

The 1958 British Birth Cohort consists of 17 638 people born in Britain during 1 week in 1958; 920 immigrants with the same birth dates were added later. Data were collected at ages 0 (birth), 7, 11, 16, 23, 33 and 45. 17 313 subjects were still alive at age 45 and of these, 9377 (54%) participated in a biomedical survey.

We illustrate IPW using data from this biomedical survey to investigate the effects of characteristics measured at or before birth and of adult adiposity (body mass index (BMI) and waist circumference at age 45) on glucose metabolism at 45. Following Thomas et al.[55], we classified subjects as having high blood glucose if their glycosylated haemoglobin (A1C) was >6% or they had type-2 diabetes. Subjects with type-1 diabetes and immigrants were excluded, the latter because their perinatal data were unavailable. After these exclusions, 5673 participants (‘complete cases’) had complete data for variables in the analysis model. If the data are missing completely at random (MCAR), the CC analysis of these 5673 will give valid inference for the population of non-immigrants still alive and free from type-1 diabetes at age 45, but may not otherwise. IPW will give valid inference, provided Equation (3) is true and the missingness model is correctly specified.

For the missingness model, we used potential predictors of missingness recorded at ages 0 and 7 identified by Atherton et al.[56] and 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 7, overweight at 7, hospitalisation prior to 7, social care prior to 7 (all yes/no) and parents’ housing tenure at 7 (owned/rented). Maths and reading scores (normal/low) and internalising and externalising behaviour (normal/intermediate/problem) at 7 and 11 were also included, as were verbal and non-verbal scores at 11 (normal/low).

Some missingness predictors were themselves missing, because not everyone attended the age-7 and age-11 visits, and even for those who did, some variables were not recorded. We dealt with visit missingness by dividing the cohort into four strata: attendees of both visits (77%); just age-7 visit (13%); just age-11 visit (4%); and neither visit (6%). A separate logistic regression for being a complete case was fitted to each stratum, using only the predictors measured at the visits attended by that stratum. The proportion of missing values in each predictor ranged from 0 to 13%, being <6% in all but two predictors. We call the nine predictors with >2% of values missing the 'moderately incomplete' predictors. A missing indicator was included in the missingness model for each of these, i.e. a binary variable taking value 1 if the corresponding predictor is missing. For variables with <2% of values missing, missing values were singly imputed as the modal value. This was done because rare binary predictors could cause instability in the weights.

In reality, it is very unlikely that missingness will depend on predictors measured at age 7 (or age 11) only in those individuals who attend the age-7 (or age-11) visit. Likewise, it is not very plausible that, for an individual who attends both visits, missingness does not depend on a predictor if that predictor is unobserved. Therefore, we do not really believe that Equation (3) holds, especially for individuals who miss the age-7 or age-11 visit. Nevertheless, it is a better approximation to reality than assuming the data are MCAR, and 77% of subjects did attend both visits.

If, as seems likely, missingness depends on predictors whose values are missing, the effect of an observed predictor on missingness may depend on which other predictors are observed: if two predictors are correlated and the first is missing, the effect of the second is altered, as it describes not only its own effect, but also part of the effect of the first. There will therefore be an interaction between one of the predictors and the missing indicator of the other. For this reason, we used forward stepwise selection to enable such interactions to be included in the missingness model. Also included in this stepwise procedure were ordinary interactions between pairs of predictors.

In more detail, we did the following. The missingness model was fitted without interactions to the stratum who attended both age-7 and age-11 visits and were still alive at age 45. Predictors with odds ratio (OR) >1.2 or <1/1.2 ('moderate predictors') were identified. For each pair of moderate predictors, a term was created for their interaction. Also, for each pair of moderate predictors in which at least one (the first, say) was moderately incomplete, a term was created for the interaction between the second predictor and the first predictor's missing indicator. Forward stepwise selection was used, with threshold $p = 0.1$, to select which of these terms to add to the existing missingness model. To avoid weight instability resulting from adding interaction terms with low frequencies, all interaction terms with <100 individuals (0.6%) taking value one were first deleted. The same procedure was repeated independently for the other three strata.

Table 1 shows results for the missingness model for the stratum attending both visits. Only predictors with $p < 0.1$ are shown. As concluded by Atherton et al.,[56] disadvantaged individuals are more likely to be missing. Breastfed <1 month; low social class; mother leaving school earlier; overweight and externalising difficulties at 7; and internalising and externalising difficulties, low maths, low reading and low non-verbal scores at 11 all significantly predicted not being complete. Weights from this model varied from 1.4 to 42.0, with mean 3.5 and 5th and 95th centiles 2.0 and 7.9.

The Hosmer–Lemeshow test and Hinkley tests were used to assess the fit of the four missingness models, one for each stratum. The p-values from these tests were all >0.4. There is therefore no indication of poor fit. We also looked specifically at individuals with the largest weights. For the 293

Table 1. ORs, SEs and *p*-values for predictors of missingness

Predictor	OR	SE	*p*
Mother's husband's social class: III/IV	0.98	0.04	0.001
(baseline=I/II) V/no husband	0.78	0.06	
Mother left school ≤ statutory age	0.87	0.04	0.003
Breastfeeding < 1 month	0.81	0.03	0.000
Overweight at 7	0.86	0.06	0.028
Internalising behav. at 7: intermediate	0.92	0.04	0.092
(baseline=normal) problem	0.89	0.06	
Externalising behav. at 7: intermediate	0.96	0.04	0.017
(baseline=normal) problem	0.82	0.06	
Low maths score at 11	0.78	0.06	0.002
Low reading score at 11	0.85	0.07	0.039
Internalising behav. at 11: intermediate	0.90	0.04	0.020
(baseline=normal) problem	0.87	0.06	
Externalising behav. at 11: intermediate	0.89	0.04	0.011
(baseline=normal) problem	0.86	0.06	
Low non-verbal score at 11	0.77	0.07	0.002
Missing indicators			
Missing hospital before 7	1.44	0.30	0.081
Missing reading score at 7	0.66	0.13	0.032
Interactions			
Class III/IV × low reading at 7	0.66	0.13	0.032
Class III/IV × low non-verbal at 11	1.36	0.24	0.078

Notes: OR > 1 means that the variable increases the probability of being a complete case. Only variables with $p < 0.1$ are shown

individuals with weight >10, the mean predicted probability of being complete was 0.08. This agreed with the fact that 24 (8%) of them actually were complete.

Table 2 shows results of the CC and IPW analyses. Weighting has increased the estimated ORs for five variables. The biggest change in OR is for pre-eclampsia, which increased by 76% of its SE. Confidence intervals (CIs) increased in width, but not substantially. No variable that was non-significant has become significant or vice versa. The findings, then, remain essentially unchanged by IPW, although there is an indication that the effect of pre-eclampsia may be stronger than suggested by the CC analysis. We stress that this analysis is intended as an illustration of IPW, not as a definitive analysis of these data. Thomas et al.[55] analyse the same data using MI, and Seaman et al.[34] combine MI and IPW. Both used all 7518 individuals with complete data on blood glucose, waist size and BMI at age 45, imputing the remaining variables in the analysis model, rather than using only the 5673 complete cases. Their results are similar to those reported here, except that their estimated ORs for short gestation and pre-eclampsia are somewhat lower.

8 Further issues

This article is limited to IPW for a univariate response and has assumed that the probability of being a complete case depends on observed data only. However, IPW can also be used with repeated measures,[57] and sensitivity analyses can allow the probability of being complete to depend on a missing response (i.e. missing not at random).[58] IPW is also used for causal inference.[59]

Table 2. ORs, SEs and 95% CIs for predictors of high blood glucose using CC and IPW

	CC				IPW			
	OR	SE	95% CI		OR	SE	95% CI	
Short gestation	1.75	0.43	1.09	2.83	1.86	0.50	1.10	3.17
Pre-eclampsia	1.91	0.54	1.09	3.32	2.32	0.70	1.28	4.19
Smoking	1.10	0.18	0.81	1.52	1.14	0.19	0.81	1.59
Pre-preg BMI	1.39	0.23	1.01	1.92	1.47	0.26	1.04	2.08
Manual SEP	1.32	0.26	0.90	1.96	1.42	0.29	0.94	2.12
Birth weight	0.74	0.07	0.61	0.89	0.73	0.07	0.60	0.89
BMI	1.07	0.03	1.01	1.13	1.06	0.03	1.00	1.12
Waist size	1.06	0.01	1.04	1.09	1.06	0.01	1.03	1.09

Notes: Binary predictors are gestational age <38 weeks, pre-eclampsia, smoking during pregnancy, pre-pregnancy BMI $\geq$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.

Finally, we mention SEs. Commonly, these are estimated for IPW using a simple sandwich estimator (e.g. in STATA use the 'robust' option). This treats weights as known, whereas they are usually estimated. In fact, the true asymptotic SEs are actually greater when true weights are used than when they are estimated.[35] So, ignoring uncertainty in the weights causes over-estimation of SEs, i.e. conservative inference. A more complicated sandwich estimator avoids this bias, but is not implemented in most software.[35]

In summary, IPW can be valuable in certain settings, but care must be taken to ensure that the missingness model is correctly specified and weights are not unstable.

Acknowledgements

We thank Chris Power and Leah Li for valuable discussions and assistance in obtaining the 1958 data, and Claudia Thomas for preparing the variables we used. The Centre for Longitudinal Studies provided the official 1958 Cohort data. The authors are funded by MRC grants U.1052.00.006 and PHSRN17.

References

1. Höfler M, Pfister H, Lieb R and Wittchen H-U. The use of weights to account for non-response and drop-out. *Soc Psychiatry Psychiatr Epidemiol* 2005; **40**: 291–299.
2. Pfeffermann D. The role of sampling weights when modeling survey data. *Int Stat Rev* 1993; **61**: 317–337.
3. Pfeffermann D. The use of sampling weights for survey data analysis. *Stat Methods Med Res* 1996; **5**: 239–261.
4. Vansteelandt S, Carpenter J and Kenward MG. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology* 2010; **6**: 37–48.
5. Alati R, Najman JM, Kinner SA, Mumun AA, Williams GM, O'Callaghan M and Bor W. Early predictors of adult drinking: A birth cohort study. *Am J Epidemiol* 2005; **162**: 1098–1107.
6. Alonso A, Segui-Gomez M, de Irala J, Sanchez-Villegas A, Beunza JJ and Martinez-Gonzalez MA. Predictors of follow-up and assessment of selection bias from dropouts using inverse probability weighting in a cohort of university graduates. *Eur J Epidemiol* 2006; **21**: 351–358.
7. Caldwell TM, Rodgers B, Clark C, Jefferis BJMH, Stansfeld SA and Power C. Lifecourse socioeconomic predictors of midlife drinking patterns, problems and abstention : Findings from the 1958 British birth cohort study. *Drug Alcohol Depend* 2008; **95**: 269–278.
8. Chikobvu P, Lombard CJ, Flisher AJ, King G, Townsend L and Muller M. Bias in a binary risk behaviour model subject to inconsistent reports and dropout in a South African high school cohort study. *Stat Med* 2009; **28**: 494–509.
9. Clark C, Rodgers B, Caldwell T, Power C and Stansfeld S. Childhood and adulthood psychological ill health as predictors of midlife affective and anxiety disorders. *Arch Gen Psychiatry* 2007; **64**: 668–678.

10. Dufouil C, Brayne C and Clayton D. Analysis of longitudinal studies with death and drop-out: a case study. *Statistics in Medicine* 2004; **23**: 2215–2226.
11. Frederiksen H, Hjelmborg J, Mortensen J, McGue M, VAupel JW and Christensen K. Age trajectories of grip strength: cross-sectional and longitudinal data among 8,342 Danes aged 46 to 102. *Ann Epidemiol* 2006; **16**: 554–562.
12. Gerberich SG, Church TR, McGovern PM, et al. An epidemiological study of the magnitude and consequences of work related violence: the Minnesota Nurse's Study. *Occup Environ Med* 2004; **61**: 495–503.
13. Jones AM, Koolman X and Rice N. Health-related non-response in the British household panel survey and European community household panel: using inverse-probability-weighted estimators in non-linear models. *J R Stat Soc, Ser A* 2006; **169**: 1–27.
14. Mustard CA, Kalcevich C, Frank JW and Boyle M. Childhood and early adult predictors of risk of incident back pain: Ontario child health study 2001 follow-up. *Am J Epidemiol* 2005; **2005**: 779–786.
15. Power C, Li L and Hertzman C. Cognitive development and cortisol patterns in mid-life: Findings from a British birth cohort. *Psychoneuroendocrinology* 2008; **33**: 530–539.
16. Priebe S, Fakhoury W, White I, et al. Characteristics of teams, staff and patients: associations with outcomes of patients in assertive outreach. *Br J Psychiatry* 2004; **185**: 306–311.
17. Rao RS, Sigurdson AJ, Doody MM and Graubard BI. An application of a weighted method to adjust for nonresponse in standardised incidence ratio analysis of cohort studies. *Biom J* 2004; **46**: 579–588.
18. Stansfeld SA, Clark C, Caldwell TM, Rodgers B and Power C. Psychosocial work characteristics and anxiety and depressive disorders in midlife: The effects of prior psychological distress. *Occup Environ Med* 2008; **65**: 634–642.
19. Stansfeld SA, Clark C, Rodgers B, Caldwell TM and Power C. Childhood and adulthood socio-economic position and midlife depressive and anxiety disorders. *Drug Alcohol Depend* 2008; **95**: 269–278.
20. Tate AR, Jones M, Hull L, et al. How many mailouts? Could attempts to increase the response rate in the Iraq war cohort study be counterproductive? *BMC Med Res Methodol* 2007; **7**: 51.
21. Stefanski LA and Boos DD. The calculus of M-estimation. *Am Stat* 2000; **56**: 29–38.
22. Schafer JL. *Analysis of incomplete multivariate data*. London: Chapman and Hall, 1997.
23. Robins JM and Gill RD. Non-response models for the analysis of non-monotone ignorable missing data. *Stat Med* 1997; **16**: 39–56.
24. Collins LM, Schafer JL and Kam C-M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods* 2001; **6**: 330–351.
25. Beunikens C, Sotto C and Molenberghs G. A simulation study comparing weighted estimating equations with multiple imputation based estimating equations. *Comput Stat Data Anal* 2008; **52**: 1533–1548.
26. Carpenter JR, Kenward MG and Vansteelandt S. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *J R Stat Soc Ser A* 2006; **169**: 571–584.
27. Robins J, Sued M, Lei-Gomez Q and Rotnitzky A. Comment: Performance of doubly-robust estimators when "inverse probability" weights are highly variable. *Stat Sci* 2007; **22**: 544–559.
28. Horton NJ and Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Stat* 2007; **61**: 79–90.
29. Kenward M and Carpenter J. Multiple imputation: Current perspectives. *Stat Methods Med Res* 2007; **16**: 199–218.
30. Von Hippel PT. How to impute interactions, squares and other transformed variables. *Sociol Methodol* 2009; **39**: 265–291.
31. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res* 2007; **16**: 219–242.
32. Goldstein H. Handling attrition and non-response in longitudinal data. *Longit Life Course Stud* 2009; **1**: 63–72.
33. Kim JK, Brick JM, Fuller WA and Kalton G. On the bias of the multiple-imputation variance estimator in survey sampling. *J R Stat Soc: Ser B (Stat Methodol)* 2006; **68**: 509–521.
34. Seaman SR, White IR, Copas AJ and Li L. *Combining multiple imputation and inverse-probability weighting*, (submitted).
35. Tsiatis AA. *Semiparametric theory and missing data*. New York: Springer, 2006, pp.206–207.
36. Kalton G and Flores-Cervantes I. Weighting methods. *J Off Stat* 2003; **19**: 81–97.
37. Austin PC. The performance of different propensity-score methods for estimating relative risks. *J Clin Epidemiol* 2008; **61**: 537–545.
38. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J and Sturmer T. Variable selection for propensity score models. *Am J Epidemiol* 2006; **163**: 1149–1156.
39. Kang JDY and Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci* 2007; **22**: 523–539.
40. Hosmer DW and Lemeshow S. *Applied logistic regression*. New York: Wiley, 1989.
41. Hinkley D. Transformation diagnostics for linear models. *Biometrika* 1985; **72**: 487–496.
42. Wang CY, Wang S, Zhao L-P and Ou S-T. Weighted semiparametric estimation in regression analysis with missing covariate data. *J Am Stat Assoc* 1997; **92**: 512–525.
43. Liu C. Robit regression a simple robust alternative to logistic and probit regression. In: Gelman A and Meng X-L (eds) *Applied Bayesian modelling and causal inference from incomplete-data perspectives*. New York: Wiley, 2004, pp.227–238.
44. Ridgeway G and McCaffrey DF. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci* 2007; **22**: 540–543.
45. Folsom RE and Witt MB. Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *Proceedings of the American Statistical Association, Social Statistics Section*. 1991, pp.197–202.
46. Cao W, Tsiasis AA and Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* 2009; **96**: 723–734.
47. Brookhart MA and Van der Laan MJ. A semi-parametric model selection criterion with applications to the marginal structural model. *Comput Stat Data Anal* 2006; **50**: 475–498.
48. Mortimer KM, Neugebauer R, van der Laan M and Tager IB. An application of model-fitting procedures for marginal structural models. *Am J Epidemiol* 2005; **162**: 382–388.
49. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc, Ser B* 1996; **58**: 267–288.

50. Cole SR and Hernan MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2000; **168**: 656–664.
51. Robins JM, Hernan MA and Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**: 550–559.
52. Paik MC and Wang C. Handling missing data by deleting completely observed records. *J Stat Plann Inference* 2009; **139**: 2341–2350.
53. Bang H and Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; **61**: 962–972.
54. Efron B and Tibshirani R. *An introduction to the bootstrap*. New York/London: Chapman and Hall, 1993.
55. Thomas C, Hypponen E and Power C. Prenatal exposures and glucose metabolism in adulthood. *Diabetes Care* 2007; **30**: 918–924.
56. Atherton K, Fuller E, Shepherd P, Strachan DP and Power C. Loss and representativeness in a biomedical survey at age 45 years: 1958 British birth cohort. *J Epidemiol Commun Health* 2008; **62**: 216–223.
57. Robins JM, Rotnitzky A and Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc* 1995; **90**: 106–121.
58. Rotnitzky A, Robins JM and Scharfstein DO. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J Am Stat Assoc* 1998; **93**: 1321–1339.
59. Hogan JW and Lancaster T. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Stat Methods Med Res* 2004; **13**: 17–48.
60. Robins JM and Wang N. Inference for imputation estimators. *Biometrika* 2000; **87**: 113–124.
61. Reilly M and Pepe M. The relationship between hot-deck multiple imputation and weighted likelihood. *Stat Med* 1997; **16**: 5–19.
62. Paik MC. The generalized estimating equations approach when data are not missing completely at random. *J Am Stat Assoc* 1997; **92**: 1320–1329.
63. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581–592.

Appendix 1

Asymptotic equivalence of IPW and MI in special cases

Robins and Wang[60] show asymptotic equivalence of MI and mean-score imputation when $M = \infty$. Reilly and Pepe[61] show, for regression of fully observed Y on fully observed $\boldsymbol{X}$ and MAR $\boldsymbol{Z}$, exact equivalence of mean-score imputation and IPW when Y and $\boldsymbol{X}$ are categorical and imputation and missingness models are saturated. Paik[62] shows, for regression of MAR Y on fully observed $\boldsymbol{X}$, exact equivalence of mean-score imputation and IPW when $\boldsymbol{X}$ is categorical and imputation and missingness models are saturated. It follows that MI and IPW are asymptotically equivalent in these two special cases.

Furthermore, IPW is not more asymptotically efficient when terms are removed from a missingness model, assuming this model remains correct when these terms are removed.[35] MI cannot become more inefficient when unnecessary terms are removed from the imputation model, as these terms merely add noise to the imputed values.[24] It follows that, when a MAR outcome is regressed on fully observed categorical covariates or a fully observed categorical outcome is regressed on fully observed categorical covariates and a MAR covariate, IPW with a correct weighting model cannot be more asymptotically efficient than MI with a correct imputation model.

Appendix 2

Difference between conditional independence assumptions of IPW and MI

MI assumes data are MAR and that the parameters of the data-generating process are distinct from those of the missingness process.[63] When data V are monotone missing, Equation (5) is equivalent to the MAR assumption combined with the three further reasonable assumptions that (1) the missingness mechanism is the same for all individuals in the sample, (2) $(V_1, M_1), \ldots, (V_n, M_n)$ are independent, where M_i is individual i's missingness pattern and (3) data V would still be MAR if the observed values of V were different (but the missingness pattern remained unchanged).[23] Hence, if IPW uses a missingness model based on Equation (5), IPW and MI make the same assumption about the conditional independence of the missingness pattern and missing values given observed values, with IPW additionally making assumptions (1)–(3).

With non-monotone missing data, the missingness model is commonly restricted to fully observed predictors, due to the practical difficulty of fitting a Markov RMM model. The MAR assumption then allows the probability of being complete to depend on variables not in the missingness model.[23] Furthermore, when data are non-monotone missing, there exist MAR mechanisms that conform to assumptions (1)–(3) but cannot be represented by Markov RMM models.[23] For these reasons, it could be argued that when data are non-monotone missing, the conditional independence assumption of IPW is stronger than that of MI. However, whereas the conditional independence assumption of IPW is a statement only about the probability of being complete, MAR is also a statement about the probability of other missingness patterns. Also, MAR mechanisms that cannot be represented by Markov RMM models are difficult to motivate in terms of causality.