Curnow *et al.*
BMC Medical Research Methodology (2024) 24:231
https://doi.org/10.1186/s12874-024-02353-9

BMC Medical Research Methodology

RESEARCH

Open Access

Multiple imputation using auxiliary imputation variables that only predict missingness can increase bias due to data missing not at random

Elinor Curnow[1,2*], Rosie P. Cornish[1,2], Jon E. Heron[1,2], James R. Carpenter[3,4] and Kate Tilling[1,2]

Abstract

Background Epidemiological and clinical studies often have missing data, frequently analysed using multiple imputation (MI). In general, MI estimates will be biased if data are missing not at random (MNAR). Bias due to data MNAR can be reduced by including other variables ("auxiliary variables") in imputation models, in addition to those required for the substantive analysis. Common advice is to take an inclusive approach to auxiliary variable selection (i.e. include all variables thought to be predictive of missingness and/or the missing values). There are no clear guidelines about the impact of this strategy when data may be MNAR.

Methods We explore the impact of including an auxiliary variable predictive of missingness but, in truth, unrelated to the partially observed variable, when data are MNAR. We quantify, algebraically and by simulation, the magnitude of the additional bias of the MI estimator for the exposure coefficient (fitting either a linear or logistic regression model), when the (continuous or binary) partially observed variable is either the analysis outcome or the exposure. Here, "additional bias" refers to the difference in magnitude of the MI estimator when the imputation model includes (i) the auxiliary variable and the other analysis model variables; (ii) just the other analysis model variables, noting that both will be biased due to data MNAR. We illustrate the extent of this additional bias by re-analysing data from a birth cohort study.

Results The additional bias can be relatively large when the outcome is partially observed and missingness is caused by the outcome itself, and even larger if missingness is caused by both the outcome and the exposure (when either the outcome or exposure is partially observed).

Conclusions When using MI, the naïve and commonly used strategy of including all available auxiliary variables should be avoided. We recommend including the variables most predictive of the partially observed variable as auxiliary variables, where these can be identified through consideration of the plausible casual diagrams and missingness mechanisms, as well as data exploration (noting that associations with the partially observed variable in the complete records may be distorted due to selection bias).

Keywords Missing data, Multiple imputation, Bias amplification, Auxiliary variable, ALSPAC

*Correspondence:
Elinor Curnow
elinor.curnow@bristol.ac.uk
Full list of author information is available at the end of the article

Introduction

Epidemiological studies often have missing data, with multiple imputation (MI) a commonly-used, flexible, and general method for analysing partially observed datasets [1]. A sufficient condition for unbiased estimation using MI is that data are either missing completely at random (MCAR) or missing at random (MAR), conditional on the observed data. In addition, imputation models must be appropriately specified and compatible with the substantive analysis model. This means that imputation models should contain the same variables and include any required non-linear terms or interactions implied by the analysis model [2] – with the "substantive model compatible" approach used in situations where it is difficult to specify a compatible model [3]. In general, MI estimates will be biased if data are missing not at random (MNAR), unless additional information is available. Table 1 provides an intuitive, practical, interpretation of these, and other, missing data terms—for a full discussion, with examples, see Chapter 1 of Carpenter et al. [4], and the more detailed description of MAR in the Discussion section.

Common MI strategies when data are suspected to be MNAR include:

1. Exploring the sensitivity of MI results to departures from the MAR assumption using a "pattern mixture" approach [6]. In this approach, the observed and missing values are allowed to differ by a value, or set of values, δ (the "sensitivity parameter").
2. Applying an MI method that can accommodate data MNAR, such as the "not-at-random fully conditional specification" procedure [7]. This is an extension of the pattern mixture approach.
3. Including a "proxy" for the partially observed variable (*i.e.* a variable that is predictive of the missing values) as an auxiliary variable (Table 1) in the imputation model [8].

Note in some MNAR settings, complete records analysis (CRA, Table 1) will yield unbiased estimates when MI will not *e.g.* if estimating the exposure coefficient from a linear regression when the exposure is MNAR and missingness is unrelated to the analysis outcome [9].

In this paper, we focus on strategy 3, including auxiliary variables in the imputation model. We highlight the consequences of using an inclusive strategy for auxiliary variable selection (*i.e.* including all variables thought to be predictive of missingness and/or the missing values) as has been suggested previously (and, anecdotally, is common practice) [10–12]. We demonstrate that when data are MNAR and the imputation model includes a predictor of missingness that is, in truth, unrelated to the partially observed variable, then the bias due to data being MNAR may be increased rather than reduced. This occurs in a similar way to bias amplification in the presence of unmeasured confounding when conditioning on variables that only influence an exposure [13].

Our motivating example is a longitudinal cohort study where we are interested in the association between

Table 1 Missing data definitions

Term	Definition
Complete Records Analysis (CRA)	Analysis is restricted to subjects who have complete data for all variables in the analysis model
Missing Completely At Random (MCAR)	The probability that data are missing is independent of the observed and missing values of variables in the analysis model, and of any related variables. Data can be MCAR if missingness is caused by a variable independent of those in the analysis model *e.g.* if missingness is for administrative reasons
Missing At Random (MAR)	Given the observed data, the probability that data are missing is independent of the true values of the partially observed variable. Any systematic differences between the observed and missing values can be explained by associations with the observed data
Missing Not At Random (MNAR)	If data are not MCAR nor MAR, data are said to be MNAR. The probability that data are missing depends on the (unobserved) values of the partially observed variable, even after conditioning on the observed data
Multiple Imputation (MI)	MI is a method for handling missing data. It consists of three steps: 1. An imputation model is fitted to the observed data (this is usually some form of regression model). The missing values are replaced with draws ("imputed") from its predictive distribution (after first perturbing the model parameters). This imputation stage is carried out multiple (M) times, to give M completed datasets 2. The analysis model is fitted to each of the M completed datasets 3. The M sets of results are combined using Rubin's rules, [5] to correctly account for the uncertainty about the missing values
Predictive Mean Matching (PMM)	PMM is an MI approach that uses an alternative method in Step 1 of the MI process: instead of imputing missing values directly from the conditional predictive distribution of the missing data given the observed data, each missing value is replaced with an observed value randomly chosen from a donor pool anchored on the conditional predicted mean
Auxiliary variable	A variable that is not in the analysis model but that is included as a predictor in the imputation model

a partially observed outcome, child's IQ, and a fully observed exposure, duration of breastfeeding. If the probability that child's IQ is missing (its "missingness") is related to neither observed nor missing values of child's IQ, given the observed data for the other analysis model variables, and all these variables are included in both analysis and imputation models, then both MI and CRA estimates of the outcome-exposure association will be unbiased [9]. On the other hand, suppose that missingness in child's IQ is caused by child's IQ itself, as depicted in the causal diagram (or directed acyclic graph, DAG) in Fig. 1 (with X, Y, and R_{ind} denoting our exposure (duration of breastfeeding), outcome (child's IQ), and missingness indicator for the outcome, respectively). In this case, child's IQ is MNAR, conditional on the fully observed exposure. Since child's IQ is the outcome of the substantive analysis, both CRA and MI estimates of the outcome-exposure association will be biased [9].

If we apply strategy 3, described above, by including the proxy "child's educational attainment score" in the imputation model for child's IQ, we may reduce the bias in the exposure-outcome association due to child's IQ being MNAR [8, 14]. This is because child's educational attainment score is highly correlated with child's IQ. However, including a predictor of missingness in the imputation model where we believe this is unrelated to child's IQ (denoted by Z in Fig. 1, *e.g.* whether the mother smoked during pregnancy) may increase the bias of the MI estimate. Note (depending on the magnitude and direction of the associations), bias due to data MNAR may also increase even if auxiliary variables are predictive of both the missing values and missingness, particularly if the auxiliary variable is a collider [15].

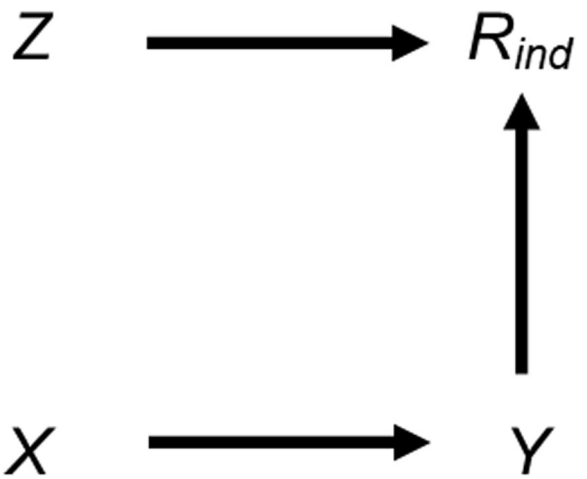

Fig. 1 Directed acyclic graph depicting the relationship between outcome Y, exposure X, missingness indicator for the outcome R_{ind}, and potential auxiliary variable Z. Lines indicate causally related variables, with arrows indicating the direction of the causal relationship; absent lines represent variables with no direct causal relation

In this paper we quantify the magnitude of the additional bias of the MI estimator of the exposure coefficient (fitting either a linear or logistic regression model) due to using an auxiliary variable that predicts missingness, but not the values of the partially observed variable itself, when data are MNAR. By "additional bias", we mean the difference between the MI estimator when including a predictor of missingness in the imputation model (as well as the other analysis model variables) and the MI estimator when including just the other analysis model variables in the imputation model (noting that both estimators will yield biased estimates of the true outcome-exposure association when data are MNAR). We consider settings in which either the outcome or exposure are MNAR, where the partially observed variable is either continuous or binary, and where missingness is caused by the partially observed variable itself and/or another related variable. We quantify the additional bias using algebraic methods and by simulation, and illustrate our results using the real data example described above.

Throughout the paper, we assume that MI is performed by replacing missing values with draws from a suitable regression model (*i.e.* a linear or logistic regression model when the partially observed variable is continuous or binary, respectively) using a linear combination of the specified predictors. We focus on this approach, rather than *e.g.* predictive mean matching (PMM, Table 1) [16] because MI using draws from a correctly specified model will generally yield more precise estimates than PMM [17]. All analyses were conducted using Stata (17.0, StataCorp LLC, College Station, TX). Stata code to perform the simulation studies and real data analysis are included in the final sections of the Supplementary Material (Sections S6 and 7).

Scenario 1. Additional bias of the MI estimator from including a predictor only of missingness in the imputation model when continuous outcome *Y* is partially observed and missingness is caused by *Y*

Methods

We first consider the setting in Fig. 1, discussed above, in more detail. This simplified setting is chosen to give insights into the more complex settings that typically occur in epidemiological practice. We are interested in the relationship between a continuous outcome Y and a continuous exposure X, with β_{YX} (the parameter of interest) denoting the exposure coefficient from a linear regression of Y on X. We assume that X is fully observed and Y is partially observed, with variable R_{ind} denoting the missingness indicator for Y ($R_{ind}=1$ if Y is observed,

and 0 otherwise) and π_1 denoting the probability that $R_{ind}=1$, or $\pi_1 = \mathrm{P}(R_{ind}=1)$. Our substantive model is simply the regression of Y on X; we do not adjust for (fully observed) continuous variable Z because it does not confound the X–Y relationship. Since Y is MNAR, with missingness caused by Y itself, the MI estimator will be biased (as will CRA), assuming the proportion of missing data is greater than zero.

Maximum additional bias of the MI estimator

Here we provide general expressions for the maximum additional bias of the MI estimator (when using X and Z as predictors in the imputation model for Y compared with just X), when continuous outcome Y is MNAR and missingness is caused by Y.

We assume that the joint distribution of Y, X, Z, and R is multivariate normal (with R denoting the latent normal variable for the binary missingness indicator variable R_{ind}), with associated univariate normal distributions defined as follows: $Y = \beta_{YX}X + \varepsilon_Y$ where $\varepsilon_Y \sim \mathrm{N}(0, \sigma_Y^2)$; $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$; $Z \sim \mathrm{N}(\mu_Z, \sigma_Z^2)$; $R = \beta_{RY}Y + \beta_{RZ}Z + \varepsilon_R$ where $\varepsilon_R \sim \mathrm{N}(0, \sigma_R^2)$. R is related to R_{ind} such that $\pi_1 = \mathrm{P}(R_{ind}=1) = \mathrm{P}(R \leq \mathrm{r}) = \Phi\left(\frac{r-\mu_R}{\sqrt{V_R}}\right)$, with $\Phi(.)$ denoting the cumulative distribution function of the standard normal distribution and μ_R and V_R denoting the mean and variance of R, respectively. We further assume that each of Y and R is a linear combination of the variables causing it plus an error term (with X and Z having no direct causes), with no interactions, all errors uncorrelated, no model mis-specification, and no measurement error, and that an ordinary least squares (OLS) estimator is used to obtain estimates in both analysis and imputation models.

Following the argument of Curnow et al. [18], we first provide general expressions for the expected value of the MI estimator in this setting, when using either X, or X and Z, to impute Y. In general, the expected value depends on the set of records with observed values of Y, *i.e.* those for which the missingness indicator, $\mathrm{R_{ind}}$, equals 1, or equivalently, those for which its underlying normal variable $R \leq r$. For example, when using X to impute Y, the expected value of the MI estimator equals the expected value of $\widehat{\beta}_{YX|R_{ind}=1}$, or $\widehat{\beta}_{YX|R\leq r}$, taking expectations first over the imputation distribution, given the set of observed values of Y, and then over this set of values itself. When there are no missing data, the expected value of the MI estimator is equal to β_{YX}, *i.e.* unbiased, (and bias will be minimal if data are nearly complete). As we detail in Supplementary Material, Sect. S1, as the proportion of missing values of Y tends to one, we can approximate, with increasing accuracy, $\widehat{\beta}_{YX|R\leq r}$ by $\widehat{\beta}_{YX|R=r}$ for some r (tending to $-\infty$). In this limiting case, the expected value of the MI estimator will tend to a maximum value of $\beta_{YX|R=r} \approx \beta_{YX|R_{ind}=1}$ (denoting the exposure coefficient from a linear regression of Y on X and R or R_{ind}). Using a similar argument, the expected value of the MI estimator will tend to a maximum value of $\beta_{YX|Z=z,R=r} \approx \beta_{YX|Z=z,R_{ind}=1}$ (denoting the exposure coefficient from a linear regression of Y on X, Z, and R or R_{ind}) when using both X and Z to impute Y. Note that $\beta_{YX|R=r}$ and $\beta_{YX|Z=z,R=r}$ do not depend on the specific values of r and z, and we use the more general forms $\beta_{YX|R}$ and $\beta_{YX|Z,R}$ hereafter. Hence, the maximum additional bias of the MI estimator (*i.e.* the difference between the maximum bias of the two MI estimators) when using X and Z as predictors in the imputation model for Y compared with just X is $\beta_{YX|Z,R} - \beta_{YX|R}$. Full derivations of this and other results in this section are included in the Supplementary Material, Section S1. Equations were verified by simulation (Supplementary Material, Section S2).

Maximum additional bias of the MI estimator in terms of the direct effect sizes

We next provide a general expression for the maximum additional bias of the MI estimator in terms of the direct effect sizes and error variances. $\beta_{YX|R}$ and $\beta_{YX|Z,R}$ can be expressed as follows:

$$\beta_{YX|R} = \beta_{YX} \times \left\{1 - \frac{\beta_{RY}^2\sigma_Y^2}{\beta_{RY}^2\sigma_Y^2 + \beta_{RZ}^2\sigma_Z^2 + \sigma_R^2}\right\} \quad (2.1)$$

And

$$\beta_{YX|Z,R} = \beta_{YX} \times \left\{1 - \frac{\beta_{RY}^2\sigma_Y^2}{\beta_{RY}^2\sigma_Y^2 + \sigma_R^2}\right\} \quad (2.2)$$

where the direct effect sizes are denoted by $\beta_{..}$, *e.g.*β_{RY} denotes the direct effect of Y on R, and the error variances are denoted by $\sigma_{.}^2$, *e.g.*σ_Y^2 denotes the error variance of Y.

Since $0 < \frac{\beta_{RY}^2\sigma_Y^2}{\beta_{RY}^2\sigma_Y^2+\sigma_R^2+\beta_{RZ}^2\sigma_Z^2} < \frac{\beta_{RY}^2\sigma_Y^2}{\beta_{RY}^2\sigma_Y^2+\sigma_R^2} < 1$ (assuming all parameters are non-zero), $\left|\beta_{YX|Z,R}\right| < \left|\beta_{YX|R}\right| < |\beta_{YX}|$, that is, the MI estimator of β_{YX} will be biased towards zero. The maximum bias will be greater in magnitude when the imputation model includes X and Z as predictors than when it includes only X.

Then the maximum additional bias of the MI estimator from including Z as a predictor is: $\beta_{YX|Z,R} - \beta_{YX|R} = \frac{-\beta_{YX}\beta_{RY}^2\beta_{RZ}^2\sigma_Y^2\sigma_Z^2}{(\beta_{RY}^2\sigma_Y^2+\sigma_R^2+\beta_{RZ}^2\sigma_Z^2)(\beta_{RY}^2\sigma_Y^2+\sigma_R^2)}$ (2.3).

Equation (2.3) shows that the magnitude of the maximum additional bias will depend on the strength of the *Y*-*X*, *R*-*Y*, and *R*-*Z* relationships, as well as on the size of the error variances. There will be no additional bias if at least one of β_{YX}, β_{RY}, or β_{RZ} is equal to zero, consistent with the underlying DAG (Fig. 1). Note that

we can also express the effect on the MI estimator of including Z as a predictor in the imputation model in terms of bias amplification (defined as the bias of $\beta_{YX|Z,R}$ divided by the bias of $\beta_{YX|R}$): when Z (as well as X) is included in the imputation model for Y, the maximum bias due to Y being MNAR is amplified by a factor of:

$$\left\{1+\frac{\beta_{RZ}^2\sigma_Z^2}{\beta_{RY}^2\sigma_Y^2+\sigma_R^2}\right\} \tag{2.4}$$

Note that if instead X was partially observed and Y was fully observed, MI would yield unbiased results (given a correctly specified imputation model) because in this case R would not be related to X after conditioning on Y. However, CRA would still be invalid because missingness depends on the analysis outcome.

Illustration of maximum additional bias of the MI estimator

We illustrate how the magnitude of the maximum additional bias, calculated using Eq. 2.3, varies with the direct effect sizes. We use a numerical example, with moderate values of the direct effect sizes β_{YX}, β_{RY}, and β_{RZ} relative to the error variances, which were all equal to one. Hence, the magnitude of the biases in our example can be interpreted as both absolute bias and the bias, relative to the error variances. Direct effect sizes were each set to 0.00, 0.25, 0.50, 0.75, or 1.00. For β_{RY} and β_{RZ}, note that these values correspond approximately to odds ratios (from a logistic regression model for R_{ind}) of 1.00, 1.50, 2.30, 3.50, or 5.30 (using the general rule for transforming a coefficient from a logistic to a probit model [19]). Figure 2 illustrates the impact of the direct effect sizes β_{YX}, β_{RY}, and β_{RZ} on various measures of bias, derived using Eqs. 2.1–2.3. Panel A depicts the

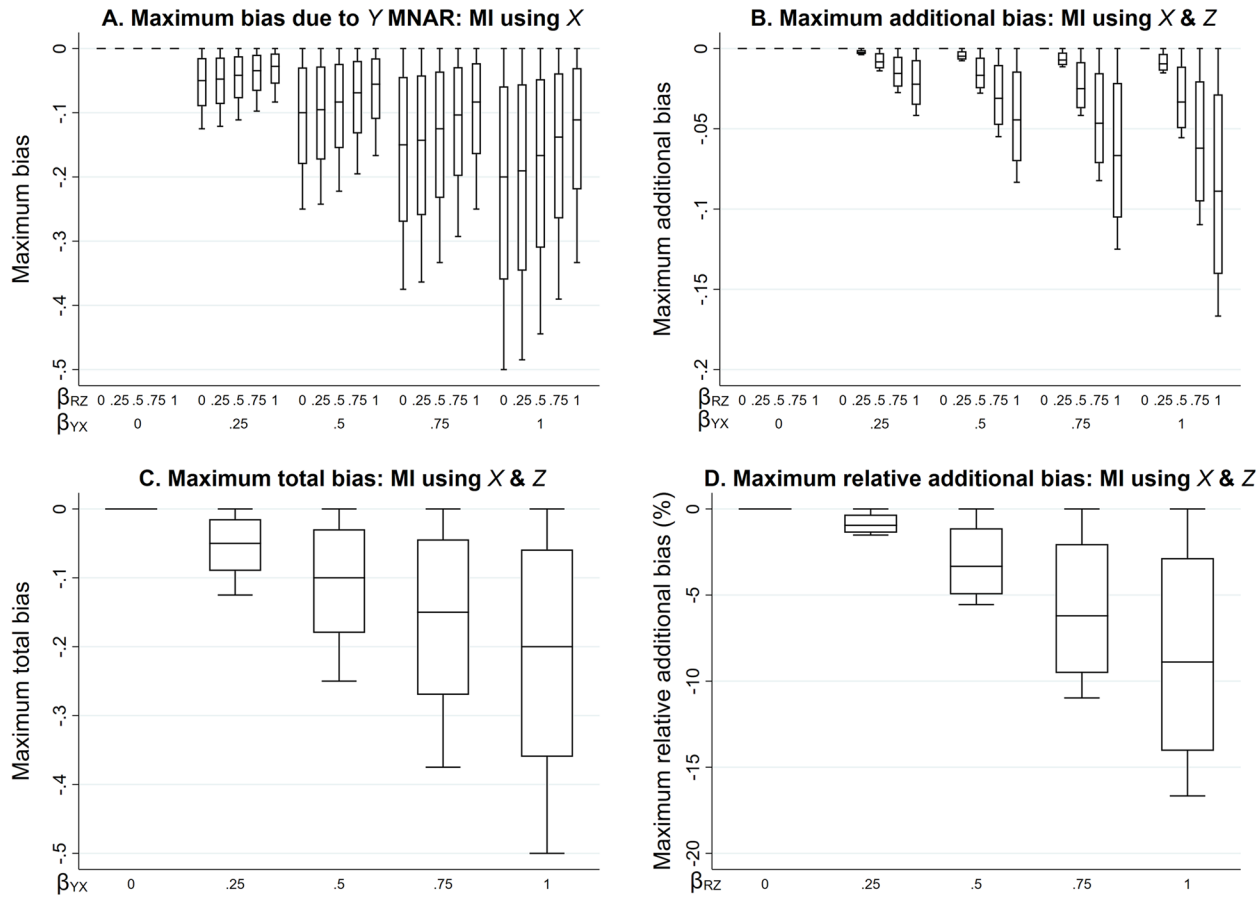

Fig. 2 Bias of the MI estimator of β_{YX} when continuous outcome Y is missing not at random, with missingness caused by Y itself, and the imputation model includes exposure X, or X and a predictor of missingness but not the missing values, Z, varying the direct effect sizes β_{YX}, β_{RY}, and β_{RZ}. Panel A depicts the maximum bias when the imputation model includes X. Panels B-D depict the maximum additional bias, maximum total bias, and maximum relative additional bias, respectively, when the imputation model includes X and Z. All bias quantities were calculated using Eqs. 2.1–2.3. The distribution of each box-plot is due to variation in β_{RY}. Note that maximum total bias depends on β_{YX} and β_{RY} but not β_{RZ}; maximum relative additional bias depends on β_{RZ} and β_{RY} but not β_{YX}

maximum bias of the MI estimator (due to Y being MNAR) when the imputation model includes only X as a predictor. Panels B-D depict the maximum additional bias (compared to the maximum bias due to Y being MNAR), the maximum total bias (the sum of the maximum bias due to Y being MNAR and the maximum additional bias), and the maximum relative additional bias (maximum additional bias multiplied by 100, divided by β_{YX}), respectively, when the imputation model includes both X and Z as predictors. The distribution of each box-plot is due to variation in β_{RY}.

Each measure of bias is equal to zero if β_{YX} is equal to zero (additionally, the maximum additional bias is equal to zero if any of the direct effect sizes are equal to zero), and negative otherwise. The maximum bias due to Y being MNAR increases in magnitude with β_{YX}, but for a given value of β_{YX}, decreases in magnitude as β_{RZ} increases. However, the maximum additional bias increases in magnitude with each of the direct effect sizes, as do the maximum total bias (which depends on β_{YX} and β_{RY} but not β_{RZ}) and the maximum relative additional bias (which depends on β_{RZ} and β_{RY} but not β_{YX}). Note that all parameters have a zero or positive value in this illustration. However, if, for example, we take the same values as mentioned above for β_{RY} and β_{RZ}, but set β_{YX} to negative values, then the measures of bias would be of the same magnitude but positive.

When the relationships are as depicted in Fig. 1, but Y is binary, the results described here still approximately apply (results obtained by simulation, see Supplementary Material, Section S3 and Figure S1). This follows by assuming that Y has an underlying normal distribution (which is valid provided the probability of each value of Y is not close to 0 or 1).

Scenario 2. Additional bias of the MI estimator from including a predictor only of missingness in the imputation model when continuous outcome *Y* or continuous exposure *X* are partially observed and missingness is related to *Y* via an unmeasured variable

We next consider the setting in which missingness of the partially observed variable (either Y or X) is related to Y via an unmeasured variable, U, as depicted in Fig. 3. We assume that the joint distribution of Y, X, Z, U, and R is multivariate normal (with R denoting the latent normal variable for the binary missingness indicator variable R_{ind}), with associated univariate normal distributions defined as follows: $Y = \beta_{YX}X + \beta_{YU}U + \varepsilon_Y$ where $\varepsilon_Y \sim \mathrm{N}(0, \sigma_Y^2)$; $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$; $Z \sim \mathrm{N}(\mu_Z, \sigma_Z^2)$; $U \sim \mathrm{N}(\mu_U, \sigma_U^2)$; $R = \beta_{RZ}Z + \beta_{RU}U + \varepsilon_R$ where $\varepsilon_R \sim \mathrm{N}(0, \sigma_R^2)$.

In this setting (given the same assumptions and using the same analysis model and MI method as in the previous scenario), we would expect the CRA estimator and the MI estimator to be biased because missingness is related to our analysis outcome Y (conditional on X), via U. However, in the special case in which partially observed variable Y is continuous and the analysis model is a linear regression, both the CRA and MI estimators (using either X, or X and Z, as predictors in the imputation model for Y) are unbiased. Proof is provided in Supplementary Material, Section S4. Note that this is not the case if Y is binary, although the bias is generally small (results obtained by simulation, see Supplementary Material, Section S3 and Figures S2-3).

When X is partially observed, the MI estimator (using either Y, or Y and Z, as predictors in the imputation model for X) will be biased because missingness is related to X, conditional on Y. The theoretical magnitude of the maximum additional bias has a more complicated form

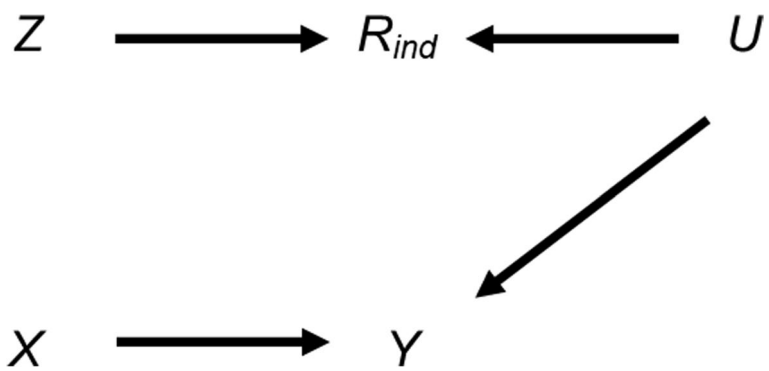

Fig. 3 Directed acyclic graph depicting the relationship between outcome Y, exposure X, missingness indicator for the outcome or exposure R_{ind}, potential auxiliary variable Z, and unmeasured variable U. Lines indicate causally related variables, with arrows indicating the direction of the causal relationship; absent lines represent variables with no direct causal relation

when X is partially observed because the imputation and analysis models are not the same. Again following the argument of Curnow et al. [18], the MI estimator will be unbiased only if unbiased estimates of all the imputation model coefficients can be obtained using records with observed values of X. However, taking the imputation model coefficient for Y as an example, we find that this coefficient is biased, taking its maximum value of $\beta_{XY|R}$ (denoting the exposure coefficient from a linear regression of X on Y and R) when the imputation model includes only Y, and $\beta_{XY|R,Z}$ (denoting the exposure coefficient from a linear regression of X on Y, Z, and R) when the imputation model includes Y and Z, as the proportion of missing values tends to one.

In terms of the direct effect sizes and error variances,

$$\beta_{XY|R} = \beta_{XY} \times \frac{1}{1 - \left\{ \beta_{YU}^2 \beta_{RU}^2 \sigma_U^4 / \left(\beta_{YX}^2 \sigma_X^2 + \beta_{YU}^2 \sigma_U^2 + \sigma_Y^2 \right) \left(\beta_{RZ}^2 \sigma_Z^2 + \beta_{RU}^2 \sigma_U^2 + \sigma_R^2 \right) \right\}} \tag{3.1}$$

and

$$\beta_{XY|R,Z} = \beta_{XY} \times \frac{1}{1 - \left\{ \beta_{YU}^2 \beta_{RU}^2 \sigma_U^4 / \left(\beta_{YX}^2 \sigma_X^2 + \beta_{YU}^2 \sigma_U^2 + \sigma_Y^2 \right) \left(\beta_{RU}^2 \sigma_U^2 + \sigma_R^2 \right) \right\}} \tag{3.2}$$

where the direct effect sizes are denoted by $\beta_{..}$, *e.g.*β_{RU} denotes the direct effect of U on R, and the error variances are denoted by $\sigma_{.}^2$, *e.g.*σ_Y^2 denotes the error variance of Y. Since $\left|\beta_{XY|R,Z}\right| > \left|\beta_{XY|R}\right| > |\beta_{XY}|$, bias of the Y coefficient will be amplified when Z is also included as a predictor in the imputation model for X (see Supplementary Material, Section S4, for derivation of these results).

Due to its complexity in the setting in which X is partially observed, an expression for the theoretical magnitude of the additional bias of the MI estimator is not derived here. However, we illustrate the effect on the MI estimate from including auxiliary variable Z in the imputation model by simulation (see Supplementary Material Section S3 for further details). Note that we refer to the MI or CRA "estimate" when describing simulation study results, rather than "estimator" (which we have used when describing algebraic results).

Figure 4 illustrates the impact of the direct effect sizes on the additional bias of the MI estimate when the imputation model includes Z as a predictor and X is partially observed (with 50% missing values). As before, the additional bias is plotted against β_{YX} and β_{RZ}. The

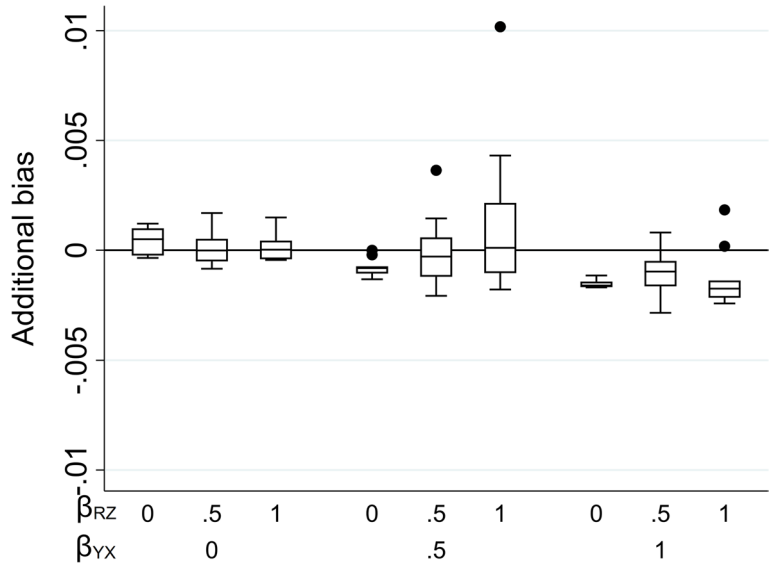

Fig. 4 Additional bias of the MI estimate of β_{YX} when continuous exposure X is missing not at random, conditional on outcome Y, and the imputation model includes Y and an auxiliary variable Z that predicts missingness but not the missing values, with missingness related to Y via an unmeasured variable U. Simulation results shown when 50% of values are missing, varying the direct effect sizes β_{YX}, β_{YU},β_{RU}, and β_{RZ}. The distribution of additional bias in each box-plot is averaged over the values of β_{YU} and β_{RU}

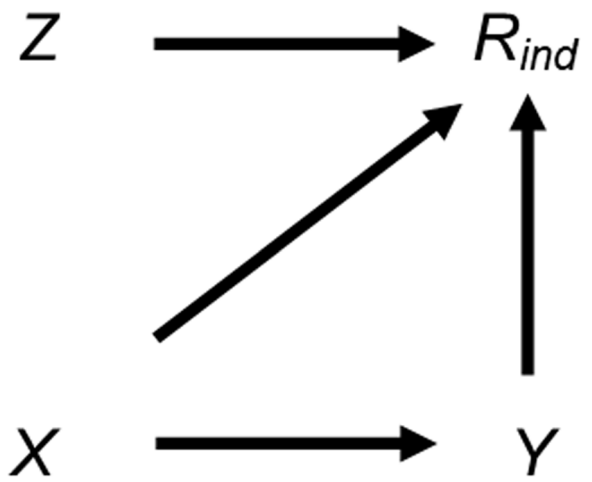

Fig. 5 Directed acyclic graph depicting the relationship between outcome Y, exposure X, missingness indicator for the outcome or exposure R_{ind}, and potential auxiliary variable Z. Lines indicate causally related variables, with arrows indicating the direction of the causal relationship; absent lines represent variables with no direct causal relation

distribution of the additional bias for each value of β_{YX} and β_{RZ} (represented as a box-plot) is due to the variation in β_{YU} and β_{RU}. Figure 4 shows that the additional bias is small, regardless of the direct effect sizes. Results are similar if X is binary (see Supplementary Material, Figure S4).

Scenario 3. Additional bias of the MI estimator from including a predictor only of missingness in the imputation model when continuous outcome *Y* or continuous exposure *X* are partially observed and missingness is caused by both *X* and *Y*

Finally, we consider the setting in which the CRA and MI estimators are biased if either Y or X are partially observed: when Y and X directly cause missingness, as per Fig. 5. We assume that the joint distribution of Y, X, Z, and R is multivariate normal (with R denoting the latent normal variable for the binary missingness indicator variable R_{ind}), with associated univariate normal distributions defined as follows: $Y = \beta_{YX}X + \varepsilon_Y$ where $\varepsilon_Y \sim \mathrm{N}(0, \sigma_Y^2)$; $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$; $Z \sim \mathrm{N}(\mu_Z, \sigma_Z^2)$; $R = \beta_{RY}Y + \beta_{RX}X + \beta_{RZ}Z + \varepsilon_R$ where $\varepsilon_R \sim \mathrm{N}(0, \sigma_R^2)$.

In this setting (given the same assumptions and using the same analysis model and MI method as in the previous scenarios), we can express both the maximum bias due to Y being MNAR (when using X as the predictor in the imputation model for Y) and the maximum additional bias (from including Z as well as X in the imputation model) in terms of the direct effect sizes and error variances (see Supplementary Material, Section S5, for derivation), as follows:

$$\text{maximum bias } = -\frac{\beta_{YX}\beta_{RY}\sigma_Y^2\left(\beta_{RY} + \frac{\beta_{RX}}{\beta_{YX}}\right)}{\beta_{RY}^2\sigma_Y^2 + \sigma_R^2 + \beta_{RZ}^2\sigma_Z^2} \quad (4.1)$$

and

$$\text{maximum additional bias } = \frac{-\beta_{YX}\beta_{RY}\beta_{RZ}^2\sigma_Y^2\sigma_Z^2\left(\beta_{RY} + \frac{\beta_{RX}}{\beta_{YX}}\right)}{(\beta_{RY}^2\sigma_Y^2 + \sigma_R^2 + \beta_{RZ}^2\sigma_Z^2)(\beta_{RY}^2\sigma_Y^2 + \sigma_R^2)} \quad (4.2)$$

where the direct effect sizes are denoted by $\beta_{..}$, *e.g.*β_{RY} denotes the direct effect of Y on R, and the error variances are denoted by $\sigma_{.}^2$, *e.g.*σ_Y^2 denotes the error variance of Y, as before. Note that in this setting, the maximum bias may be towards or away from zero, depending on the sign and magnitude of the direct effects, relative to the magnitude of the error variances. However, the maximum additional bias will always be in the same direction as the maximum bias, and will amplify the bias by a factor of

$$\left\{1 + \frac{\beta_{RZ}^2\sigma_Z^2}{\beta_{RY}^2\sigma_Y^2 + \sigma_R^2}\right\} \quad (4.3)$$

Note that this is identical to the amplification factor in Scenario 1 (although the bias due Y being MNAR in this setting may be greater or smaller than in Scenario 1, depending on the sign and magnitude of $\frac{\beta_{RX}}{\beta_{YX}}$).

When X is partially observed, the maximum additional bias of the Y coefficient in the imputation model for X (*i.e.* in addition to the bias due to X being MNAR) from including Z as a predictor in the imputation model is equal to:

$$\beta_{XY}\beta_{RX}\left\{\frac{\beta_{RY}\sigma_Y^2}{\beta_{YX}} + \beta_{RX}\sigma_X^2\left(1 - \frac{\beta_{YX}^2\sigma_X^2}{\beta_{YX}^2\sigma_X^2 + \sigma_Y^2}\right)\right\} \times \left\{\frac{1}{\beta_{RX}^2\sigma_X^2 + \beta_{RZ}^2\sigma_Z^2 + \sigma_R^2 - \{\beta_{YX}^2\beta_{RX}^2\sigma_X^4/(\beta_{YX}^2\sigma_X^2 + \sigma_Y^2)\}} - \frac{1}{\beta_{RX}^2\sigma_X^2 + \sigma_R^2 - \{\beta_{YX}^2\beta_{RX}^2\sigma_X^4/(\beta_{YX}^2\sigma_X^2 + \sigma_Y^2)\}}\right\} \quad (4.4)$$

As in the previous scenario, we explored the effect of this additional bias on the MI estimate by simulation, due to the complexity of the theoretical expression for the maximum additional bias when X is partially observed (see Supplementary Material Section S3 for further details).

Figures 6 and 7 illustrate, respectively, the impact of the direct effect sizes on the maximum additional bias when Y is partially observed, calculated using Eq. 4.2, and the additional bias when 50% of values of X are

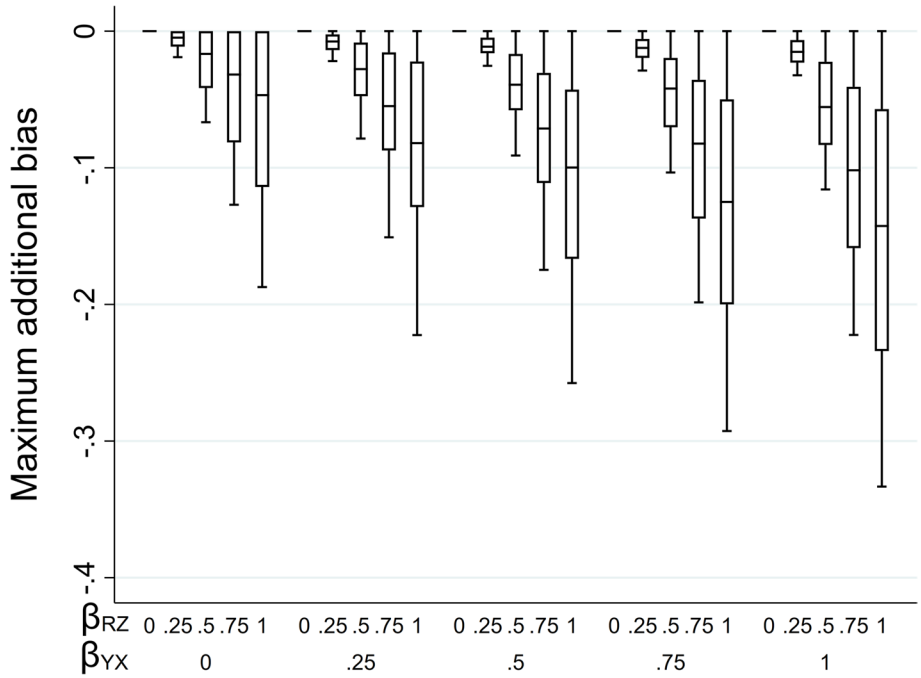

Fig. 6 Maximum additional bias of the MI estimator of β_{YX} when continuous outcome Y is missing not at random, with missingness caused by Y and X, and the imputation model includes exposure X and a predictor of missingness but not the missing values, Z, varying the direct effect sizes β_{YX}, β_{RY}, β_{RX}, and β_{RZ}. Maximum additional bias was calculated using Eq. 4.2. The distribution of maximum additional bias in each box-plot is averaged over the values of β_{RY} and β_{RX}

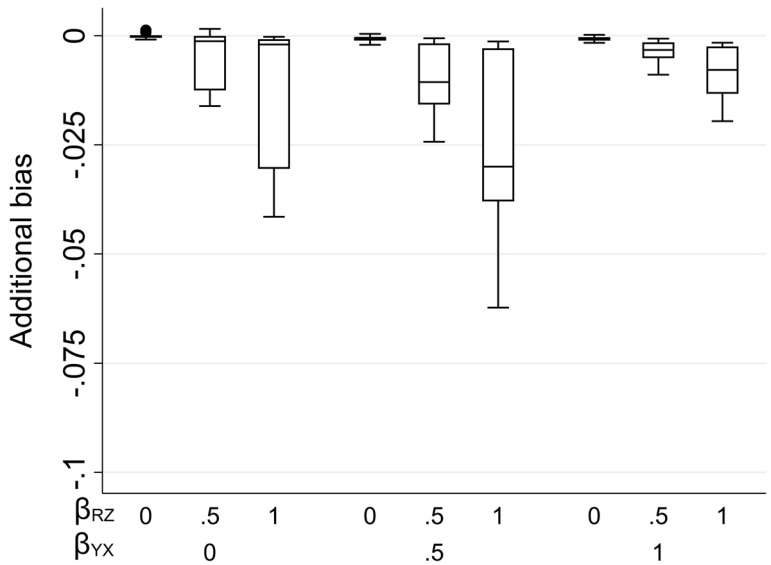

Fig. 7 Additional bias of the MI estimate of β_{YX} when continuous exposure X is missing not at random, with missingness caused by Y and X, and the imputation model includes exposure Y and a predictor of missingness but not the missing values, Z. Simulation results shown when 50% of values are missing, varying the direct effect sizes β_{YX}, β_{RY}, β_{RX}, and β_{RZ}. The distribution of additional bias in each box-plot is averaged over the values of β_{RY} and β_{RX}

missing, estimated by simulation, when the imputation model includes Z as a predictor. The distribution of each box-plot is due to the variation in β_{RY} and β_{RX}. When Y is partially observed, Fig. 6 shows that the maximum additional bias is negative and increases in magnitude with the direct effect sizes (and is larger than in Scenario

1 for the same direct effect sizes). When X is partially observed, Fig. 7 shows that the additional bias is negative and increases in magnitude with β_{RZ}, as well as with β_{YX} when $\beta_{YX} \leq 0.5$. However, the additional bias is smaller in magnitude when $\beta_{YX} = 1$. Results for partially observed Y are similar if Y is binary (see Supplementary Material, Figure S5). If X is binary, additional bias when X is partially observed increases with both β_{YX} and β_{RZ} (see Supplementary Material, Figure S6). The difference between results when X is continuous or binary may be due to the choice of distribution of X in each case: in the continuous case, X is normally distributed, with mean equal to 0 and variance equal to 1; in the binary case, X takes values of 0 or 1 with probability 0.5 (equivalent to a mean of 0.5 and a variance of 0.25).

Real data example

Methods

We illustrate this situation using data from the Avon Longitudinal Study of Parents and Children (ALSPAC). ALSPAC is a prospective study which recruited pregnant women with expected dates of delivery between 1st April 1991 and 31st December 1992, in the Bristol area of the UK [20, 21]. We use data from the initial recruitment phase, in which 14,541 pregnant women enrolled, resulting in 14,062 live births (13,988 alive at one year). This study uses data from all singletons and twins, where neither the mother nor child had withdrawn consent at the time of analysis (N=13,923). Children and their mothers have been followed up since birth through questionnaires, clinics, and linkage to routine datasets. ALSPAC has a searchable data dictionary: http://www.bristol.ac.uk/alspac/researchers/our-data/, describing all available data; the (previously-published) questionnaires and clinical measures used in our analysis are also described (see https://www.bristol.ac.uk/alspac/researchers/our-data/questionnaires/ and https://www.bristol.ac.uk/alspac/researchers/our-data/clinical-measures/). Ethical approval was obtained from the ALSPAC Ethics and Law Committee and local research ethics committees. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time.

Here, our substantive model of interest is a linear regression of child's IQ at age 15 years (*IQ15*) on breastfeeding duration (*bf*: categorised as never/<3 months versus 3 months plus). Guided by previous studies [8, 14], we adjust for six confounders of the breastfeeding-IQ relationship, namely child's sex, mother's educational level (whether the child's mother held a post-16 years qualification or not), mother's occupational social class (professional, managerial, or non-manual skilled occupation vs. manual skilled, semi-skilled, or unskilled occupation), mother's age and parity (number of previous births), and housing tenure (whether the family home was owned/mortgaged, privately rented, or rented from the local council or a housing association).

IQ15 was not reported for 8913 (64%) participants in the study. Previous studies [8, 14] used linked educational attainment data to explore the missingness mechanism for *IQ15*. They found that *IQ15* was more likely to be missing for individuals with lower educational attainment (highly correlated with *IQ15*), which suggests *IQ15* is MNAR. We explore the consequences of performing MI when *IQ15* is likely to be MNAR, focusing particularly on the effect of including an auxiliary variable that is predictive of missingness but not the missing values of *IQ15*. From previous studies [14, 22], we identified an auxiliary variable which potentially has these properties. Our chosen auxiliary variable is whether the mother smoked during the first trimester of pregnancy (*matsmok*). Note that there were also missing values for *bf*, confounders, and *matsmok*: *bf* was missing for 1406 (10%) individuals, values of one or more confounders were missing for 4394 (32%) individuals (although child's sex and maternal age were fully observed), and *matsmok* was missing for 817 (6%) individuals. For simplicity, and purely for illustrative purposes, we assume that *bf*, confounders, and *matsmok* are MAR, conditional on the observed data.

In Fig. 8, black lines depict our hypothesised relationships between *IQ15*, *bf*, confounders (with confounders collectively denoted by C – for simplicity, we do not depict the relationships between individual confounders and/or missingness indicators for variables other than *IQ15*), potential auxiliary variable *matsmok*, and missingness indicator R_{IQ15} (a binary variable indicating whether *IQ15* is observed). Here, we assume the setting is similar to that depicted in our theoretical Scenario 1 *i.e.* we assume missingness is caused by *IQ15* but not by our exposure, *bf*, or confounders. As in all real data studies, we cannot rule out the existence of unmeasured variable(s) (denoted by U in Fig. 8), which may be related to the analysis model variables and/or their missingness (with these potential relationships denoted by grey lines in Fig. 8). Hence, in reality, there may be further bias due to unmeasured confounding and/or data MNAR beyond that explored here.

We first assessed whether the hypothesised relationships between *IQ15*, R_{IQ15}, *bf*, and *matsmok* were plausible by exploring the relationships in the observed data. We then applied our equation (Eq. 2.4) for maximum bias amplification due to including predictor of missingness *matsmok* in the imputation model for *IQ15*. We assumed (without loss of generality) that R had a mean of zero and a variance of one. Therefore, we

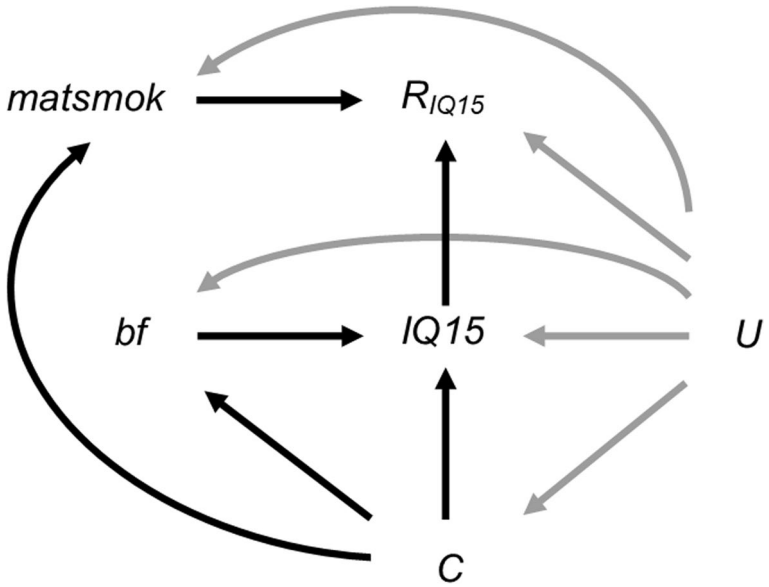

Fig. 8 Directed acyclic graph depicting the relationship between child's IQ at age 15 years (IQ15), duration of breastfeeding (bf), confounders of the IQ15-bf relationship (C), whether the mother smoked during the first trimester of pregnancy (matsmok), missingness indicator R_{IQ15} (a binary variable indicating whether IQ15 is observed), and unmeasured variable(s) U. Lines indicate causally related variables, with arrows indicating the direction of the causal relationship. Black lines depict the assumed causal relationships (including those assumed in theoretical Scenario 1); grey lines depict additional relationships that are plausible in our real data example; absent lines represent variables with no direct causal relationship

used the following version of Eq. 2.4: maximum bias amplification $= 1 + \frac{\beta_{RZ}^2 \sigma_Z^2}{1-\beta_{RZ}^2 \sigma_Z^2 - \beta_{RY}^2 \beta_{YX}^2 \sigma_X^2}$

where, in our setting, X denotes *bf* and Z denotes *matsmok*. Coefficient β_{RZ} and the product $\beta_{RY}\beta_{YX}$ were estimated as 0.6 × the coefficient for *matsmok* and *bf*, respectively, from a logistic regression of R_{IQ15} on *matsmok*, *bf*, and confounders (as before, multiplying by 0.6 to transform the coefficients to the equivalent coefficients from a probit regression of the underlying normal variable R [19]). We estimated $\sigma_X^2 = \text{Var}(X)$ and $\sigma_Z^2 = \text{Var}(Z)$ using the normal approximation to the binomial because X and Z were binary. We assumed that the estimates used in our maximum bias amplification equation were unbiased (which may not have been the case if there were unmeasured confounders of the relationship between *matsmok*, *bf*, and R_{IQ15}).

We compared our estimate of the maximum bias amplification to both the CRA estimate and MI estimates using no auxiliary variables or using *matsmok* as an auxiliary variable. We used MI by chained equations [23] to impute missing values of *IQ15*, *bf*, confounders, and (where used) *matsmok*, including all other variables as predictors in the imputation model for each partially observed variable. We used a linear regression model to impute *IQ15*, logistic regression to impute *bf*, binary confounders, and *matsmok*, ordered logistic regression to impute parity, and multinomial logistic regression to impute housing tenure. We used 20 iterations in the imputation step and a large number of imputations (100) to ensure we obtained stable estimates of the exposure coefficient and its SE.

Results

The estimated association between *matsmok* and *IQ15*, adjusted for *bf* and confounders, was -0.79 (95% CI: -1.88, 0.31). The wide CI suggests that *matsmok* is only weakly predictive of *IQ15*, conditional on *bf* and confounders. We would expect some association between *matsmok* and *IQ15* in the observed data via the *matsmok*—R_{IQ15}—*IQ15* pathway *i.e.* due to collider/selection bias because we are conditioning on R_{IQ15}. Estimates of the coefficient (*i.e.* the logarithm of the odds ratio) for *matsmok* and *bf* from a logistic regression of R_{IQ15} on *matsmok*, *bf*, and confounders, were -0.39 (95% CI: -0.51, -0.27) and 0.44 (95% CI: 0.35, 0.53), respectively. This suggests that, conditional on the confounders and each other, *matsmok* and *bf* are strongly predictive of missingness of *IQ15*. These results, combined with our prior knowledge of the data, suggest that inclusion of *matsmok* in the imputation model for *IQ15* may amplify any bias due to *IQ15* being

MNAR. Note that it is not possible to check for a direct relationship between *IQ15* and its missingness using the observed data due to perfect prediction (because all observed values of *IQ15* have $R_{IQ15}=1$). However, the observed relationship between *bf* and R_{IQ15} is consistent with our assumption (based on our prior knowledge) that missingness depends on *IQ15* itself *i.e.* via the *bf*—*IQ15*—R_{IQ15} pathway (although this observed relationship could also imply *bf* is a direct cause of missingness of *IQ15*).

Substituting values based on the observed data into our equation (with coefficient β_{RZ} and the product $\beta_{RY}\beta_{YX}$ estimated as -0.23 and 0.26, respectively, based on coefficient estimates given above, and additionally, using estimates of Var(Z) and Var(X) of 0.18 and 0.25, respectively), we estimated that including *matsmok* in the imputation model for *IQ15* would amplify any bias in the *bf* coefficient due to *IQ15* being MNAR by 1% towards the null.

Analysis results (Table 2) confirmed that MI estimates of the *bf* coefficient were very similar, regardless of whether auxiliary variable *matsmok* was used in the MI procedure. The MI estimate based on *matsmok* was slightly smaller than the MI estimate based only on analysis model variables, as predicted by our equation and consistent with results in the theoretical Scenario 1. Both MI estimates were smaller than the CRA estimate. Based on the conclusions from previous studies [8, 14], both MI and CRA estimates under-estimate the true magnitude of the association. Using *matsmok*, a predictor of missingness but not of *IQ15* itself, as an auxiliary variable amplifies rather than reduces any bias, albeit the size of the bias amplification is small in this particular setting. The magnitude of bias amplification would be larger in our real data setting if the relationship between our auxiliary variable, *matsmok*, and missingness of *IQ15* was much stronger than the relationship between our exposure, *bf*, and missingness of *IQ15*. This can be seen more clearly if we express our equation for bias amplification, above, as:

$1+\frac{1}{(1/\beta_{RZ}^2\sigma_Z^2)-1-(\beta_{RY}^2\beta_{YX}^2\sigma_X^2/\beta_{RZ}^2\sigma_Z^2)}$ and also note that the terms $\beta_{RZ}^2\sigma_Z^2$ and $\beta_{RY}^2\beta_{YX}^2\sigma_X^2$ represent the squared correlations of *matsmok* and *bf* with missingness of *IQ15*, respectively. In our real data setting, these expressions are of very similar magnitude (*i.e.* the magnitudes of our estimates of both coefficient β_{RZ} and the product $\beta_{RY}\beta_{YX}$, and Var(Z) and Var(X), are very similar). Hence, including *matsmok* in the imputation model for *IQ15* makes little difference to the MI estimate.

Table 2 Relationship between child's IQ at age 15 years and duration of breastfeeding, estimated using different analysis strategies

Duration of breastfeeding	Mean change in child's IQ at age 15: estimate (SE)[a]		
	CRA (*N* = 4,115)	MI, no auxiliary variables (*N* = 13,923)	MI, including auxiliary variable[b] (*N* = 13,923)
Never/ < 3 months	-	-	-
3 months plus	3.75 (0.40)	3.57 (0.35)	3.54 (0.37)

SE Standard error, *CRA* Complete records analysis, *MI* Multiple imputation

[a] Adjusted for mother's educational level, occupational social class, age, parity, and housing tenure, and child's sex

[b] Whether mother smoked during first trimester of pregnancy

Discussion

In this paper, we quantify, algebraically and by simulation, the magnitude of the additional bias of the MI estimator, in addition to any bias due to data MNAR, from including a predictor of missingness but not the missing values themselves in the imputation model. We have derived algebraic expressions for the maximum additional bias when a continuous outcome is partially observed. We have demonstrated that if missingness is caused by the outcome, the additional bias can be substantial, relative to the magnitude of the exposure coefficient (and also if the outcome is binary). Furthermore, if missingness is caused by the outcome and the exposure, the additional bias can be even larger, when either the (continuous or binary) outcome or exposure is partially observed. In both situations, we have shown that the magnitude of the additional bias depends on the relative magnitude of the relationships between the exposure and outcome, and between each of the exposure, outcome, and potential auxiliary variable and missingness, as well as on the proportion of missing data.

In addition, when a continuous analysis model outcome Y is partially observed and linear regression models are fitted (for both analysis and imputation), we have demonstrated algebraically the, perhaps surprising, result: if missingness is only related to Y via another variable U (where U causes Y and its missingness but is only related to exposure X and confounders via Y), then both CRA and MI will be unbiased even if U is not included in the analysis and imputation models. Furthermore, in this scenario, the bias of the MI estimate is likely to be small when binary Y (fitting a logistic regression model) or (continuous or binary) X is partially observed.

A strength of our approach is that we have considered a range of commonly-occurring scenarios, in which the partially observed variable is either the analysis model outcome or the exposure, as well as either continuous or binary. By using both algebra and simulation, we have been able to provide a detailed illustration of the magnitude of bias due to including auxiliary variables that only predict missingness, and how this is related to the magnitude and sign of individual assocations between exposure, outcome, auxiliary variables, and missingness. A limitation of our study is that we have only considered simple models, without interactions or non-linear relationships. However, since our general argument is based on a "missingness" DAG [24, 25], which does not make any distributional assumptions, our findings can be applied to more complex models (*e.g.* including an exposure-confounder interaction), to avoid using MI in a way which may increase bias. Note that the magnitude and direction of additional bias may be different from those suggested by our equations in this case, particularly if either the analysis or missingness model includes interactions.

A further limitation of our study is that in each of our scenarios, only a single variable has missing values. In this case, imputation using draws from a suitable conditional distribution gives a valid imputation from the joint distribution. Given multiple missingness, the chained equations and joint modelling approaches can be made equivalent for multivariate normal data, or approximately equivalent in many cases for binary and categorical data [4, 26, 27]. Thus, we expect our results to apply more generally in multiple missingness settings, regardless of whether a chained equations or joint modelling approach is used. If multiple missingness is handled using MI by chained equations (as we did in the real data example), each imputation model only considers one variable to have missing values, as here. In this case, auxiliary variables should be considered separately for each imputation model, because an auxiliary variable may be predictive of one partially observed variable (and/or its missingness), but not another. Note that in the case of multiple partially observed variables, the MAR assumption may imply different causes of missingness depending on the patterns of missing data. This may be implausible and/or difficult to accommodate in the imputation scheme in practice. In this situation, we recommend focusing on assessing the validity of the MAR assumption for the most common missing data patterns and/or variables with the most missing data. Less common missing data patterns can often be assumed to be missing completely at random—it is unlikely to change the final conclusions if this assumption is incorrect [4, 28].

In summary, we conclude that, whilst auxiliary variables have the potential to improve precision of MI estimates and reduce bias due to data MNAR, the naïve and commonly used strategy of including all available auxiliary variables should be avoided. Any auxiliary variables that, in truth, cause missingness but are independent of the partially observed variable may cause additional bias, over and above any bias due to data MNAR. As with bias amplification in confounding [29], it is possible that variables that are weakly associated, rather than completely independent, of the partially observed variable may also inflate the bias due to data MNAR—this is an area for future research. Note that, in practice, it is generally not possible to determine whether a variable is weakly predictive, rather than independent, of the partially observed variable. This is both due to finite sampling variation and because this requires knowledge of the missing values themselves. Furthermore, auxiliary variables that are only weakly predictive of the partially observed variables can increase the standard error of the MI estimate [10]. Therefore, although it is important to identify predictors of missingness to inform analysis strategy (*e.g.* to determine whether CRA is likely to be valid), our results show that such variables should not necessarily be included as predictors in the imputation models unless they also predict the partially observed variable. Given a choice of potential auxiliary variables, we recommend including the variables most predictive of the partially observed variable as auxiliary variables in the imputation model (in addition to all variables required for the analysis model) in order to minimise the risk of amplifying any bias due to data being MNAR. These variables can be identified through consideration of the plausible casual diagrams and missingness mechanisms, as well as data exploration (noting that associations with the partially observed variable in the complete records may be distorted due to selection bias).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02353-9.

Supplementary Material 1

Acknowledgements
We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

Publication history
This manuscript was previously published in medRxiv: doi: https://doi.org/10.1101/2023.10.17.23297137.

Authors' contributions
EC, RPC, JEH, JRC, and KT contributed to the study conception and design. EC performed the algebraic analysis, analyzed the simulation study and real data, and wrote the first draft of the manuscript. All authors commented on previous versions of the manuscript, and read and approved the final manuscript.

Funding
This work is directly funded by the UK Medical Research Council (grant no MR/V020641/1).
Elinor Curnow, Jon Heron, Rosie Cornish, and Kate Tilling work in the Medical Research Council Integrative Epidemiology Unit at the University of Bristol which is supported by the UK Medical Research Council (grant no MC_UU_00032/02) and the University of Bristol. James Carpenter is also supported by the UK Medical Research Council (grant no MC_UU_00004/04). The UK Medical Research Council and the Wellcome Trust (grant no 217065/Z/19/Z), and the University of Bristol currently provide core funding for ALSPAC. Data collection is funded from a wide range of sources: a comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). This publication is the work of the authors, who will serve as guarantors for the contents of this paper. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
Stata code to verify algebraic results, and also to generate and analyse the data as per the simulation studies is included in Supplementary Material, Section S6. Stata code to perform the real data analysis is included in Supplementary Material, Section S7. The real data are not publicly available due to privacy restrictions. Requests to access these datasets should be directed to alspac-data@bristol.ac.uk.

Declarations

Ethics approval and consent to participate
Ethical approval was obtained from the Avon Longitudinal Study of Parents and Children Ethics and Law Committee and local research ethics committees. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the Avon Longitudinal Study of Parents and Children Ethics and Law Committee at the time.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK. [2]Medical Research Council Integrative Epidemiology Unit at the University of Bristol, University of Bristol, Bristol, UK. [3]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, University of London, London, UK. [4]Medical Research Council Clinical Trials Unit at University College London, University of London, London, UK.

Received: 6 November 2023 Accepted: 26 September 2024
Published online: 07 October 2024

References

1. Carpenter JR, Smuk M. Missing data: a statistical framework for practice. Biom J. 2021;63(5):915–47.
2. Curnow E, Carpenter JR, Heron JE, Cornish RP, Rach S, Didelez V, et al. Multiple imputation of missing data under missing at random: compatible imputation models are not sufficient to avoid bias if they are mis-specified. J Clin Epidemiol. 2023;160:100–9.
3. Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. Stat Methods Med Res. 2015;24(4):462–87.
4. Carpenter JR, Kenward MG, Bartlett JW, Morris TP, Quartagno MW, Angela M. Multiple Imputation and its Application 2e. Chichester: Wiley; 2023.
5. Rubin DB. Multiple imputation for nonresponse in surveys. New York, USA: Wiley; 1987.
6. Kenward MG, Goetghebeur EJ, Molenberghs G. Sensitivity analysis for incomplete categorical data. Stat Model. 2001;1(1):31–48.
7. Tompsett DM, Leacy F, Moreno-Betancur M, Heron J, White IR. On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. Stat Med. 2018;37(15):2338–53.
8. Cornish R, Macleod J, Carpenter J, Tilling K. Multiple imputation using linked proxy outcome data resulted in important bias reduction and efficiency gains: a simulation study. Emerg Themes Epidemiol. 2017;14(14):1–13.
9. Hughes R, Heron J, Sterne J, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int J Epidemiol. 2019;48:1294–304.
10. Collins LM, Schafer JL, Kam C-M. A Comparison of Inclusive and Restrictive Strategies in modern missing data procedures. Psychol Methods. 2001;6(4):330–51.
11. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol. 2017;9(null):157–66.
12. Enders CK. Multiple imputation as a flexible tool for missing data handling in clinical research. Behav Res Ther. 2017;98:4–18.
13. Steiner PM, Kim Y. The mechanics of omitted variable bias: bias amplification and cancellation of offsetting biases. J Causal Inference. 2016;4(2):20160009.
14. Cornish RP, Tilling K, Boyd A, Davies A, Macleod J. Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years. Int J Epidemiol. 2015;44(3):937–45.
15. Thoemmes F, Rose N. A cautious note on auxiliary variables that can increase bias in missing data problems. Multivar Behav Res. 2014;49(5):443–59.
16. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. BMC Med Res Methodol. 2014;14:75.
17. Vansteelandt S, Carpenter JR, Kenward MG. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. Methodology. 2010;6(1):37–48.
18. Curnow E, Tilling K, Heron JE, Cornish RP, Carpenter JR. Multiple imputation of missing data under missing at random: including a collider as an auxiliary variable in the imputation model can induce bias. Front Epidemiol. 2023;3:1237447.
19. Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. New York, USA: Cambridge University Press; 2006.
20. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort profile: the 'children of the 90s'; the index offspring of the Avon Longitudinal Study of Parents and Children (ALSPAC). Int J Epidemiol. 2013;42(1):111–27.
21. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, et al. Cohort profile: the avon longitudinal study of parents and children: ALSPAC mothers cohort. Int J Epidemiol. 2013;42:97–110.
22. Breslau N, Paneth N, Lucia VC, Paneth-Pollak R. Maternal smoking during pregnancy and offspring IQ. Int J Epidemiol. 2005;34(5):1047–53.
23. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res. 2007;16:219–42.
24. Daniel RM, Kenward MG, Cousens SN, Stavola BLD. Using causal diagrams to guide analysis in missing data problems. Stat Methods Med Res. 2012;21(3):243–56.
25. Lee KJ, Carlin JB, Simpson JA, Moreno-Betancur M. Assumptions and analysis planning in studies with missing data in multiple variables: moving beyond the MCAR/MAR/MNAR classification. Int J Epidemiol. 2023;52(4):1268–75.
26. Liu J, Gelman A, Hill J, Su YS, Kropko J. On the stationary distribution of iterative imputations. Biometrika. 2013;101(1):155–73.

27. Hughes RA, White IR, Seaman SR, Carpenter JR, Tilling K, Sterne JAC. Joint modelling rationale for chained equations. BMC Med Res Methodol. 2014;14:28+.
28. Seaman S, Galati J, Jackson D, Carlin J. What is meant by "missing at random"? Stat Sci. 2013;28(2):257–68, 12.
29. Ding P, Vanderweele TJ, Robins JM. Instrumental variables as bias amplifiers with general outcome and confounding. Biometrika. 2017;104(2):291–302.

Publisher's Note