Complete case logistic regression with a dichotomised continuous outcome: a simulation study

Rosaleen Peggy Cornish (✉ rosie.comish@bristol.ac.uk)
University of Bristol

Jonathan William Bartlett
University of Bath

John Macleod
University of Bristol

Kate Tilling
University of Bristol

Research Article

Keywords: Logistic regression, complete case analysis, missing data, multiple imputation, ALSPAC, data linkage, proxy

Posted Date: September 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-911187/v1

License: ☑️ ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Complete case logistic regression with a dichotomised continuous outcome: a simulation study

Rosaleen Peggy Cornish*1,2, Jonathan William Bartlett3,4, John Macleod1,3, Kate Tilling1,2.

*Corresponding author

1. Population Health Sciences, Bristol Medical School, University of Bristol, Oakfield House, Oakfield Grove, Bristol, BS8 2BN.
2. MRC Integrative Epidemiology Unit, University of Bristol.
3. Department of Mathematical Sciences, Faculty of Science, University of Bath, BA2 7AY.
4. NIHR ARC West, 9th Floor, Whitefriars, Lewins Mead, Bristol, BS1 2NT.
Abstract

A complete case logistic regression will give a biased estimate of the exposure odds ratio only if there is a multiplicative interaction between the exposure and outcome with respect to the probability of missingness – whereas linear regression with a continuous outcome is biased in more scenarios, including when only the outcome causes missingness. It is not clear whether a complete case logistic regression will give a biased estimate of the odds ratio if missingness depends on a continuous outcome but this outcome is dichotomised for the analysis – a common situation in epidemiology.

Methods

We investigated this using a simulation study and data from the Avon Longitudinal Study of Parents and Children (ALSPAC), a UK birth cohort. We also examined whether any bias could be reduced by including a proxy for the binary outcome as an auxiliary variable in multiple imputation.

Results

There was negligible bias in the exposure odds ratio when the probability of being a complete case was independently associated with the exposure and (continuous) outcome but important bias in the presence of an interaction, particularly at high levels of missing data. Inclusion of the proxy led to significant bias reductions when this had high sensitivity and specificity in relation to the study outcome.

Conclusions

The robustness of logistic regression to missing data is maintained even when the outcome is a binary version of a continuous outcome. Bias due to an interaction between the exposure and outcome in their effect on selection could be reduced by including proxies for
the missing outcome as auxiliary variables in MI. If such proxies are available, we would recommend using MI over a complete case analysis because, in practice, it would be difficult to rule out an interaction.

Keywords: Logistic regression, complete case analysis, missing data, multiple imputation, ALSPAC, data linkage, proxy.

Background

Epidemiological studies often suffer from missing data arising through non-response. This results in a loss of power and may induce bias. One of the most common approaches to addressing missing data is to carry out a complete case analysis, in which the analysis is restricted to individuals with complete data on all variables in the analysis model.

As highlighted previously, a complete case logistic regression will produce asymptotically unbiased estimates of the exposure odds ratio (OR) as long as the probability of being a complete case depends multiplicatively on the outcome and the exposure [1]. In other words, if R is the response indicator (such that R=1 for complete cases and R=0 otherwise) and Y, X, and C are the outcome, exposure and confounders, respectively, the complete case exposure OR is asymptotically unbiased provided \( P(R=1|Y,X,C) = f(X,C) \times g(Y,C) \) for some functions \( f(X,C) \) and \( g(Y,C) \).

However, although many outcomes in medicine are binary (for example, diagnosed with Type II diabetes), diagnoses are often based on one or more underlying continuous measures whereby, for example, a person is diagnosed if they exceed a defined threshold.
Many diseases are not simply present or absent and, as such, the binary measure is a crude version of an underlying continuous outcome. For example, although tools exist to determine whether an individual meets a diagnostic threshold for depression, it is widely acknowledged that mental disorders such as depression are best measured on a continuum [2]. Although such outcomes are often treated as binary variables in epidemiological studies, it is likely that missingness – if related to the outcome – would vary across levels of the underlying continuous measure (for example, symptom severity) rather than being only associated with whether an individual meets the diagnostic threshold. It is not clear to what extent this might bias estimates of the exposure OR obtained using a complete case analysis.

In the current study, based on data from the Avon Longitudinal Study of Parents and Children (ALSPAC), we examine the association between smoking in pregnancy and offspring depression using a complete case logistic regression. We also use measures of depression derived from linked general practitioner (GP) data to explore the likely missing data mechanism and as auxiliary variables in multiple imputation (MI) and compare the estimate obtained from MI to that obtained in the complete case analysis. We then present results from a simulation study, based on this example, examining bias in the log OR when missingness in the (binary) outcome is associated with an underlying continuous measure. Finally, we explore conditions under which using an imperfect proxy of the outcome as an auxiliary variable in MI reduces (or increases) bias. We vary the sensitivity of the proxy in terms of predicting the study outcome, the proportion of missing data, and the extent to which the outcome is missing not at random (MNAR).

Methods
ALSPAC provided the motivating example for the simulation study. ALSPAC is a birth cohort which recruited c14,500 pregnant women living in and around Bristol, a city in south west England, in the early 1990s. Detailed data were collected during pregnancy and the offspring have been followed up since birth. Further details are given elsewhere [3]. ALSPAC has a searchable data dictionary and variable search tool [4].

Linkage to GP data

In ALSPAC, informed parental consent was mandatory until age 16. When the children reached legal adulthood, they were sent 'fair processing' materials describing ALSPAC’s intended use of their health and administrative records and were given means to consent or object. Data were not extracted for participants who objected, or who were not sent fair processing materials. Linkage to GP data is described in Additional file 1.

Analysis of ALSPAC data

The outcome was a binary measure of depression (meets ICD-10 criteria for a diagnosis, yes/no, derived from the revised Clinical Interview Schedule (CIS-R) [5], completed during a study clinic attended when participants were 18 years. Note that the CIS-R also can be used to generate a (numerical) depression score, but this was not used in our ALSPAC analysis. The exposure was smoking in pregnancy (yes/no); this was based on questionnaire data collected during pregnancy and shortly after birth. We adjusted for the following confounders measured during pregnancy: maternal age, parity and educational level, maternal and paternal antenatal anxiety and depression, family occupational social class, housing tenure (home owned/mortgaged, privately rented, rented from the local council or...
a housing association), and number of rooms in the home (excluding bathrooms). We also
adjusted for sex because of its strong association with the outcome.

From the linked GP data we derived three binary measures of depression: whether or not an
individual had a (1) current, (2) historical, or (3) future record of a diagnosis, symptoms or
treatment for depression (henceforth referred to as current, historical or future depression).

These measures, and their association with the binary depression indicator defined using the
CIS-R, have been described previously [6]. Briefly, current refers to the period 6 months
either side of the month in which the CIS-R was completed, historical refers to any time prior
to this period, and future any time following this period.

Logistic and log-link binomial regression were used to examine associations with missingness
in ALSPAC-measured depression. Logistic regression – using a complete case analysis and MI
using chained equations – was used to examine the association between smoking in
pregnancy and offspring depression. In addition to all the variables included in the
substantive model (smoking in pregnancy, binary CIS-R depression status and covariates
described above), the MI models included the following auxiliary variables: the three
measures of depression derived from GP data and whether the mother had ever smoked
(collected via questionnaire in early pregnancy but referring to lifetime smoking). Stata’s mi
impute chained command was used to carry out the imputations; 100 datasets were
imputed with a burn-in of 20 iterations.

Simulation study
Simulated datasets. We first simulated complete datasets of 10,000 observations (to approximately match the numbers in ALSPAC with complete baseline covariates). Missing data were then simulated in a separate process. The variables simulated were analogous to: depression (binary outcome), a numerical depression score, smoking in pregnancy (exposure) and current GP-recorded depression (linked proxy). For simplicity, we did not simulate covariates. The variables were simulated such that their marginal distributions – and relationships between them – were similar to those observed in ALSPAC. Smoking in pregnancy was simulated with probability 0.25 of having smoked. The continuous depression score for individual \(i\) was simulated as a standard normal variable (normal with mean 0, variance 1) dependent on smoking such that:

\[
\text{Depression score}_i = \mu + \omega \times (\text{smoke}_\text{preg}_i) + \varepsilon_i
\] (1)

where \(\text{smoke}_\text{preg}\) is smoking in pregnancy, coded 0/1, and \(\varepsilon\) is the random error, following a normal distribution with mean 0 and variance calculated to give the score a variance of 1.

The binary depression measure was created using a logistic function (Equation 2):

\[
\text{p}
\_]\_\text{deps}_i = \frac{1}{1 + \exp(\varphi \times (\text{depression score}_i - \pi))}
\] (2)

with \(\varphi\) and \(\pi\) chosen using trial and error to give prevalences of 7.5% and 15%, and where \(\text{p}
\_]\_\text{deps}_i\) represents the probability that an individual was classified as having depression (in the study data). Note that the values of \(\varphi\) were similar (6.1 and 6.5 for prevalences of 7.5% and 15%, respectively), so the strength of association between the continuous and binary
measure of depression was also similar in these two scenarios. A logistic function was used rather than dichotomising the continuous outcome so that the analysis (and imputation) model was correctly specified. Thus, the depression score was a linear function of the exposure and the binary depression measure was a logistic function of the depression score, which correctly implies that the binary depression measure could be written as a logistic function involving the exposure. The parameters in Equation 2 were such that this logistic function was very steep (Supplementary Figure S1, additional file 2). As a sensitivity analysis we repeated a subset of the simulations but this time dichotomising the continuous outcome at a threshold, as might be done in practice (e.g. for defining hypertension).

The analysis model is given by Equation 3.

\[
\text{Logit}(p_{\text{dep}_i}) = \beta_0 + \beta_1 \times (\text{smoke}_{\text{preg}_i})
\]  

(3)

The regression coefficient \(\omega\) for smoking in pregnancy from Equation 1 was set to give an OR for depression of 1.5 (comparing smokers to non-smokers).

For the purposes of this analysis, the study measure of depression was taken as the reference standard. Thus, the linked (binary) GP measure of depression was created – using a logistic function – to give different sensitivities and specificities in relation to the study’s binary measure (Equation 4).

\[
p_{\text{GPdep}_i} = \frac{1}{1 + \exp(\rho \times (\text{depression score}_i - \theta))}
\]  

(4)
Values of $\rho$ and $\theta$ were chosen (using trial and error) to give sensitivities of 0.25 and 0.75 and a specificity of 97.5%.

Generating the missing data. We created missing data only in the outcome, (both the continuous and binary version); this was simulated as MNAR in two ways:

(i) Missingness probability dependent multiplicatively on the exposure and continuous outcome but not their interaction

Two different probabilities were generated using logistic regression (Equations 5 and 6):

\[
\text{logit}(p_{1i}) = \alpha_1 + \delta \times \text{depression score}_i
\]  
\[
\text{logit}(p_{2i}) = \alpha_2 + \gamma \times \text{smok_preg}_i
\]

From these, we created two Bernoulli random variables $R_1$ and $R_2$ (with probabilities $p_1$ and $p_2$). The outcome was classified as being observed if both $R_1$ and $R_2$ were equal to 1 – and missing otherwise. The values of $\alpha_1$ and $\alpha_2$ were chosen using trial and improvement to give specific percentages of missing data and the values of $\delta$ chosen to vary the degree to which the outcome was MNAR. $\gamma$ was fixed as $\ln(0.75)$.

(ii) Missingness dependent multiplicatively on the exposure, continuous outcome and their interaction

The probabilities were generated from the logistic model shown in Equation 7, so that the logarithm of the probability of missingness depended on exposure, outcome and their interaction. As above, the values of $\alpha$ were chosen to produce given percentages of missing data. In these scenarios with an interaction, $\delta$, $\gamma$ and $\tau$ were fixed at $\ln(0.9)$, $\ln(0.7)$ and $\ln(1.1)$, respectively. Note that this interaction on the logit scale implies a multiplicative
interaction between the exposure and outcome with respect to the probability of missingness, such that a complete case analysis is not expected to be (asymptotically) unbiased.

\[
\text{logit}(P(\text{observed}_i)) = \alpha + \delta \times \text{depression score}_i + \gamma \times \text{smoke}_i + \tau 
\times \text{depression score}_i \times \text{smoke}_i
\] (7)

Scenarios investigated. The following six factors were varied in the simulations.

Factor 1: (2 levels): Outcome (depression) prevalence: 7.5%, 15%

Factor 2 (4 levels): Percent missing outcome data: 20%, 40%, 60%, 80%.

Factor 3 (2 levels): Degree to which outcome was MNAR: OR for observing depression for a one SD increase in depression score = 0.9, 0.75 (\(\delta\) from Equation (5)).

Factor 4 (2 levels): Sensitivity of GP depression measure in determining study binary depression = 25%, 75%.

Factor 5 (2 levels): Interaction (yes/no) between exposure and continuous outcome with respect to the log probability of missingness (described above).

Factor 6 (2 levels): Percent missingness in linked GP depression measure (0% or 25%).

In the scenarios without the interaction, we simulated every possible combination of Factors 1-4 (32 scenarios). In the set that included the interaction (Factor 5) we varied Factors 1, 2 and 4 (16 scenarios). Finally, in the set of scenarios where 25% missingness was introduced in the linked variable, only Factor 2 was varied; the other factors were fixed: interaction
present, prevalence of exposure 15% and sensitivity of GP measure 75%. For each scenario we simulated 1,000 datasets.

Statistical analysis. We estimated the log OR for depression on smoking in pregnancy using logistic regression. We used both a complete case analysis and MI – in which the missing (binary) study measure of depression was imputed using logistic regression from the exposure and the GP measure of depression. For each simulated dataset, 100 datasets were imputed.

The estimates obtained from these analyses were compared to the true log OR. The bias was estimated as $\overline{\ln OR} - \ln OR$, where $\overline{\ln OR}$ was the estimated log OR averaged over the 1000 simulated datasets. This was converted to percentage bias. We also calculated the mean squared error (MSE) and the empirical standard error, the standard deviation of the point estimates for the log OR. For MI, we also calculated the fraction of missing information (FMI) and the percent increase in precision compared to the complete case analysis, given by the variance of the log OR obtained using a complete case analysis divided by the variance obtained from MI.

The simulations and all data analysis were carried out in Stata 15.0.

Results

Analysis of ALSPAC data

There were 14,684 enrolled participants alive at one year who had not subsequently withdrawn from the study. Of these, ALSPAC had no National Health Service (NHS) number
for 23 individuals and 95 explicitly dissented to linkage to their health records. This analysis is based on the remaining 14,566. Of these, 11,227 (77%) had data on smoking in pregnancy, 4,537 (31%) had depression data, and 2,718 (19%) were complete cases (individuals with data on smoking in pregnancy, depression, and covariates, but not necessarily linked GP data). In addition, among the 14,566 individuals, 10,560 (72%) had sufficient GP data to generate at least one depression measure. Further details of the available data are given in Table 1. Complete cases were more likely to be female, have a mother who was nulliparous, older, and who did not smoke during pregnancy; higher socio-economic position (measured by maternal education and other factors) was also associated with being a complete case (Supplementary Table S1). In contrast, characteristics of those with GP data were similar to those among all individuals (Supplementary Table S1).

Table 1: Completeness of ALSPAC data by availability of GP data

<table>
<thead>
<tr>
<th>Complete data on:</th>
<th>Linked GP data</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Smoking status in pregnancy</td>
<td>Depression status (CIS-R)</td>
</tr>
<tr>
<td>Covariates</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>2,201</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2,923</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>280</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>830</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2,196</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>478</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>1,472</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10,560</td>
</tr>
</tbody>
</table>

ALSPAC: Avon Longitudinal Study of Parents and Children; CIS-R: Revised Clinical Interview Schedule; GP: General Practitioner.

Association between ALSPAC-measured and GP-recorded depression. Table 2 shows the relationship between the three GP depression outcomes and CIS-R-defined depression. Most
individuals (98%) without depression according to the CIS-R did not have a current GP record for depression. However, only just over a quarter of individuals with CIS-R-measured depression had a current GP depression record (Table 2). The results were similar for historical depression. Future depression had a higher sensitivity but lower specificity.

Table 2: ALSPAC depression according to GP measures of depression

<table>
<thead>
<tr>
<th>GP measurea</th>
<th>CIS-R diagnosis of depression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Current diagnosis or symptoms or treatment</td>
<td>3012 (97.7%)</td>
</tr>
<tr>
<td>Future diagnosis or symptoms or treatment</td>
<td>2500 (79.6%)</td>
</tr>
<tr>
<td>Historical diagnosis or symptoms or treatment</td>
<td>3233 (96.2%)</td>
</tr>
</tbody>
</table>

ALSPAC: Avon Longitudinal Study of Parents and Children; CIS-R: Revised Clinical Interview Schedule; GP: General Practitioner

aThe denominators vary because the numbers with historical, current and future data on depression are different.

After mutual adjustment (for the other GP depression variables), the GP measures were all strongly associated with CIS-R-defined depression: OR = 5.04; 95% CI (3.11, 8.17); 3.14 (2.37, 4.17); and 2.31 (1.44, 3.69), for current, future and historical depression, respectively.

Predictors of missing ALSPAC-measured depression data. Supplementary Table S2 shows the association between covariates and missingness in CIS-R depression. Since the majority of missing data was in the outcome (depression), these factors were the same as those associated with being a complete case. Table 3 shows the associations between the GP measures of depression and missing CIS-R depression among those with complete data on smoking in pregnancy, covariates, and GP data (n=4,468). Using logistic regression, and after adjusting for covariates (including the exposure, smoking in pregnancy), individuals with a
future depression record were more likely to have missing CIS-R depression data; the
association was weaker with current and historical depression. This suggests that the
outcome, depression, is likely to be MNAR conditional on the exposure and covariates; the
addition of the auxiliary variables (GP-recorded depression) should give a better
approximation to missing at random (MAR).

Table 3: Association between GP-recorded depression and missingness in ALSPAC depression

<table>
<thead>
<tr>
<th>Variable</th>
<th>Odds ratio (OR) (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Historical diagnosis or symptoms or treatment</td>
<td>1.13 (0.87, 1.47)</td>
<td>P=0.4</td>
</tr>
<tr>
<td>Current diagnosis or symptoms or treatment</td>
<td>1.23 (0.90, 1.69)</td>
<td>P=0.2</td>
</tr>
<tr>
<td>Future diagnosis or symptoms or treatment</td>
<td>1.32 (1.14, 1.53)</td>
<td>P&lt;0.001</td>
</tr>
</tbody>
</table>

ALSPAC: Avon Longitudinal Study of Parents and Children; GP: General Practitioner

*AAdjusted for all covariates, including smoking in pregnancy (the exposure)

There was no evidence for an interaction between smoking in pregnancy and current GP-
recorded depression with respect to missingness in CIS-R-measured depression [risk ratio
(RR) for interaction between smoking in pregnancy and current depression = 0.97 (0.58,
1.61), P=0.9; and RR for interaction with future depression = 0.88 (0.69, 1.13), P=0.3, when
added to a binomial regression model including a restricted set of covariates (sex, mother’s
education, mother’s age, parity, housing, social class, and number of rooms)]. These
covariates were selected on the basis of their strength of association with missing
depression data; only a restricted set of covariates could be included because models
including additional covariates did not converge. Based on this, there is no evidence to reject
the condition needed to ensure an unbiased estimate of the OR from the complete case

logistic regression, apart from the fact that this was not the CIS-R measure of depression but
its proxy, GP-recorded depression. However, we note the confidence intervals for these interactions are quite wide. In the multiply imputed data, a log-link regression for missingness in CIS-R depression showed no evidence for an interaction between smoking in pregnancy and CIS-R depression (RR for interaction = 0.97 (0.65, 1.44), \(P=0.9\)). Thus, under an assumption that CIS-R depression is MAR given the covariates and the GP depression variables, there was again no evidence to reject the assumption required for unbiasedness of the complete case OR estimate.

Relationship between smoking in pregnancy and offspring depression. Table 4 gives the ORs for depression comparing offspring of mothers who smoked during pregnancy to offspring of non-smokers obtained using the complete case analysis and MI. The complete case estimates were closer to the null than those obtained using MI; MI resulted in increased precision.

<table>
<thead>
<tr>
<th>Analysis approach</th>
<th>Crude OR (95% CI)</th>
<th>Adjusted(^a) OR (95% CI)</th>
<th>Gain in precision(^b) (adjusted log OR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete case (n=2,718)</td>
<td>1.72 (1.20, 2.46)</td>
<td>1.36 (0.92, 2.12)</td>
<td>n/a</td>
</tr>
<tr>
<td>MI(^c) (n=14,566)</td>
<td>1.86 (1.44, 2.40)</td>
<td>1.46 (1.06, 2.01)</td>
<td>287%</td>
</tr>
</tbody>
</table>

MI: Multiple imputation; OR: odds ratio
\(^a\)Adjusted for sex, mothers age, parity, & education, family occupational social class, maternal and paternal antenatal depression and anxiety, housing tenure and number of rooms in house.
\(^b\)Variance (log OR) from complete case analysis / variance (log OR) from MI, expressed as a percentage decrease/increase.
\(^c\)Including linked GP data and ever smoked (from ALSPAC) as auxiliary variables.

Simulation study results
Figure 1 and Supplementary Table S3 show the percent bias in the log OR (error bars in Figure 1 are 1.96 times the Monte Carlo error) when the log probability of missingness depended linearly on the exposure and a function of the outcome (i.e. no multiplicative interaction with respect to the probability of missingness). Similarly, Figure 2 and Supplementary Table S4 show the percent bias when there was an interaction between the exposure and continuous outcome in the missingness model.

Figures 1 and 2 about here

**Complete case analysis.** There was no evidence for bias in the complete case estimate of the log OR when the log probability of missingness depended linearly on the exposure and a function of the continuous outcome (Figure 1 and Supplementary Table S3). This was also the case when the binary outcome was obtained by dichotomising the continuous outcome (Supplementary Table S5). When an interaction between the exposure and outcome with respect to the log probability of missingness was introduced (simulated by including an interaction term in the logistic model for missingness), the bias increased, ranging from 7-8% when 20% of the outcome data were missing to around 35% when 80% were missing. This bias (and the Monte Carlo error) was marginally lower when the prevalence of the exposure was 15% compared to when it was 7.5% (Figure 2).

**Imputation of binary outcome.** In scenarios with no interaction, imputing the binary outcome gave estimates that were similar – on average across the simulated datasets – to the complete case estimates; a stronger proxy (greater sensitivity) generally resulted in slightly lower bias than with a weaker proxy (Figure 1 and Supplementary Table S3). When the
interaction was introduced, there were small reductions in bias (compared to the complete case analysis) when the proxy had low sensitivity but greater reductions when it had a higher sensitivity; precision was also increased (Figure 2 and Supplementary Table S4). Similarly, for a given percentage of missing data, the FMI was lowest when the proxy had high sensitivity (Supplementary Tables S3 and S4).

**Missingness in the linked outcome.** When missingness was introduced in the linked outcome, the reductions in bias and gains in precision were lower than those seen in the equivalent scenarios in which the linked outcome was fully observed (Supplementary Tables S6).

**Discussion**

Our simulations suggest that when the log probability of missingness in a binary outcome depends linearly on the exposure and a function of the underlying continuous outcome, then estimates of the log OR for exposure will be subject to little or no bias; in contrast, if there is an interaction between the exposure and outcome in the dependence of the log probability of missingness on these variables, the bias could be substantial. In the absence of an interaction, including an imperfect proxy for the missing binary outcome as an auxiliary variable in MI will result in similar (but more precise) estimates of the log OR compared to the complete case analysis, particularly when the proxy has high sensitivity (and specificity).

If an interaction is present, imputing the binary outcome will lead to relatively large bias reductions if the proxy has high sensitivity (and specificity); otherwise, the bias reductions are likely to be small. Although a standard implementation of MI assumes MAR, in this case we were imputing the outcome using logistic regression; as such, although the intercept would be expected to be estimated with bias, the log OR for exposure would be
approximately unbiased under the same conditions as the complete case analysis. When there is an interaction, the reductions in bias in MI relative to the complete case analysis result from getting closer to MAR.

In the ALSPAC example, we used three (rather than one) linked proxies for the outcome, with sensitivities 22% (historical), 26% (current) and 55% (future depression) and specificities 96%, 98%, and 80%, respectively. These proxies were independently associated with ALSPAC-measured depression so would be better than a single proxy with sensitivity 25% but may not predict ALSPAC-measured depression as accurately as a single proxy with sensitivity 75%. Although there was no evidence for an interaction between the exposure and outcome with respect to the probability of missingness, this analysis used the proxy outcome measure, GP-recorded depression, rather than the CIS-R measure of depression. However, in the imputed data there was no interaction between the exposure and the CIS-R measure of depression in a log link model for missingness. A key difference in the ALSPAC analysis, however, was the fact that most of the covariates – including the exposure – were also partially observed, with some also potentially MNAR. Thus, since the MI estimate of the odds ratio in this example was slightly higher than the complete case estimate, the simulations suggest that the MI estimate could be subject to a small amount of (upward) bias due to a violation of the MAR assumption. Of course, there are likely to be other sources of bias in the estimate - most notably, residual confounding [7]. Similarly, and as discussed by Bartlett et al, the estimate from both the complete case analysis and MI would also be biased if the outcome model were incorrectly specified [1].
In terms of using auxiliary data, our findings are in line with previous research showing that inclusion of auxiliary variables in MI can increase precision and reduce bias as long as the correlation between the auxiliary variable(s) and the variable with missing data is reasonably high [8-10].

Our study has several limitations. In particular, the simulations did not match the data example exactly. Firstly, the ALSPAC example also included covariates, many of which were subject to missing data themselves and predictors of being a complete case, whereas covariates were not included in the simulations. Secondly, in ALSPAC, GP data were not available for all participants. Although the distribution of most characteristics was similar among the subgroup with GP data compared to among the overall sample, individuals living in owned or mortgaged accommodation were more likely than those in rented accommodation to have linked GP data (Supplementary Table S1).

Conclusions

In summary, when a continuous outcome is MNAR a complete case analysis will result in a biased estimate of the exposure-outcome association. Our results suggest that, in contrast, if this outcome is dichotomised or if the underlying continuous outcome is not measured (such that only a binary form is available) and a complete case logistic regression is used, this is likely to produce estimates that are subject to little or no bias if the log probability of missingness is an additive function of the exposure and continuous outcome. Future work could examine whether this result holds with different (non-linear) missingness mechanisms. If this condition does not hold, however, a complete case analysis will result in a biased estimate of the exposure odds ratio. This bias could be reduced by including one or more
proxies for the missing outcome as auxiliary variables in MI. If such proxies are available, we
would recommend using MI over a complete case analysis because, in practice, it would be
difficult to rule out an interaction. Note that it would also be important to carry out
sensitivity analyses to explore the robustness of the findings to any assumptions made about
the missing data mechanism.

List of abbreviations
ALSPAC: Avon Longitudinal Study of Parents and Children
CIS-R: revised clinical interview schedule
GP: general practitioner
MAR: missing at random
MI: multiple imputation
MNAR: missing not at random
NHS: National Health Service
OR: odds ratio
RR: risk ratio

Figure titles
Figure 1: Mean percent bias in log odds ratio when the log probability of missingness
depended linearly on exposure and a function of outcome
Footnote: Error bars are 1.96 x the Monte Carlo error
Figure 2: Mean percent bias in log odds ratio when the log probability of missingness
depended on exposure, outcome and their interaction
Footnote: Error bars are 1.96 x the Monte Carlo error
References


Declarations

Ethics approval and consent to participate

Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees (NHS Haydock REC: 10/H1010/70). All procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. Informed consent for the use of questionnaire and clinic data was obtained from participants following recommendations of the ALSPAC Ethics and Law Committee at the time. Study participants who complete questionnaires consent to the use of their data by approved researchers. Up until age 18 an overarching informed parental consent was used to indicate parents were happy for their child (the study participant) to take part in ALSPAC. Consent for data collection and use was implied via the written completion and return of questionnaires. Study participants have the right to withdraw their consent for specific elements of the study, or from the study as a whole, at any time. At age 18, study children were sent ‘fair processing’ materials describing ALSPAC’s intended use of their health and administrative records and were given clear means to consent or object via a written form. Data were not extracted for participants who objected, or who were not sent fair processing materials.

Consent for publication

Not applicable

Availability of data and materials
Due to ALSPAC data access permissions, the authors do not have the authority to share the study data analysed in this study, but any researcher can apply to use ALSPAC data, including the variables used in this investigation. Information about access to ALSPAC data is given on their website: (http://www.bristol.ac.uk/alspac/researchers/access/). The code used to generate the simulated datasets is available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Medical Research Council (MR/L012081). The UK Medical Research Council and the Wellcome Trust (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. Data collection is funded from a range of sources. KT and RC work in the MRC Integrative Epidemiology Unit which receives funding from the UK Medical Research Council and the University of Bristol (MC_UU_00011/3). JM is partly funded by the National Institute for Health Research Collaboration West (NIHR ACR West) at University Hospitals Bristol and Weston NHS Foundation Trust, UK.

Authors’ contributions

RC and KT conceived and designed the study, with input from JW. RC ran the simulations and conducted the analyses. RC, KT and JW interpreted the results. RC wrote the first draft of the manuscript with substantial contributions from KT and JW. RC, KT, JW, and JM revised and edited the manuscript. All authors read and approved the final manuscript.
Acknowledgements

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole Avon Longitudinal Study of Parents And Children (ALSPAC) team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.
Figures

Figure 1

Mean percent bias in log odds ratio when the log probability of missingness depended linearly on exposure and a function of outcome. Footnote: Error bars are 1.96 x the Monte Carlo error.
Figure 2

Mean percent bias in log odds ratio when the log probability of missingness depended on exposure, outcome and their interaction Footnote: Error bars are 1.96 x the Monte Carlo error

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile2.docx
- Additionalfile3.docx