Biometrical Journal **52** (2010) 61, zzz–zzz / DOI: 10.1002/bimj.200100000

Estimating treatment effects with partially observed covariates using outcome regression with missing indicators

Helen A. Blake[*,1,2], **Clémence Leyrat** [1,3], **Kathryn E. Mansfield** [3], **Laurie A. Tomlinson** [3], **James Carpenter** [1,4], and **Elizabeth J. Williamson** [1,5]

[1] Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

[2] Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London, WC1H 9SH, UK

[3] Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

[4] MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, 90 High Holborn, London, WC1V 6LJ, UK

[5] Health Data Research UK, 215 Euston Road, London, NW1 2BE, UK

Received zzz, revised zzz, accepted zzz

Missing data is a common issue in research using observational studies to investigate the effect of treatments on health outcomes. When missingness occurs only in the covariates, a simple approach is to use missing indicators to handle the partially observed covariates. The missing indicator approach has been criticised for giving biased results in outcome regression. However, recent papers have suggested that the missing indicator approach can provide unbiased results in propensity score analysis under certain assumptions. We consider assumptions under which the missing indicator approach can provide valid inferences, namely: (1) no unmeasured confounding within missingness patterns; either (2a) covariate values of patients with missing data were conditionally independent of treatment; or (2b) these values were conditionally independent of outcome; and (3) the outcome model is correctly specified: specifically, the true outcome model does not include interactions between missing indicators and fully observed covariates. We prove that, under the assumptions above, the missing indicator approach with outcome regression can provide unbiased estimates of the average treatment effect. We use a simulation study to investigate the extent of bias in estimates of the treatment effect when the assumptions are violated and we illustrate our findings using data from electronic health records. In conclusion, the missing indicator approach can provide valid inferences for outcome regression, but the plausibility of its assumptions must first be considered carefully.

Key words: average treatment effect; missing confounder data; missing covariate data; missing indicator; outcome regression

1 Introduction

Observational studies are a valuable source of information for research investigating the efficacy and safety of treatments in practice. We focus on scenarios where we want to estimate the effect of treatment on a health outcome. However, a common challenge when using observational data is how to deal with missing data. If not handled appropriately, missing data can lead to bias and a loss of efficiency (Bartlett et al., 2015). When using observational data for research, missing data is often an issue in variables that may be considered as potential confounders, such as smoking status or ethnicity.

The simplest approach for handling partially observed covariates is complete record analysis (also called complete case analysis), where patients with missing data are excluded from analysis. Although complete

*Corresponding author: e-mail: Helen.Blake@lshtm.ac.uk

record analysis can provide unbiased results (Bartlett et al., 2015), this approach will typically lead to a loss of efficiency due to the exclusion of information. Furthermore, if patients with complete records are not representative of the population of interest, results from a complete record analysis may not be generalizable to the population of interest (Little and Rubin, 2002; Pigott, 2001).

A popular alternative missing data method is multiple imputation, where missing values are imputed multiple times with plausible values in order to create multiple 'complete' imputed datasets. After analysing each dataset, the results are combined using Rubin's Rules to obtain an overall treatment effect estimate (Rubin, 1976; Carpenter and Kenward, 2013). Although multiple imputation is very powerful, it can be fairly complex and standard implementation requires the assumption that data are missing at random (i.e. the probability of being missing depends on observed data and, given these, does not depend on unobserved data) (Carpenter and Kenward, 2013; Seaman et al., 2013). The plausibility of the missing at random assumption should be considered when implementing multiple imputation (Sterne et al., 2009). In addition, imputing missing values in standard multiple imputation relies on parametric assumptions (Little and Rubin, 2002), the plausibility of which should also be considered (Nguyen et al., 2017).

Another simple way of dealing with partially observed covariates is to use missing indicators — variables which indicate whether the covariate is missing or observed. For a continuous covariate, missing observations are replaced with a fixed value, say 0, and a missing indicator is added to the analysis model, alongside the continuous variable. For a categorical covariate, the missing indicator approach is equivalent to adding a 'missing' category to the variable.

The use of missing indicators to handle missing covariates in outcome regression has been criticised in the literature for being "ad hoc" (Vach and Blettner, 1991; Greenland and Finkle, 1995), and for giving biased results (Groenwold et al., 2012; Jones, 1996). However, the missing indicator approach is often used to deal with missing covariates (Knol et al., 2010) and has been recommended as a missing data method for propensity score analysis (Stuart, 2010). Related methods, incorporating the last-observation-carried-forward approach, have been studied in the context of non-systematic monitoring of covariates in settings with time-varying treatments (Hernán et al., 2009; Kreif et al., 2018). Furthermore, our recent work in the propensity score context suggests that the missing indicator approach can provide unbiased estimates under certain assumptions (Blake et al., 2019). In propensity score analysis, we want to model the relationship between the covariates and the treatment, whereas in outcome regression, we wish to model the relationship between the covariates and the outcome. So, we need to investigate whether the validity of our findings in the propensity score context also holds for outcome regression. Therefore, in this paper we consider whether our work can be extended to the context of outcome regression.

We begin in Section 2 by describing the basic principles of the missing indicator approach and the assumptions underlying its validity. In Section 3, we prove that the missing indicator approach can give unbiased estimates of the treatment effect in outcome regression and show how our work fits in with the literature. In Section 4, we explore the extent of bias in the estimation of the treatment effect when these assumptions are violated. In Section 5, we apply the missing indicator approach in multivariable outcome regression to an illustrative example. We conclude with a discussion in Section 6.

2 Background

2.1 Notation and potential outcome framework

Let Z be a binary variable indicating treatment allocation (or exposure status, etc. depending on context) and let Y represent the observed outcome variable. In this paper, we will concentrate on missing data in covariates and assume that treatment Z and outcome Y are fully observed, as the missing indicator method does not accommodate missing data on the outcome or exposure.

To enable us to describe the assumptions underlying the missing indicator approach, we refer to the potential outcome framework, developed by Rubin (1974) for causal inference from observational data.

We let $Y(z)$ represent the potential outcome that would be observed if Z was set equal to the value z ($z = 0, 1$).

We focus on a scenario with two confounders: a fully observed confounder C and a partially observed confounder X. The missing indicator R equals 1 if X is observed, and $R = 0$ if X is missing.

The confounder values can be partitioned as $\{X_{obs}, X_{mis}\}$, where X_{obs} is the set of X values that are observed and X_{mis} is the set of missing X values (i.e. X_{mis} contains the true unobserved X values). For each patient with $R = 1$, X_{obs} is equal to X and X_{mis} is empty. For each patient with $R = 0$, $X_{mis} = X$ and X_{obs} is empty. Additionally, we define $X^* = X$ when $R = 1$, and $X^* = 0$ when $R = 0$. Note that an alternative approach would be to define X_{obs} instead as RX (which is equivalent to X^*) and X_{mis} as $(1 - R)X$. However, for the purposes of this paper, we use the X_{obs} and X_{mis} notation, following the literature on which our theory builds (D'Agostino and Rubin, 2000; Mattei, 2009).

Our estimand of interest is the average treatment effect (ATE): $\mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$. To estimate the treatment effect, we make the following standard assumptions for causal inference with complete data: strongly ignorable treatment allocation, no interference, consistency, and positivity.

The Strongly Ignorable Treatment Allocation (SITA) assumption — an important assumption in causal inference using observational data — is that there is no unmeasured confounding (Rosenbaum and Rubin, 1983). In a scenario with two confounders, C and X, the SITA assumption can be written as:

$$Z \perp Y(1), Y(0) | C, X. \tag{1}$$

Under the assumption of no interference, the treatment status of one patient does not affect the potential outcomes of another patient (Höfler, 2005; Hernán and Robins, 2018). Assuming consistency, the observed outcome of a patient is equal to the potential outcome corresponding to the treatment they actually received, i.e. if $Z = z$ then $Y = Y(z)$ (Hernán and Robins, 2018). Finally, under positivity, all patients have a non-zero probability of being assigned to each value of treatment, given their characteristics (Little and Rubin, 2000; Hernán and Robins, 2018)

2.2 The missing indicator approach

The missing indicator approach is a simple method of dealing with partially observed covariates. When using outcome regression, the missing indicator approach allows patients with missing data to be used for the estimation of the treatment effect on the outcome, given covariates.

For a continuous partially observed covariate, the missing indicator approach in outcome regression replaces missing covariate values with some fixed value: the same value (for example, 0) is used for all participants with that covariate missing. Both the modified covariate and the missing indicator R are then included in the analysis model. For a categorical partially observed covariate, the missing indicator approach is equivalent to adding a 'missing' category to the variable. The regression coefficient for treatment can then be used to obtain an estimate of the treatment effect using appropriate transformations (eg. the identity function for linear regression).

For example, using the missing indicator approach for linear regression, the analysis model is $E[Y] = \alpha_0 + \alpha_1 Z + \alpha_2 C + \alpha_3 X^* + \alpha_4 R$, where $X^* = X$ when $R = 1$, and $X^* = 0$ when $R = 0$, and where α_1 is the regresssion coefficient corresponding to our estimate of the ATE.

We note that, in the propensity score context, the missing indicator approach allows patients with missing data to contribute to the estimation of the propensity score (i.e. the probability of receiving treatment, given patient characteristics). So, missing indicators are included in the propensity score model, rather than the outcome model (which only includes treatment allocation and the propensity score as covariates).

2.2.1 Assumptions underlying the missing indicator approach

Our recent work in the context of propensity score analysis has shown that the missing indicator approach relies on four assumptions (Blake et al., 2019). In this paper, we extend this work by investigating whether

these four assumptions also underlie the validity of the missing indicator approach in outcome regression, in order to understand when this approach is appropriate in practice.

The first assumption is that there is no unmeasured confounding within missingness patterns, i.e. within each subgroup of patients who have information recorded on the same variables (Mattei, 2009). We call this the missingness Strongly Ignorable Treatment Allocation (mSITA) assumption, due to the similarity to the SITA assumption (equation (1)). Mathematically:

$$\text{mSITA:} \quad Z \perp Y(z) | C, X, R \quad \text{for } z = 0, 1. \tag{2}$$

We call the second and third assumptions the Conditionally Independent Treatment (CIT) assumption and the Conditionally Independent Outcomes (CIO) assumption, respectively. The CIT assumption is that missing confounder values are conditionally independent of treatment, given the observed confounder values and the missing indicator, while the CIO assumption is that missing confounder values are conditionally independent of the potential outcomes (Mattei, 2009).

$$\text{CIT:} \quad Z \perp X_{mis} | C, X_{obs}, R. \tag{3a}$$

$$\text{CIO:} \quad Y(z) \perp X_{mis} | C, X_{obs}, R \quad \text{for } z = 0, 1. \tag{3b}$$

Note that in scenarios with partially observed confounders, the mSITA, CIT and CIO assumptions replace the SITA assumption with respect to identification of the causal estimand.

The fourth assumption in the propensity score context is that the propensity score model is correctly specified; in particular, we assume that the true propensity score model does not include an interaction between the missing indicator R and the fully observed confounder C (CR interaction). In other words, the effect of the fully observed confounder on treatment allocation is assumed to be the same for all missingness patterns.

The analogue assumption for outcome regression is that the outcome model is correctly specified and, in particular, the true outcome model does not include a CR interaction. The plausibility of this correct specification assumption is context-dependent and can be assessed in the data at hand, allowing the possibility of adapting the model in order to ensure the outcome model is correctly specified.

We can obtain valid inferences from the missing indicator approach in propensity score analysis under the following sufficient assumptions: (i) the mSITA assumption holds; (ii) either the CIT or the CIO assumption holds; and (iii) the propensity score model is correctly specified (Blake et al., 2019). In this paper, we extend this work to the outcome regression context, demonstrating in Section 3 that we can use the missing indicator approach with outcome regression to obtain valid inferences under the following assumptions: (i) the mSITA assumption holds; (ii) either the CIT or the CIO assumption holds; and (iii) the outcome model is correctly specified.

2.2.2 Plausibility of the assumptions underlying the missing indicator approach

In missing data methodology, when deciding if a particular method is appropriate, it is important to consider the way in which data becomes missing, i.e. the missingness mechanism. Rubin's taxonomy (Rubin, 1976) is commonly used to classify data as being missing completely at random, missing at random, or missing not at random (Carpenter and Kenward, 2013; Little and Rubin, 2002).

The plausibility of the assumptions in Section 2.2.1 rely instead on the underlying structure of the data (i.e. the causal associations between variables), rather than the missingness mechanisms (Blake et al., 2019). For example, the CIT and CIO assumptions together mean that the partially observed confounder does not confound the relationship between treatment and outcome when it is missing (Blake et al., 2019). So, either the confounder-treatment relationship is absent in individuals who have missing confounder values or the confounder-outcome relationship is absent in individuals who have missing confounder values. Hence, key violations of the CIT or CIO assumptions occur when the missing confounder values affect treatment allocation or the outcome, respectively.

If we believe that the SITA assumption (i.e. no unmeasured confounding) holds in full data, then the mSITA assumption says that additionally conditioning on missingness patterns does not introduce bias. One key way in which this can be violated is when there are: shared unmeasured common causes between outcome and missingness, and unmeasured common causes between treatment and missingness. This is an example of M-bias, which has been discussed extensively in the literature (Greenland, 2003; Pearl, 2009).

The correct specification assumption would be violated if the effects of fully observed confounders on the outcome varied by missingness pattern. Unlike this parametric assumption, which can be tested in the data, the mSITA, CIT and CIO assumptions are not testable. Instead, researchers should use substantive knowledge of the given clinical setting to determine the plausibility of the mSITA, CIT and CIO assumptions.

The first step to assess the plausibility of these assumptions would be to consider whether it is clinically plausible that X is only a confounder when it is observed. If so, and if key violations of the assumptions can be ruled out, then researchers can construct a causal diagram to represent the underlying structural assumptions for the given clinical setting (Blake et al., 2019). This causal diagram should include the missing indicator R. The next step is to convert this causal diagram to incorporate potential outcomes (Richardson and Robins, 2013; Balke and Pearl, 1994; Shpitser and Pearl, 2007). Then, the d-separation rule – which determines whether variables are conditionally independent given a set of other variables (Pearl, 1995; Richardson and Robins, 2013) – can be applied to the causal diagram to assess whether the mSITA assumption holds. In order to assess the CIT and CIO assumptions, the causal diagram should be restricted to patients with $R = 0$ and modified to reflect why it is plausible that X is only a confounder when it is observed (Blake et al., 2019). The d-separation rule can then be applied to this final causal diagram to assess the CIT and CIO assumptions.

For example, consider a simple scenario with a partially observed confounder X and a fully observed confounder C, where C also has causal effects on both X and R. Further suppose that the X-Z relationship is absent in patients with missing X values. Hence, it is plausible that X is only a confounder when it is observed. Figure 1 shows a causal diagram representing this scenario, constructed in the form of a single world intervention graph in order to incorporate potential outcomes (Richardson and Robins, 2013). Applying the d-separation rule to Figure 1, as previously described (Richardson and Robins, 2013; Blake et al., 2019), we find that Z is conditionally independent of $Y(z)$ given C, X and R. Hence, the mSITA assumption holds in this example. In order to be able to assess the CIT and CIO assumptions, we modify Figure 1, by restricting to patients with $R = 0$ and removing the arrow from X to Z in order to encode the assumption that the X-Z relationship is absent in patients with $R = 0$. Figure 2 shows this modified causal diagram. Applying the d-separation rule to this diagram, we find that the CIT assumption holds and that the CIO assumption is violated. Hence, in this scenario, the mSITA and CIT assumptions hold and the missing indicator approach is considered appropriate.

When there are multiple partially observed confounders, R becomes a vector of the missing indicators, whilst X_{obs} now represents all of the sets of observed confounder values and X_{mis} represents all sets of missing confounder values. Assuming that the missingness of these confounders are not associated with each other or with the other confounders, we can assess the CIT and CIO assumptions for each confounder separately, but whilst conditioning on all sets of observed confounder values and all fully observed confounders. An assumption only holds if it holds for every confounder. Issues may arise if the missing indicator of one confounder changes the missing values of another confounder; however, this seems unlikely. For complex scenarios, we recommend constructing a causal diagram that incorporates all relevant substantive knowledge and a missing indicator for each partially observed confounder, and then using software such as Dagitty (Textor et al., 2011) to assess the plausibility of the assumptions.

3 Unbiased estimation of the average treatment effect

In this section we prove that, under the four assumptions given in Section 2.2.1, the missing indicator approach in outcome regression can give an unbiased estimate of the average treatment effect (ATE). We also explore how this result relates to the findings in the literature that the missing indicator approach gives biased results (Jones, 1996), and how the assumptions relate to prior literature.

The target estimand is: ATE $= \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$. We can rewrite this as:

$$\text{ATE} = \mathrm{E}\big[\mathrm{E}\big(Y(1)|C, X_{obs}, R\big) - \mathrm{E}\big(Y(0)|C, X_{obs}, R\big)\big],$$

which can then be written as:

$$\text{ATE} = \sum \big[\sum y\mathrm{P}\big(Y(1) = y|C, X_{obs}, R\big) - \sum y\mathrm{P}\big(Y(0) = y|C, X_{obs}, R\big)\big]. \tag{4}$$

Below in Section 3.1, we show that if the mSITA assumption holds and either the CIT assumption or the CIO assumption holds, then:

$$\mathrm{E}[Y(z)|C, X_{obs}, R] = \mathrm{E}[Y(z)|Z, C, X_{obs}, R] \quad (\text{for } z = 0, 1). \tag{5}$$

Hence, we can rewrite equation (4) as:

$$\begin{aligned}\text{ATE} &= \sum \big[\sum y\mathrm{P}\big(Y(1) = y|Z, C, X_{obs}, R\big) - \sum y\mathrm{P}\big(Y(0) = y|Z, C, X_{obs}, R\big)\big] \\ &= \mathrm{E}\big[\mathrm{E}\big(Y(1)|Z, C, X_{obs}, R\big) - \mathrm{E}\big(Y(0)|Z, C, X_{obs}, R\big)\big].\end{aligned}$$

Under the consistency assumption (Section 2.1), this is:

$$\text{ATE} = \mathrm{E}\big[\mathrm{E}\big(Y|Z = 1, C, X_{obs}, R\big) - \mathrm{E}\big(Y|Z = 0, C, X_{obs}, R\big)\big]. \tag{6}$$

So, we can model the relationship between the outcome and C, X_{obs}, R in each of the two treatment groups and — assuming that the outcome model is correctly specified — we can substitute estimates of the conditional expectations in equation (6) to obtain an unbiased estimate of the ATE. Thus, under the assumptions given in Section 2.2.1, we can get an unbiased estimate of the treatment effect by modelling the relationship between outcome and treatment, given confounders and the missing indicator.

The missing indicator approach suggests a particular parametric specification of the outcome model, at this stage. In particular, missing indicators are added as main effects only, thereby encoding the assumption that there are no interactions between the missing indicators and fully observed confounders. These parametric modelling assumptions can be assessed using the data at hand, although it is unclear whether such checks are common in practice.

3.1 Proof of equation (5)

We first suppose the mSITA and CIT assumptions hold (equations (2) and (3a), respectively). For $z = 0, 1$, we can write $\mathrm{E}[Y(z)|C, X_{obs}, R]$ (from equation (5)) in summation notation:

$$\begin{aligned}\sum y\mathrm{P}\big(Y(z) = y|C, X_{obs}, R\big) &= \sum\sum y\mathrm{P}\big(Y(z) = y, X_{mis}|C, X_{obs}, R\big) \\ &= \sum\sum y\mathrm{P}\big(Y(z) = y|X_{mis}, C, X_{obs}, R\big)\mathrm{P}(X_{mis}|C, X_{obs}, R).\end{aligned} \tag{7}$$

Under the mSITA assumption, the first probability in equation (7) can be written as $\mathrm{P}\big(Y(z) = y|Z, X_{mis}, C, X_{obs}, R\big)$, and under the CIT assumption, the second probability can be written

as $\mathrm{P}(X_{mis}|Z, C, X_{obs}, R)$. So, for $z = 0, 1$, equation (7) becomes:

$$\begin{aligned}
&\sum\sum y\mathrm{P}\big(Y(z) = y|Z, X_{mis}, C, X_{obs}, R\big)\mathrm{P}(X_{mis}|Z, C, X_{obs}, R) \\
&\quad = \sum\sum y\mathrm{P}\big(Y(z) = y, X_{mis}|Z, C, X_{obs}, R\big) \\
&\quad = \sum y\mathrm{P}\big(Y(z) = y|Z, C, X_{obs}, R\big). \quad \square
\end{aligned}$$

Alternatively, if the mSITA and CIO assumptions (equations (2) and (3b)) hold, for $z = 0, 1$, we write:

$$\sum y\mathrm{P}\big(Y(z) = y|C, X_{obs}, R\big) = \sum y\frac{\mathrm{P}\big(Y(z) = y|C, X_{obs}, R\big)}{\mathrm{P}\big(Z|C, X_{obs}, R\big)}\sum \mathrm{P}\big(Z, X_{mis}|C, X_{obs}, R\big), \tag{8}$$

where the denominator is strictly positive under the positivity assumption.

Now, we can write:

$$\sum \mathrm{P}\big(Z, X_{mis}|C, X_{obs}, R\big) = \sum \mathrm{P}\big(Z|X_{mis}, C, X_{obs}, R\big)\mathrm{P}\big(X_{mis}|C, X_{obs}, R\big). \tag{9}$$

Under the mSITA assumption, the first probability in equation (9) can be written as $\mathrm{P}\big(Z|Y(z) = y, X_{mis}, C, X_{obs}, R\big)$, and under the CIO assumption, the second probability can be written as $\mathrm{P}(X_{mis}|Y(z) = y, C, X_{obs}, R)$. So, for $z = 0, 1$, equation (9) becomes:

$$\begin{aligned}
\sum \mathrm{P}\big(Z, X_{mis}|C, X_{obs}, R\big) &= \sum \mathrm{P}\big(Z|Y(z) = y, X_{mis}, C, X_{obs}, R\big)\mathrm{P}\big(X_{mis}|Y(z) = y, C, X_{obs}, R\big) \\
&= \sum \mathrm{P}\big(Z, X_{mis}|Y(z) = y, C, X_{obs}, R\big) \\
&= \mathrm{P}\big(Z|Y(z) = y, C, X_{obs}, R\big).
\end{aligned}$$

Hence, we can write equation (8) as:

$$\begin{aligned}
&\sum y\frac{\mathrm{P}\big(Y(z) = y|C, X_{obs}, R\big)}{\mathrm{P}\big(Z|C, X_{obs}, R\big)}\mathrm{P}\big(Z|Y(z) = y, C, X_{obs}, R\big) \\
&\quad = \sum y\frac{\mathrm{P}\big(Y(z) = y, Z|C, X_{obs}, R\big)}{\mathrm{P}\big(Z|C, X_{obs}, R\big)} \\
&\quad = \sum y\mathrm{P}\big(Y(z) = y|Z, C, X_{obs}, R\big). \quad \square
\end{aligned}$$

3.2 Connections to prior work on the missing indicator approach

Jones (1996) assumed that the true outcome model is a linear regression model with a fully observed covariate Z, a single partially observed covariate X and independent normal errors ϵ:

$$Y = \beta_0 + \beta_1 Z + \beta_2 X + \epsilon, \tag{10}$$

where ϵ is independent of (Z, X, R). Correspondingly, the missing indicator approach can be represented mathematically as:

$$E[Y] = \gamma_0 + \gamma_1 Z + \gamma_2 X^* + \gamma_3 R. \tag{11}$$

Jones (1996) showed that the least squares estimator of γ_1 is biased for β_1, noting that the least squares estimator is unbiased when the sample covariance of Z and X, for patients missing X, is zero. If the CIT assumption holds, this condition holds, since treatment allocation is independent of the confounder for those patients with missing confounder values.

The true outcome model assumed by Jones (1996) in equation (10) leads to the CIO assumption being violated as the outcome is dependent on the missing confounder values. However, if the CIO assumption does hold, then the true outcome model instead resembles the parametric model corresponding to the missing indicator approach in equation (11) (i.e. the true model is $Y = \beta_0 + \beta_1 Z + \beta_2 X^* + \beta_3 R + \epsilon$), and it is simple to show that the least squares estimator is unbiased.

Hence, our findings are compatible with Jones's findings (1996). We have additionally shown that the missing indicator approach can give unbiased estimates when the mSITA and CIO assumptions hold (regardless of whether the CIT assumption additionally holds).

3.3 Connection to alternative statements of assumptions in the literature

The missing indicator method has been recommended for propensity score analysis (Stuart, 2010), based on work in relation to the missingness pattern approach within propensity score analysis (Mattei, 2009; D'Agostino and Rubin, 2000). This approach involves modelling the propensity score separately for each pattern of missing confounder data and can be thought of as a generalisation of the missing indicator method.

In Section 2.2.1, our statement of the mSITA, CIT and CIO assumptions follows Mattei (2009), who states assumptions sufficient for valid inference for the missingness pattern approach. Our assumptions differ from Mattei (2009) in that our version of the CIO assumption is slightly weaker, and requires the conditional independence statement to hold separately for each potential outcome, rather than jointly for the pair of potential outcomes as in the Mattei (2009) original presentation.

D'Agostino and Rubin (2000) instead provide the following assumption, sufficient for valid inference in the missingness pattern approach:

$$Z \perp (Y(0), Y(1), X_{mis}) | X_{obs}, R. \tag{12}$$

The mSITA and CIT assumptions imply that equation (12) holds. However, mSITA and CIO can hold while equation (12) is violated. Thus Mattei (2009) gives a wider set of assumptions under which the missingness pattern approach provides valid inference.

There are strong connections between the missingness pattern approach and other work exploring non-systematic monitoring of time-varying covariates (Hernán et al., 2009; Kreif et al., 2018). These papers suggest a version of the "no unmeasured confounding assumption" which, in the single time-point exposure setting, can be written as:

$$Z \perp Y(z) | X_{obs}, R. \tag{13}$$

Note that here, we have omitted fully observed confounders in order to simplify our discussion of the connections between alternative statements of assumptions.

If the D'Agostino assumption holds, then assumption (13) holds. Further, if either the mSITA and CIT assumptions hold, or the mSITA and CIO assumptions hold, then assumption (13) holds. Compared to the D'Agostino assumption (12), therefore, the mSITA, CIT and CIO assumptions can be seen as a wider set of assumptions under which variants of missingness-pattern-type approaches can produce valid inference. The statement in equation (13) thus provides the most general of the three sets of assumptions.

Kreif et al. (2018) focus on the scenario where the partially missing (non-systematically monitored) covariate is key to the treatment decision process and thus when the clinician does not have this covariate information, they must rely on the last measurement available. Therefore, in their setting – in contrast to our scenario – the covariate always contributes to the treatment decision, whether as an up-to-date measurement or as the last available measurement. However, both settings lead to a causal structure which satisfies a CIT-type assumption.

Here, we assume that full-data inference is the goal, i.e. if we were able to obtain full data then we would. Kreif et al. (2018), in contrast, treat the monitoring process (which induces the missing covariate

data) as an intrinsic part of the setting, and as an attribute of interest in its own right. In particular, in time-varying settings the optimal treatment combination may depend on the intended monitoring process. This makes the inferential goals of Kreif et al. (2018) quite different to those laid out in the current paper. In particular, the set of assumptions we focus on (mSITA, CIT and CIO) require the investigator to consider the confounding structure in the full data setting and how missingness arises in that setting (mSITA), and then to subsequently explore how this structure may change when missing confounder values are present (CIT/CIO). We have found this two-step process useful in considering plausibility of assumptions in real-life settings. Furthermore, although identification of the ATE can be proved using a modification of assumption (13), it may sometimes be more helpful to separate the relationship between the missing confounder values and treatment allocation from the relationship between the missing confounder values and the potential outcomes by considering the CIT and CIO as two separate sub-assumptions.

All three sets of assumptions make it clear that they are likely to be satisfied in a setting where missing confounder values are unavailable to the individual making the treatment decision and thus do not affect treatment. However, only the first version, with CIT and CIO as two separate sub-assumptions, makes it clear that there is another quite different set of scenarios in which missingness-pattern-type methods may provide valid inference.

4 Simulation Study

In this simulation study, we explored the extent of the bias introduced into the treatment effect estimation when each of the key assumptions is violated. Source code to reproduce the results is available as Supporting Information on the journal's web page (http://onlinelibrary.wiley.com/doi/xxx/suppinfo).

4.1 Data-generating mechanisms

We considered 81 data-generating mechanisms. For each, datasets of sample size $n = 500$ were generated. The data-generating mechanisms differed according to which of the assumptions hold. A factorial design was used to consider all possible combinations of each assumption having no violation, a weak violation or a strong violation.

We let U_Z represent a common cause between treatment and the missing indicator, U_Y represent a common cause between the outcome and the missing indicator, and e represent error in the outcome regression model. We generated U_Z, U_Y and e from independent standard Normal distributions.

Two binary confounders X and C were generated using Binomial distributions: $X \sim \text{Bin}(1, 0.67)$ and $C \sim \text{Bin}(1, 0.58)$. To create missing data in X, we generated a missing indicator $R \sim \text{Bin}(1, \text{P}(R = 0))$, where: $\text{logit}(\text{P}(R = 0)) = -0.5 + 1.48 \cdot U_Z + 1.36 \cdot U_Y$.

We also generated a binary treatment allocation variable $Z \sim \text{Bin}(1, \text{P}(Z = 1))$, where:

$$\text{logit}(\text{P}(Z = 1)) = -1.2 + \alpha U_Z + 1.38XR + \beta X(1 - R) + 2R + 1.69C.$$

The observed proportion of treated patients varied between 62.2% and 86.2%, depending on the data-generating mechanism. We generated a continuous outcome using the regression model:

$$Y = 1 - 2.35Z - 2.2\alpha U_Y - 1.55XR + \gamma X(1 - R) + 1.8R - 1.7C + \delta CR + 3e.$$

where $\alpha \in \{0, 0.125, 1.25\}$, $\beta \in \{0, 0.138, 1.38\}$, $\gamma \in \{0, -0.155, -1.55\}$ and $\delta \in \{0, -0.42, -4.2\}$. If $\alpha = 0$, then the mSITA assumption holds. Similarly, if $\beta = 0$, $\gamma = 0$, or $\delta = 0$, then, respectively, the CIT assumption holds, the CIO assumption holds, or the outcome model is correctly specified. For each parameter, the smaller and larger non-zero values represent, respectively, a weak violation and a strong violation of the corresponding assumption.

Data were simulated using Stata 14.2 with 5000 simulation repetitions per data-generating mechanism.

4.2 Methods

Each simulated data set was analysed using the missing indicator approach with multivariable linear regression, by creating a new version of the partially observed binary covariate with a third 'missing' category. Our estimand is the average treatment effect, as estimated using the treatment coefficient from a linear regression model. Our performance measure of interest is absolute bias of the ATE: $\frac{1}{5000}\sum_{i=1}^{5000}\hat{\theta}_i - \theta$, where $\hat{\theta}_i$ is the estimated treatment effect from the ith repetition, and θ is the true treatment effect.

4.3 Results

In Figure 3, the left-hand panel presents the absolute bias in the estimated treatment effect for eight scenarios, depicting all possible combinations of the mSITA, CIT and CIO assumptions holding or not. The dark bars show scenarios where the mSITA assumption (required for valid inference) holds. The light bars show scenarios where it does not. As expected from our theory above, if the mSITA assumption is violated (light bars), bias is present. The four sets of bars show combinations of the CIT and CIO assumptions holding or not, for scenarios where the mSITA assumption holds (dark bars); bias is present only when both CIT and CIO are violated. The right-hand panel of Figure 3 shows the same eight scenarios, but with a parametric violation of assumptions: the outcome model fitted assumes no interaction between the missingness indicator and the fully observed confounder C, but in truth this interaction does exist. Violation of this parametric assumption leads to bias in all eight scenarios.

Figure 4 shows a number of scenarios in which the outcome model is correctly specified but the other three assumptions (mSITA, CIT and CIO) may be violated. The three panels show – from left to right – increasing levels of violation of the CIO assumption. Within the three panels, the three sets of bars show – from left to right – increasing levels of violation of the CIT assumption. Within each set of bars, the bars show – from left to right – increasing levels of violation of the mSITA assumption. Large bias is seen when either a strong violation of the mSITA assumption is present, or when strong violations of both the CIT and CIO assumptions are present.

Figure 5 shows the same scenarios as Figure 4, but with weak violations of the parametric assumptions (weak CR interactions present but not included in the fitted model) shown in the top panel, and strong violations of the parametric assumptions shown in the bottom. Weak violations of the parametric assumptions induced additional small amounts of bias compared to Figure 4. Strong violations of the parametric assumptions induced large amounts of biases under most settings.

The missing indicator approach gives unbiased estimates of the treatment effect when the mSITA assumption holds, the outcome model is correctly specified and either one, or both, of the CIT and CIO assumptions hold. When both the CIT and CIO assumptions are violated, the missing indicator approach gives biased results, whether or not the other two assumptions hold. The worst bias occurs when both the mSITA assumption and the correct specification assumption is violated and the CIT assumption holds, whether or not the CIO assumption holds. In general, having the mSITA assumption violated results in larger biases for the settings explored in the simulation study. In addition, incorrectly specifying the outcome model, i.e. failing to include an interaction between the fully observed confounder C and the missing indicator R in the true outcome model, generally results in larger biases than when the outcome model is correctly specified.

When the outcome model is correctly specified, weak violations of the other assumptions results in similar biases compared to when the assumptions hold (Figure 4). Similar results were found when considering data-generating mechanisms where the true outcome model includes a weak CR interaction and when considering data-generating mechanisms with a strong CR interaction (Figures 5a and 5b respectively). In general, having a weak CR interaction resulted in similar or larger biases compared to scenarios where the outcome model is correctly specified.

5 Application to illustrative example

5.1 Study description

Our illustrative example is a cohort study using electronic health records data from the UK Clinical Practice Research Datalink and the Hospital Episode Statistics (Mansfield et al., 2016). The cohort study aimed to investigate the association between risk of acute kidney injury (AKI) and use of angiotensin-converting enzyme inhibitors or angiotensin receptor blockers (ACEI/ARBs), compared to other antihypertensive drugs. An important covariate in the study was chronic kidney disease, which was categorised into stages based on a continuous measure of kidney function called the estimated glomerular filtration rate (eGFR). Lower values of eGFR indicate worse kidney function.

Data were obtained for 570 586 new adult users of antihypertensive drugs between 1997 and 2014. Follow-up began at first prescription of ACEI/ARBs, beta blockers, calcium channel blockers, or diuretics. The treatment of interest was prescription of ACEI/ARBs. Our outcome of interest was kidney function within 2 months of first prescription of an antihypertensive drug, as measured using eGFR (Levey et al., 2009). Due to conditions of the data use agreement, we can no longer access the eGFR data after treatment initiation, so we have simulated this variable, based on observed relationships in prior studies (see Appendix for details). As a result, the 'true' treatment effect is known.

In this study there were a number of fully observed potential confounders: age, sex, chronic comorbidities, other antihypertensive or diuretic drugs, and calendar period. In addition, two potential confounders were partially observed: ethnicity, which had 59.0% missing data; and baseline eGFR category, which had 52.9% missing data.

In this example, only 21% of patients had complete data for both ethnicity and baseline eGFR category; the majority of patients records would be discarded, leading to a loss of efficiency, if complete record analysis was used for this example. Furthermore, standard multiple imputation may not be appropriate since the missing at random assumption is questionable: baseline eGFR category is more likely to be measured for patients with worse kidney function. The assumptions underlying the missing indicator approach seem reasonable in this context. First, the mSITA assumption would be violated if there are any unobserved common causes between missingness of baseline eGFR category and treatment allocation or the outcome. In this example, it seems plausible that any such common causes, such as age or chronic comorbidities, are measured and able to be included in the analysis model. In addition, predictors of missingness in ethnicity seem unlikely to also be predictors of prescription decisions. Thus the mSITA assumption seems plausible here.

Second, it is plausible to assume that information about a patient's baseline eGFR category is unlikely to influence the clinician's decision to prescribe if this information is not available to the clinician (eg. if a kidney function test had not been ordered beforehand). In practice, proxy information about a patient's baseline eGFR category may be available to the clinician (but not to researchers using electronic health records). However, this is likely to reflect poor kidney function for only a small proportion of the whole study population. In addition, it is plausible that a clinician would ensure information on patient's ethnicity is recorded if they believe that this information is an important factor in their decision whether or not to prescribe ACEI/ARBs. Thus we believe that the CIT assumption is plausible.

Third, it seems plausible that the effect of the other fully observed risk factors on AKI would not vary according to whether or not ethnicity and baseline eGFR category were measured. Furthermore, this assumption can be tested in the data.

Fourth, the CIO assumption does not seem plausible in this context — since baseline kidney function remains a risk factor for change in eGFR, whether or not baseline eGFR category is measured. Since we can obtain valid inferences from the missing indicator approach when just one of the CIT and CIO assumptions hold (in addition to the mSITA and correct specification assumptions holding), the CIO assumption being violated is not an issue here; the mSITA, CIT and correct specification assumptions seem plausible and thus the missing indicator approach is considered appropriate.

5.2 Method

We applied linear regression, adjusted for ethnicity, baseline eGFR category and fully observed confounders (age category, sex, chronic comorbidities, and calendar period), to obtain estimates of the treatment effect comparing patients prescribed ACEI/ARBs at start of follow-up time exposed to ACEI/ARBs versus patients not prescribed ACEI/ARBs at baseline. To handle missing data in ethnicity and baseline eGFR category, we applied complete record analysis and the missing indicator approach. Analysis was conducted in Stata 14.2.

5.3 Results

Our results are given in Table 1. The missing indicator approach uses all missingness patterns; in addition to the 121 527 patients with complete data, 112 142 patients had missing data for baseline eGFR category, 147 011 have ethnicity missing, and 189 906 had missing data for both. Using the missing indicator approach, the estimated treatment effect was closer to the true treatment effect than the estimate from complete record analysis. In addition, the complete record analysis estimate has a wider confidence interval due to the exclusion of over 75% of the patient records. When interactions between the missing indicators and the fully observed confounders are added into the regression model, the results do not change much compared to the missing indicator approach (-0.6575, 95% CI: [-0.7424, -0.5567]), and so there is no evidence of a violation of the parametric assumption.

6 Discussion

In this paper, we have shown that the missing indicator approach in outcome regression is unbiased when (i) there is no unmeasured confounding within missingness patterns; (ii) either confounder values of patients with missing data are conditionally independent of treatment assignment, or these missing confounder values are conditionally independent of the outcome; and (iii) the effect of fully observed confounders on the outcome is the same for all missingness patterns. We have applied the missing indicator approach to an illustrative example using routinely collected data, a key area in which the method's underlying assumptions may be plausible (Blake et al., 2019).

An advantage of the missing indicator approach for outcome regression is that it is easy to implement and, unlike complete record analysis, avoids discarding much information when the proportion of missing data is large. In addition, the missing indicator approach may be appropriate in situations where multiple imputation is not, as the missing indicator approach does not rely on the conventional classification of missingness mechanisms. Whereas standard implementation of multiple imputation is guaranteed to be valid when data are missing at random, the CIT and CIO assumptions are not about the missingness mechanism, but are rather about whether the partially observed covariate confounds the relationship between treatment and outcome when it is missing. When either the CIT or the CIO assumption holds, the relationships between variables among patients with observed data are not the same as those among patients with missing data, and so multiple imputation may not be appropriate. In contrast, the missing indicator approach may be unbiased under missing not at random mechanisms, and biased under some missing completely at random mechanisms.

The missing indicator approach has been criticised in the missing data methodology literature as being 'ad hoc' (Greenland and Finkle, 1995) and biased (Groenwold et al., 2012; Jones, 1996). We have shown that the missing indicator approach can give unbiased results under certain assumptions. Researchers seeking to use the missing indicator approach should first consider whether these assumptions seem plausible within the context of a given clinical setting, with the help of causal diagrams. In our simulation study, we considered scenarios with a single partially observed variable. Our suggested approach to handling multiple partially observed confounders within the missing indicator framework requires the assumption that the missingness of one confounder does not affect the missing values of another confounder. In practice,

researchers should carefully consider the plausibility of such an assumption, in addition to considering the plausibility of the mSITA, CIT, CIO, and correct specification assumptions. If the assumptions underlying the missing indicator approach are found to not be appropriate, then researchers should consider whether the assumptions underlying complete record analysis or multiple imputation are more appropriate in the given scenario.

The missing indicator approach is a method for handling missing covariate data, but cannot handle missing data on the outcome or treatment allocation. Further work is required to extend the approach to handle other missing data, perhaps by combining with other methods such as multiple imputation. Another limitation of the missing indicator approach, in the context of propensity score analysis, is that estimation issues may arise if there are many missingness patterns and some of these patterns have low sample size. (Qu and Lipkovich, 2009) proposed a pattern-pooling algorithm to ensure sufficient sample size for estimation in propensity score analysis. Further work is needed to explore the impact of low sample size in missingness patterns in the context of outcome regression and whether this impact can be alleviated by using pattern-pooling algorithms. A limitation of our simulation study is that we did not assess the impact of changing the proportion of missing data. However, when the assumptions do not hold, bias is expected to increase with the proportion of missing data. Furthermore, in this paper, we have focused on linear regression. We believe that our theoretical results can be extended to risk difference estimation and Poisson regression; further work is required to confirm this. Careful consideration would be required to translate these results to the odds ratio setting due to non-collapsibility issues.

In conclusion, the missing indicator approach for outcome regression can be applied in a principled way and can give valid results under a particular set of assumptions, but researchers must first consider whether these assumptions seem plausible in the clinical setting of interest. We end by noting that standard application of the missing indicator approach makes rather strong parametric assumptions about absence of interactions between missing indicators and fully observed confounders; we recommend that checking these assumptions in the data at hand should form part of routine practice when applying this approach.

Acknowledgements

HAB was supported by the Economic and Social Research Council [Grant Number ES/J5000/21/1]. CL was supported by the Medical Research Council [Project Grant MR/M013278/1]. LAT was supported by a Wellcome Trust intermediate clinical fellowship [Grant Number 101143/Z/13/Z]. EJW was supportred by Health Data Research UK [Grant Number EPNCZO90], which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. Ethics approval was given by the London School of Hygiene and Tropical Medicine Research Ethics Committee [Reference: 15880] and by the Clinical Practice Research Datalink Independent Scientific Advisory Committee [ISAC Protocol Number 14_208A2].

Conflict of interest statement

The authors have declared no conflict of interest.

Appendix

A. Simulating kidney function for the illustrative example

We simulated a continuous outcome $Y = \mathbf{X}\beta + e$ where $\mathbf{X}$ denotes the design matrix, β represents the vector of regression coefficients and e denotes the vector of error terms, where $e \sim (0, 14.65)$. The

design matrix contains the vector with all entries equal to 1 and the following variables: prescription of ACEI/ARBs at baseline; diabetes mellitus status at baseline; hypertension status at baseline; cardiac failure status at baseline; arrhythmia status at baseline; ischaemic heart disease status at baseline; sex; ageband at baseline; calendar period at baseline; ethnicity; and baseline eGFR category. The regression coefficients are given in Table 2.

References

Balke, A. and Pearl, J. (1994). Probabilistic evaluation of counterfactual queries. pages 230–237. Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence.

Bartlett, J. W., Harel, O., and Carpenter, J. R. (2015). Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *Am J Epidemiol*, 182(8):730–736.

Blake, H. A., Leyrat, C., Mansfield, K. E., Seaman, S., Tomlinson, L. A., Carpenter, J., and Williamson, E. J. (2019). Propensity scores using missingness pattern information: a practical guide. Under review. arXiv preprint arXiv:1901.03981 [stat.ME].

Carpenter, J. and Kenward, M. (2013). *Multiple Imputation and Its Application*. Statistics in Practice. Wiley, Chichester.

D'Agostino, R. B. and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *J Am Stat Assoc*, 95(451):749–759.

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306.

Greenland, S. and Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, 142(12):1255–1264.

Groenwold, R. H., White, I. R., Donders, A. R. T., Carpenter, J. R., Altman, D. G., and Moons, K. G. (2012). Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *Can Med Assoc J*, 184(11):1265–1269.

Hernán, M. and Robins, J. (2018). *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming.

Hernán, M. A., McAdams, M., McGrath, N., Lanoy, E., and Costagliola, D. (2009). Observation plans in longitudinal studies with time-varying treatments. *Statistical Methods in Medical Research*, 18(1):27–52.

Höfler, M. (2005). Causal inference based on counterfactuals. *BMC Med Res Methodol*, 5:28–28.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *J Am Stat Assoc*, 91(433):222–230.

Knol, M. J., Janssen, K. J., Donders, A. R. T., Egberts, A. C., Heerdink, E. R., Grobbee, D. E., Moons, K. G., and Geerlings, M. I. (2010). Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *J Clin Epidemiol*, 63(7):728 – 736.

Kreif, N., Sofrygin, O., Schmittdiel, J., Adams, A., Grant, R., Zhu, Z., van der Laan, M., and Neugebauer, R. (2018). Evaluation of adaptive treatment strategies in an observational study where time-varying covariates are not monitored systematically. arXiv preprint arXiv:1806.11153 [stat.ME].

Levey, A. S., Stevens, L. A., Schmid, C. H., Zhang, Y. L., Castro 3rd, A. F., Feldman, H. I., Kusek, J. W., Eggers, P., Van Lente, F., Greene, T., Coresh, J., and CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) (2009). A new equation to estimate glomerular filtration rate. *Ann Intern Med*, 150(9):604–612.

Little, R. J. and Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annu Rev Public Health*, 21:121–145.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley.

Mansfield, K. E., Nitsch, D., Smeeth, L., Bhaskaran, K., and Tomlinson, L. A. (2016). Prescription of renin–angiotensin system blockers and risk of acute kidney injury: a population-based cohort study. *BMJ Open*, 6(12).

Mattei, A. (2009). Estimating and using propensity score in presence of missing background data: An application to assess the impact of childbearing on wellbeing. *Stat Methods Appl*, 18(2):257–273.

Nguyen, C. D., Carlin, J. B., and Lee, K. J. (2017). Model checking in multiple imputation: an overview and case study. *Emerg Themes Epidemiol*, 14(1):8.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, 2nd edition.

Pigott, T. D. (2001). A review of methods for missing data. *Educ Res Eval*, 7(4):353–383.

Qu, Y. and Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Stat Med*, 28(9):1402–1414.

Richardson, T. and Robins, J. (2013). Technical report 128. Single World Intervention Graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. http://www.csss.washington.edu/Papers/wp128.pdf.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*, 66(5):688–701.

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

Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013). What is meant by "missing at random"? *Statist. Sci.*, 28(2):257–268.

Shpitser, I. and Pearl, J. (2007). What counterfactuals can be tested. pages 352–359. Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338. doi:10.1136/bmj.b2393.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.*, 25(1):1–21.

Textor, J., Hardt, J., and Knüppel, S. (2011). DAGitty: A graphical tool for analyzing causal diagrams. *Epidemiology*, 22:745.

Vach, W. and Blettner, M. (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *Am J Epidemiol*, 134(8):895–907.

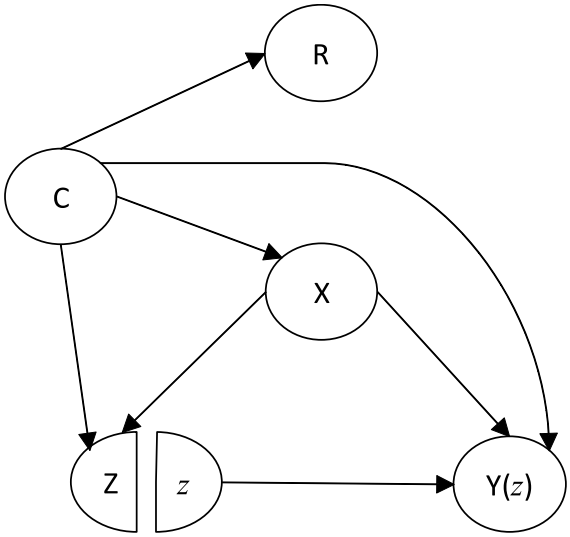

Figure 1 A causal diagram for a simple scenario with a partially observed confounder X and a fully observed confounder C, incorporating the missing indicator R. $Y(z)$ is the potential outcome resulting from intervening to set treatment Z to a particular value z.

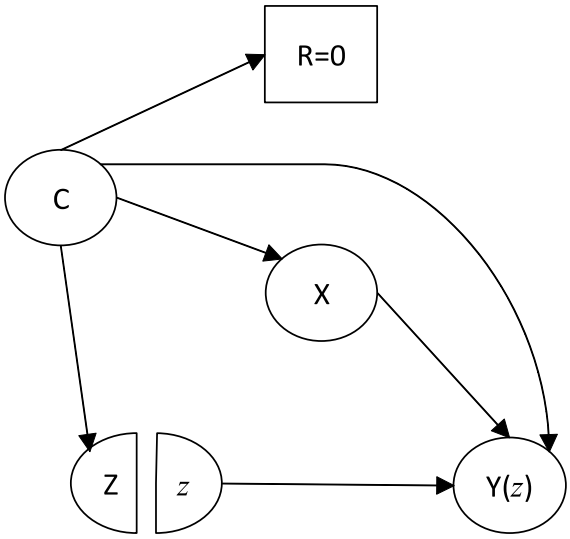

Figure 2 A causal diagram for a simple scenario with a partially observed confounder X and a fully observed confounder C, modified to assess the CIT and CIO assumptions. The square box around $R = 0$ indicates restriction to individuals with missing X values. $Y(z)$ is the potential outcome resulting from intervening to set treatment Z to a particular value z.

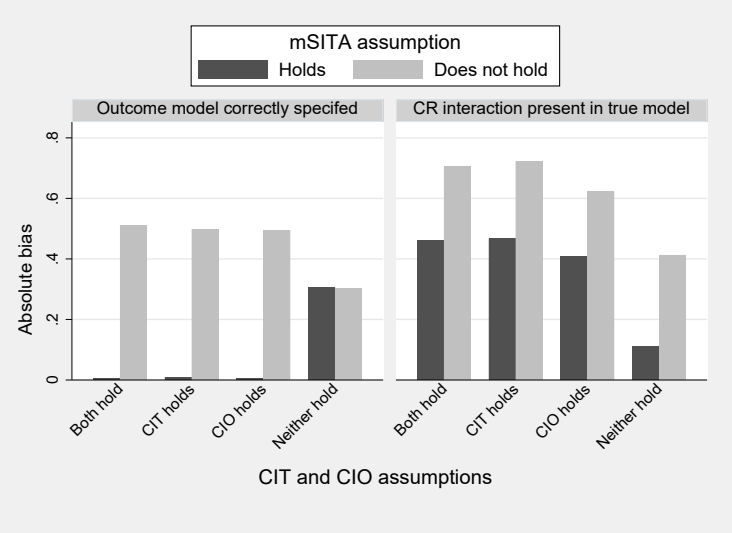

Figure 3 Results from a simulation study showing the absolute bias in the estimated treatment effect when using the missing indicator approach for multiple linear regression under different data-generating mechanisms, which vary according to: (i) whether the mSITA assumption holds; (ii) whether the CIT assumption holds; (iii) whether the CIO assumption holds; and (iv) whether there is an interaction between the fully observed confounder C and the missing indicator R in the true outcome model. True treatment effect: -2.35. Sample size: $n = 500$. Number of replications: 5000.

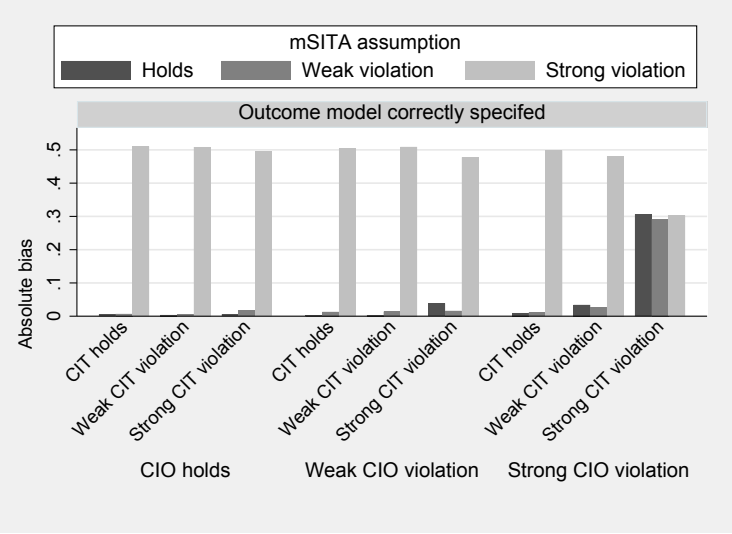

Figure 4 Results from a simulation study showing the absolute bias in the estimated treatment effect when using the missing indicator approach for multiple linear regression under different data-generating mechanisms, which vary according to whether there is no violation, a weak violation or a strong violation of: (i) the mSITA assumption, (ii) the CIT assumption, and (iii) the CIO assumption. For all data-generating mechanisms, the outcome model is correctly specified. True treatment effect: -2.35. Sample size: $n = 500$. Number of replications: 5000.

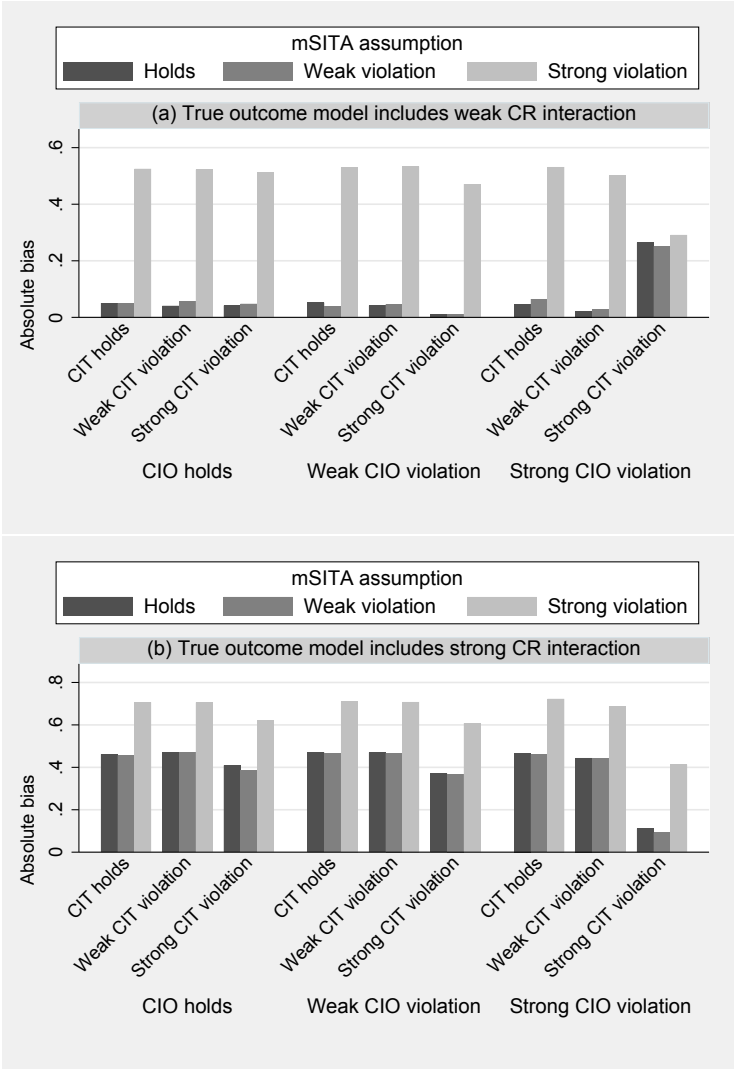

Figure 5 Results from a simulation study showing the absolute bias in the estimated treatment effect when using the missing indicator approach for multiple linear regression under different data-generating mechanisms, which vary according to whether there is no violation, a weak violation or a strong violation of: (i) the mSITA assumption, (ii) the CIT assumption, and (iii) the CIO assumption. For all data-generating mechanisms, the true outcome model contains either a weak interaction (5a) or a strong interaction (5b) between the fully observed confounder C and the missing indicator R. True treatment effect: -2.35. Sample size: $n = 500$. Number of replications: 5000.

Table 1 Estimated treatment effects (mean differences) and 95% confidence intervals (CIs) using linear regression to compare the effect on (simulated) kidney function of being prescribed ACEI/ARBs at start of follow-up versus not begin prescribed ACEI/ARBs at baseline. True treatment effect: -0.6831

Missing data method	Treatment effect (95% CI)	Number of patients analysed
Complete record analysis	-0.6150 $(-0.7977, -0.4324)$	121 527
Missing indicator approach	-0.6496 $(-0.7424, -0.5567)$	570 586

Table 2 Regression coefficients for using baseline characteristics to simulate an outcome variable measuring kidney function within two months of prescription of antihypertensive drugs. ACEI/ARBs: angiotensin-converting enzyme inhibitors or angiotensin receptor blockers. eGFR: estimated glomerular filtration rate.

Coefficient	Variable	Coefficient	Variable
-0.6831	ACEI/ARBs prescription	1.3974	calendar period $2001-2004$
0.4847	diabetes mellitus	2.7825	calendar period $2005-2008$
-5.5041	hypertension	4.2181	calendar period $2009-2011$
-1.9321	cardiac failure	4.9409	calendar period $2012-2014$
-1.6349	arrhythmia	4.1883	ethnicity recorded as south asian
-3.4547	ischaemic heart disease	-2.6490	ethnicity recorded as black
-1.6717	female	2.7238	ethnicity recorded as other
-12.8473	$45 \leq$ age < 55	3.3971	ethnicity recorded as mixed
-17.6097	$55 \leq$ age < 60	0.1647	ethnicity missing
-20.0686	$60 \leq$ age < 65	-36.7126	baseline eGFR < 30
-22.1784	$65 \leq$ age < 70	-25.1941	$30 \leq$ baseline eGFR < 45
-24.1881	$70 \leq$ age < 75	-16.3931	$45 \leq$ baseline eGFR < 60
-26.5288	$75 \leq$ age < 85	-4.4043	baseline eGFR missing
-25.6283	age ≥ 85	94.0335	constant

