

Doubly robust estimators of causal exposure effects with missing data in the outcome, exposure or a confounder[‡]

E. J. Williamson,^{a,b,*†} A. Forbes^a and R. Wolfe^a

We consider the estimation of the causal effect of a binary exposure on a continuous outcome. Confounding and missing data are both likely to occur in practice when observational data are used to estimate this causal effect. In dealing with each of these problems, model misspecification is likely to introduce bias. We present augmented inverse probability weighted (AIPW) estimators that account for both confounding and missing data, with the latter occurring in a single variable only. These estimators have an element of robustness to misspecification of the models used. Our estimators require two models to be specified to deal with confounding and two to deal with missing data. Only one of each of these models needs to be correctly specified. When either the outcome or the exposure of interest is missing, we derive explicit expressions for the AIPW estimator. When a confounder is missing, explicit derivation is complex, so we use a simple algorithm, which can be applied using standard statistical software, to obtain an approximation to the AIPW estimator. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: double-robustness; inverse probability weighting; G-computation; confounding; missing data; observational study

1. Introduction

Missing data is a common problem in medical studies. Many analytic approaches have been proposed for obtaining correct inferences in the presence of missing data. Multiple imputation is an increasingly popular strategy [1]. Inverse probability weighting is an alternative approach, where fully observed data is up-weighted to represent the full data. Li *et al.* [2] and Seaman *et al.* [3] gave comprehensive reviews of weighting methods. Other missing data approaches, including full likelihood methods, can be adopted but have been less frequently used because of the difficulty of extending these methods to general scenarios. A fairly recent approach to missing data combines the imputation and inverse-weighting strategies to form ‘doubly robust’ estimators, so-called because they are consistent if either the imputation model or the model for missingness, but not both, is misspecified [4].

Motivated by increasing the precision of inversely weighted estimators, Robins, Rotnitzky and others [5–7] proposed improved inversely weighted estimators for scenarios with missing covariates and outcomes, approaching the problem within the framework of semiparametric theory. These ideas led to the investigation of the robustness properties of these estimators, and the introduction of ‘doubly robust’ augmented inverse probability weighted (AIPW) estimators [8]. Carpenter *et al.* [9] applied ideas within these papers, comparing doubly robust AIPW estimators to estimators obtained using multiple imputation. Tsiatis [10] further developed the ideas of Robins, Rotnitzky and others and applied them to a variety of missing data scenarios and study designs. Kang and Schafer [4] discussed a variety of doubly robust estimators for scenarios with missing outcome data, including some that fall

^aDepartment of Epidemiology and Preventive Medicine, Monash University, Australia

^bCentre for Molecular, Environmental, Genetic and Analytic Epidemiology, School of Population Health, University of Melbourne, Australia

*Correspondence to: E. J. Williamson, Department of Epidemiology and Preventive Medicine, Monash University, The Alfred Centre, 99 Commercial Road, Melbourne, 3004 Victoria.

†E-mail: ewi@unimelb.edu.au

‡Supporting information may be found in the online version of this article.

outside the class of AIPW estimators, focussing particularly on the properties of these estimators under dual model misspecification. Bang and Robins [11] presented a doubly robust estimator for missing outcome data, which is attractively simple but does not always have ideal statistical properties.

These doubly robust approaches to missing data can also be useful in fully observed data when a treatment or exposure effect is of interest. Each subject is viewed as having two (or more) potential outcomes, one under each possible exposure level; only one of which was actually observed [12]. By viewing the non-observed potential outcomes as missing values, the missing data approaches discussed earlier can be applied. To obtain a doubly robust estimator of exposure effect in fully observed data, two models are required: a model for the outcome conditional on any confounding variables and a model for the exposure given the confounding variables. The latter model is often termed the propensity score model [13]. Lunceford and Davidian [14] and Austin [15] demonstrated that these two models can be combined to give an AIPW estimator that is robust to the misspecification of either the outcome or the propensity score model. Kang and Schafer [4] discussed a variety of doubly robust estimators for this scenario, both AIPW and non-AIPW estimators.

In observational studies, because the form of the statistical models used to estimate exposure effects, missing values or missingness mechanisms are typically uncertain, doubly robust approaches have highly desirable statistical properties. However, the problems of confounding and missing data typically occur simultaneously. Few authors have investigated the application of doubly robust estimators to situations with both confounding and missing data. Davidian *et al* [16] presented a doubly robust AIPW estimator of the causal effect of an exposure in the presence of missing outcome data, in a pre-test–post-test scenario where additional information is available on variables intermediate between exposure and outcome, framing their calculation of the estimator within semiparametric theory.

The objective of this paper, therefore, is to obtain doubly robust estimators for scenarios where either the outcome, the exposure or a confounder is partially missing, to examine their robustness properties and to gain an intuitive understanding of how the robustness properties are achieved. The paper is structured as follows. In Section 2, we review methods for estimating causal effects in the absence of missing data. In Section 3, we extend these methods to situations with missing data in a single variable and draw connections between previously proposed estimators and our approach. We apply our new estimators to a non-randomized comparison by using the data from a randomized trial in anaesthesia in Section 4. Section 5 presents results from a simulation study. We end by discussing our findings in Section 6.

2. Estimating causal effects in fully observed observational data

We begin by considering the situation in which the data are fully observed. There are various ways in which causal effects can be estimated in the presence of confounding. We describe three such methods: G-computation, inverse probability-of-treatment weighting (IPW) and a third approach, which combines these two methods, doubly robust estimation.

2.1. Assumptions and notation

We consider a typical cohort study, with a continuous outcome Y , a binary exposure of interest X taking values 0 (unexposed) and 1 (exposed) and measured confounders $\mathbf{Z} = (Z_1, \dots, Z_p)$. The confounders $\mathbf{Z}$ are assumed to temporally precede the exposure X . In the regression models that follow, $\mathbf{Z}$ is additionally assumed to contain a constant. The sample data is $\{(Y_i, X_i, \mathbf{Z}_i); i = 1, \dots, n\}$. Each subject has two potential outcomes: Y_0 , the outcome that would occur if X were set to 0, and Y_1 , the outcome that would occur if X were set to 1. Only one of these can ever be observed. We assume that when the observed exposure is $X = x^*$, the observed outcome is equal to the appropriate potential outcome $Y = Y_{x^*}$ (the consistency assumption [17]). Thus, $Y = XY_1 + (1 - X)Y_0$.

We assume that the data are an independent identically distributed sample from an (near) infinite population. The assumption of independence implies the stable unit treatment value assumption [18], which states that the potential outcomes for one individual are unaffected by the exposure status of another. We assume that all subjects have a non-zero probability of having either exposure; one of the potential outcomes would otherwise not be defined (the positivity assumption [19]). Finally, we make the exchangeability or strongly ignorable treatment assignment (SITA) assumption [13], which states that there are no unmeasured confounding variables. Formally, $(Y_1, Y_0) \perp\!\!\!\perp X | \mathbf{Z}$, where we use $\perp\!\!\!\perp$ to denote conditional independence [20].

For an individual, the causal effect of exposure to X is $Y_1 - Y_0$. Averaging this over the population of interest, the causal estimand is $\Delta = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$. We describe the estimation of the marginal mean $\mu_1 = \mathbb{E}[Y_1]$. Similar methods can be applied to the estimation of $\mu_0 = \mathbb{E}[Y_0]$, from which an estimate of Δ can be derived.

2.2. G-computation

Our first approach to the estimation of μ_1 , G-computation, can be thought of as a form of standardization [21, 22], on the basis of the equality $\mathbb{E}[Y_1] = \mathbb{E}_{\mathbf{Z}}[\mathbb{E}[Y_1 | \mathbf{Z}]]$. Under the consistency and SITA assumptions described earlier, $\mathbb{E}[Y_1 | \mathbf{Z}] = \mathbb{E}[Y | X = 1, \mathbf{Z}]$. Suppose $\mathbb{E}[Y | X = 1, \mathbf{Z}] = \mathbf{Z}\boldsymbol{\gamma}$ with $\boldsymbol{\gamma} = (\gamma_0, \dots, \gamma_p)^T$. The parameter $\boldsymbol{\gamma}$ can be estimated using, for example, a linear regression model of observed outcome Y on the measured confounders $\mathbf{Z}$ among exposed subjects (those with $X = 1$). Then, $\mathbb{E}[Y_1 | \mathbf{Z}]$ can be estimated from this model for each subject, whether exposed or unexposed, by $\hat{Y}_1 = \mathbf{Z}\hat{\boldsymbol{\gamma}}$, using hats to indicate sample estimates. To estimate μ_1 , we must take the expectation of the estimated $\mathbb{E}[Y_1 | \mathbf{Z}]$ over the distribution of the measured confounders, $\mathbf{Z}$. This can be estimated by taking the sample average of these estimates,

$$\hat{\mu}_1^{\text{gc}} = \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i \hat{\boldsymbol{\gamma}} = \frac{1}{n} \sum_{i=1}^n \hat{Y}_{1i}$$

where $\hat{Y}_{1i} = \mathbf{Z}_i \hat{\boldsymbol{\gamma}}$ is an estimate of the potential outcome Y_1 for subject i .

Alternatively, we could fit the model $\mathbb{E}[Y | X, \mathbf{Z}] = \mathbf{Z}\boldsymbol{\gamma}^* + \eta X$ and then estimate $\mathbb{E}[Y_1 | \mathbf{Z}]$ by setting $X = 1$. If this regression model was used to estimate both μ_1 and μ_0 , then the estimate of Δ , the population-average exposure effect, would simply be $\hat{\eta}$, the estimated regression coefficient for the exposure variable X . This is commonly used as an estimate of the exposure effect in medical and epidemiological research. It does, however, make the fairly strong assumptions that (i) the effect of the exposure, $(Y_1 - Y_0)$, is the same for all individuals and (ii) the relationship between Y_1 and $\mathbf{Z}$ is the same as that between Y_0 and $\mathbf{Z}$. The latter assumption can be lessened by including interaction terms between the confounders and exposure in the regression model.

2.3. Inverse probability-of-treatment weighting

Inverse probability-of-treatment weighting provides an alternative method for estimating μ_1 . We denote the probability of treatment (or exposure) conditional on the measured confounders, also known as the propensity score [13], by $e(\mathbf{Z}) = \Pr(X = 1 | \mathbf{Z})$. Suppose that the binary exposure X follows a logistic model parametrized by $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_p)^T$, so that $\log\{e(\mathbf{Z})/(1-e(\mathbf{Z}))\} = \mathbf{Z}\boldsymbol{\alpha}$. The parameter $\boldsymbol{\alpha}$ can be estimated from a standard logistic regression model by using maximum likelihood, giving estimates $\hat{\boldsymbol{\alpha}}$. The propensity score is estimated as

$$\hat{e} = \hat{e}(\mathbf{Z}) = \frac{\exp(\mathbf{Z}\hat{\boldsymbol{\alpha}})}{1 + \exp(\mathbf{Z}\hat{\boldsymbol{\alpha}})}$$

Inverse probability-of-treatment weighting derives from the equality $\mathbb{E}[Y_1] = \mathbb{E}[YX/e(\mathbf{Z})]$. To see this, by conditioning on $\mathbf{Z}$, we have $\mathbb{E}[YX/e(\mathbf{Z})] = \mathbb{E}[\mathbb{E}[YX/e(\mathbf{Z}) | \mathbf{Z}]] = \mathbb{E}[\mathbb{E}[Y_1 | \mathbf{Z}]\mathbb{E}[X | \mathbf{Z}]/e(\mathbf{Z})]$, where the latter equality follows from the consistency and SITA assumptions. Because $\mathbb{E}[X | \mathbf{Z}] = e(\mathbf{Z})$, the right-hand side cancels to give $\mathbb{E}[\mathbb{E}[Y_1 | \mathbf{Z}]] = \mathbb{E}[Y_1]$, from which we have $\mathbb{E}[YX/e(\mathbf{Z})] = \mathbb{E}[Y_1]$ as desired. This relationship prompts various inversely weighted estimators, including

$$\hat{\mu}_1^{\text{ipw}} = \left(\sum_{i=1}^n \frac{Y_i X_i}{\hat{e}_i} \right) \left(\sum_{i=1}^n \frac{X_i}{\hat{e}_i} \right)^{-1} \quad (1)$$

where $\hat{e}_i = \hat{e}(\mathbf{Z}_i)$ is the estimated propensity score for subject i . It is useful to note that the estimator $\hat{\mu}_1^{\text{ipw}}$ can be obtained as the constant term in a weighted linear regression model for Y including no covariates, weighting subject i by $X_i/\hat{e}_i$. A similar regression model, but weighting by $(1 - X_i)/(1 - \hat{e}_i)$, gives an estimate of μ_0 . The IPW estimate of the average causal effect, Δ , can, therefore, be obtained as the coefficient for X in a weighted linear regression model for Y including only X , weighting exposed subjects by $1/\hat{e}_i$ and unexposed subjects by $1/(1 - \hat{e}_i)$. However, such a regression model will not produce appropriate standard errors because it treats the estimated propensity scores as known probabilities. Either bootstrapping or a correction to the standard error must be used to obtain valid 95% confidence intervals [14].

2.4. Doubly robust estimation

The estimator $\hat{\mu}_1^{\text{ipw}}$ is a consistent estimator of μ_1 [14] but can have large variance if some propensity scores are close to zero. It also requires the propensity score model (which we will call the ‘ e -model’) to be correctly specified. Similarly, the estimator $\hat{\mu}_1^{\text{gc}}$ requires the outcome model (the y -model) to be correctly specified. Modified IPW estimators, AIPW estimators, have been proposed. These have the desirable property of being consistent estimators of μ_1 if either the y -model or the e -model is correctly specified [5]. Such estimators are called ‘doubly robust’ because they are robust to the misspecification of either the outcome or the propensity score model (but not both). One AIPW estimator is given by [14]

$$\hat{\mu}_1^{\text{aipw}} = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{Y_i X_i}{\hat{e}_i} - \frac{(X_i - \hat{e}_i)}{\hat{e}_i} \hat{Y}_{1i} \right\} \quad (2)$$

where $\hat{Y}_{1i}$ and $\hat{e}_i$ are defined as in Sections 2.2 and 2.3, respectively. We demonstrate and discuss the double-robustness property for a more general scenario in Section 3.3. Equation (2) can be re-arranged to give

$$\hat{\mu}_1^{\text{aipw}} = \frac{1}{n} \sum_{i=1}^n \hat{Y}_{1i} - \frac{1}{n} \sum_{i=1}^n w_i (\hat{Y}_{1i} - Y_i), \quad \text{where } w_i = \frac{X_i}{\hat{e}_i} \quad (3)$$

The first term in Equation (3) is simply the G-computation estimator, $\hat{\mu}_1^{\text{gc}}$ —the mean predicted potential outcome Y_1 (the outcome that would occur under $X = 1$) taken over the whole sample, those both exposed and unexposed. The error in estimating a subject’s potential outcome Y_1 by $\hat{Y}_1$ is the difference $\hat{Y}_1 - Y_1$. For exposed subjects $Y = Y_1$, thus, the error $\hat{Y}_1 - Y_1$ is observed and is equal to $\hat{Y}_1 - Y$. For unexposed subjects, the error is unobserved (because $Y \neq Y_1$ if $X = 0$), so the second term in Equation (3) takes the observed error among exposed subjects and re-weights the exposed sub-sample (by X/e) to reflect the confounder distribution in the whole sample, which provides an estimate of the error in the whole sample. Thus, this second term can be seen as a ‘bias correction’ [4].

2.5. Estimating equations for the doubly robust estimator

To extend the doubly robust estimator to scenarios with missing data, it is useful to note that calculating the estimator $\hat{\mu}_1^{\text{aipw}}$ is equivalent to solving a set of simultaneous estimating equations, which we now describe. To obtain this estimator, we begin by estimating the propensity score from a standard logistic regression model with exposure as the dependent variable, as described in Section 2.3. The maximum likelihood estimates of the propensity score parameters, $\hat{\alpha}$, are obtained by setting the derivative of the log-likelihood to zero. This derivative is equal to $\mathbf{U}_3$

$$\mathbf{U}_3 = \mathbf{Z}^\top \left(X - \frac{\exp(\mathbf{Z}\alpha)}{1 + \exp(\mathbf{Z}\alpha)} \right) = \mathbf{Z}^\top (X - e)$$

$\mathbf{U}_3$, and all other estimating equations, are written as column vectors of the appropriate size. The propensity score parameters $\hat{\alpha}$ are obtained as the value of α that solves $\sum_i \mathbf{U}_{3i} = 0$, where $\mathbf{U}_{3i}$ is $\mathbf{U}_3$ evaluated at the X and $\mathbf{Z}$ values of subject i , with the summation taken over the whole sample. As described in Section 2.2, to estimate the potential outcomes Y_1 , we fit a linear regression model of Y on $\mathbf{Z}$ among the exposed subjects ($X = 1$). This is equivalent to solving the estimating equations $\sum_i \mathbf{U}_{2i} = 0$, where

$$\mathbf{U}_2 = X \mathbf{Z}^\top (Y - \mathbf{Z}\gamma)$$

Then, the AIPW estimator $\hat{\mu}_1^{\text{aipw}}$ can be obtained by solving the estimating equation $\sum_i U_{1i} = 0$, where

$$U_1 = \frac{YX - (X - \hat{e}) \mathbf{Z}\hat{\gamma}}{\hat{e}} - \mu_1$$

leading to (2). Thus, in fully observed data, estimates of a vector of unknown parameters $\theta^\top = (\mu_1^{\text{aipw}}, \gamma^\top, \alpha^\top)$ can be obtained by simultaneously solving the estimating equations $\sum_i \mathbf{U}_i = 0$, where $\mathbf{U}^\top = (U_1, \mathbf{U}_2^\top, \mathbf{U}_3^\top)$. We note that $\mathbb{E}[\mathbf{U}] = 0$ at the true value of θ . This process can be easily extended to additionally estimate μ_0^{aipw} and hence the causal exposure effect.

3. Estimating causal effects when some variables are partially missing

Typically, one or more of the measured variables will have some missing values. In this section, we consider scenarios where one variable (the outcome, exposure or one of the confounders) is partially missing.

3.1. Missing data notation and terminology

In standard missing data terminology, data are described as being ‘missing completely at random’ if the probability of missingness does not depend on Y , Z or $\mathbf{X}$. Missing at random (MAR) means that the probability of missingness depends only on the fully observed component of the data, and missing not at random signifies that the missingness depends on the unobserved component of the data after conditioning on the fully observed component. For all AIPW estimators described in the following, missing data are assumed to be MAR.

We let $\mathbf{W}$ denote the data; in our scenario $\mathbf{W} = (Y, X, \mathbf{Z})$, a subset of which is fully observed ($\mathbf{V}$) with the remaining variable partially missing (Q), so $\mathbf{W} = (\mathbf{V}, Q)$. Complete cases are defined as subjects who have fully observed data. We define a complete case indicator, R , taking the value 1 if the subject has fully observed data and 0 otherwise. Assuming that data are MAR, the probability of being a complete case is defined as $\pi = \Pr(R = 1 \mid \mathbf{V})$.

3.2. Complete case analysis

When a variable is partially missing, the simplest approach is to perform a complete case analysis. This involves discarding any subjects with missing data (subjects with $R = 0$) and addressing confounding by analysing the remaining fully observed subset of data by using a method such as G-computation, IPW or a standard outcome regression model. This approach will typically lead to inefficient estimation of parameters because the data are discarded and often to biased estimates. Unless the data are missing completely at random, a complete case analysis will rarely produce unbiased estimates of a marginal mean such as $\mathbb{E}[Y_1]$. However, in many situations, the regression coefficients from a model for $\mathbb{E}[Y \mid X, \mathbf{Z}]$ will be unbiased. Under the SITA assumption, provided that the functional form of the regression model is correct, the estimates of the regression coefficients from this model will be unbiased provided that missingness does not depend on the outcome Y after conditioning on fully observed components of X and $\mathbf{Z}$. This has been demonstrated for scenarios where covariates are partially missing using simulation [23], and in general the theory of directed acyclic graphs [24].

Therefore, whenever missingness does not depend on Y (note that this includes some scenarios where X or $\mathbf{Z}$ are missing not at random), unbiased estimates of $\mathbb{E}[Y \mid X, \mathbf{Z}]$, and therefore $\mathbb{E}[Y_x \mid \mathbf{Z}]$, for $x = 0, 1$, can be obtained using complete case analyses. Then, $\mathbb{E}[Y_1 - Y_0 \mid \mathbf{Z}]$ can be (empirically) averaged over the distribution of $\mathbf{Z}$ to estimate the desired causal effect, $\Delta = \mathbb{E}[Y_1 - Y_0]$, provided that $\mathbf{Z}$ is fully observed. If a variable of $\mathbf{Z}$ is partially missing, the causal effect can still be estimated under an assumption of constant exposure effect across the population.

3.3. Constructing an augmented inverse probability weighted estimator

We briefly outline a method of constructing an AIPW estimator for missing data. Suppose that in the absence of missing data, a consistent estimator of a possibly vector-valued unknown parameter of interest θ —in our example this will be the vector $\theta^\top = (\mu_1^{\text{aipw}}, \gamma^\top, \alpha^\top)$ as discussed in Section 2.5—could be obtained by the value $\hat{\theta}$ of θ that solves the estimating equation

$$\sum_{i=1}^n \mathbf{U}(\mathbf{W}_i; \theta) = 0 \quad (4)$$

where consistency of $\hat{\theta}$ is achieved by choosing $\mathbf{U}$ such that $\mathbb{E}[\mathbf{U}] = 0$ at the true value of θ [10, p. 235]. Pre-multiplying $\mathbf{U}$ by the indicator of fully observed data, R_i , within the summation of Equation (4) would produce the (possibly biased) complete case estimator, and pre-multiplying by R_i/π_i would produce an IPW-type estimator. A doubly robust AIPW estimator can be obtained by solving the estimating equation [10, pp. 226–227]

$$\sum_{i=1}^n \mathbf{U}^\dagger(\mathbf{W}_i; \boldsymbol{\theta}) = 0, \quad \text{where} \quad \mathbf{U}^\dagger(\mathbf{W}_i; \boldsymbol{\theta}) = \frac{R_i}{\pi_i} \mathbf{U}(\mathbf{W}_i; \boldsymbol{\theta}) - \frac{(R_i - \pi_i)}{\pi_i} \mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta}) | \mathbf{V}_i] \quad (5)$$

where the expectation $\mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta}) | \mathbf{V}_i]$ is taken over the true distribution of the data, at the true value of $\boldsymbol{\theta}$. In practice, the probability of fully observed data π and the expectation $\mathbb{E}[\mathbf{U} | \mathbf{V}]$ are estimated from the data, and Equation (5) is solved after substituting in these estimates [10, pp. 229–230]. Typically, π is estimated using predicted values from a logistic regression model of R on $\mathbf{V}$. We call the model used to estimate π the ‘ π -model’. $\mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta}) | \mathbf{V}_i]$ is usually estimated by assuming a parametric model for the conditional distribution of the partially missing variable Q given the fully observed variables $\mathbf{V}$. We give examples of this process in our simulation study (Section 5).

The estimator $\hat{\boldsymbol{\theta}}$ obtained from solving (5) has an interesting property, which we will call the ‘double-robustness’ property: $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$ if either the π -model is correct or the expectation $\mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta}) | \mathbf{V}]$ is correct. To demonstrate this double-robustness property, we note that $\hat{\boldsymbol{\theta}}$ will be consistent if the expectation of the estimating equation is zero at the true value of the parameters [10, p. 235]. Thus, $\hat{\boldsymbol{\theta}}$ will be consistent if $\mathbb{E}[\mathbf{U}^\dagger(\mathbf{W}_i; \boldsymbol{\theta})] = 0$ as defined in (5) with the expectation taken at the true value of the parameter $\boldsymbol{\theta}$. We now demonstrate that this is true when either the π -model or the expectation $\mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta}) | \mathbf{V}]$ is correct.

If the π -model is correct, conditioning on the fully observed variables $\mathbf{V}$ and noting that $\mathbb{E}\left[\frac{R}{\pi} | \mathbf{V}\right] = 1$ gives

$$\begin{aligned} \mathbb{E}[\mathbf{U}^\dagger(\mathbf{W}; \boldsymbol{\theta})] &= \mathbb{E}_{\mathbf{V}} \left[\mathbb{E} \left[\frac{R}{\pi} \middle| \mathbf{V} \right] \mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta}) | \mathbf{V}] \right] - \mathbb{E}_{\mathbf{V}} \left[\mathbb{E} \left[\frac{(R - \pi)}{\pi} \middle| \mathbf{V} \right] \mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta}) | \mathbf{V}] \right] \\ &= \mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta})] - 0 = 0 \end{aligned}$$

where the last equality follows because $\mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta})] = 0$ at the true value of $\boldsymbol{\theta}$, so if the π -model is correct, we have $\mathbb{E}[\mathbf{U}^\dagger(\mathbf{W}_i; \boldsymbol{\theta})] = 0$, which implies that the estimator obtained by solving (5) is consistent. This is true irrespective of the correct specification of the function $\mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta}) | \mathbf{V}]$.

Now suppose the expectation $\mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta}) | \mathbf{V}]$ is correct. We have

$$\mathbb{E}[\mathbf{U}^\dagger(\mathbf{W}; \boldsymbol{\theta})] = \mathbb{E} \left[\frac{R}{\pi} \{ \mathbf{U}(\mathbf{W}; \boldsymbol{\theta}) - \mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta}) | \mathbf{V}] \} \right] + \mathbb{E} \left[\frac{\pi}{\pi} \mathbf{U}(\mathbf{W}; \boldsymbol{\theta}) \right] = 0$$

by conditioning the first term on $\mathbf{V}$, so if $\mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta}) | \mathbf{V}]$ is correct, we have $\mathbb{E}[\mathbf{U}^\dagger(\mathbf{W}_i; \boldsymbol{\theta})] = 0$, which implies that the estimator obtained by solving (5) is consistent. This remains true even if we replace π by an incorrect probability.

So, assuming that solving (4) gives a consistent estimator of $\boldsymbol{\theta}$ in fully observed data, we have shown that if either the π -model or the expectation $\mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta})]$ is correct, then solving Equation (5) will produce a consistent estimator of $\boldsymbol{\theta}$.

3.4. Augmented inverse probability weighted estimator of μ_1 when the outcome Y is partially missing

When the outcome Y is partially missing, we can construct an AIPW estimator as outlined in Section 3.3. The full data are $\mathbf{W} = (Y, X, \mathbf{Z})$, and the fully observed variables are $\mathbf{V} = (X, \mathbf{Z})$. The missingness indicator R takes the value 1 if Y is observed and 0 otherwise, and $\pi = \Pr(R = 1 | X, \mathbf{Z})$. We assume that missingness follows a logistic model parametrized by $\boldsymbol{\xi} = (\xi_0, \xi_1, \dots, \xi_{p+1})^\top$, so that $\log\{\pi/(1 - \pi)\} = \mathbf{V}\boldsymbol{\xi}$. Then, π is estimated by

$$\hat{\pi} = \hat{\pi}(\mathbf{V}) = \frac{\exp(\mathbf{V}\hat{\boldsymbol{\xi}})}{1 + \exp(\mathbf{V}\hat{\boldsymbol{\xi}})}$$

In fully observed data, the estimator $\hat{\mu}_1^{\text{aipw}}$ is obtained by solving the estimating equations $\sum_i \mathbf{U}_i = 0$, where $\mathbf{U}^\top = \{\mathbf{U}_1, \mathbf{U}_2^\top, \mathbf{U}_3^\top\}$, as defined in Section 2.5. To calculate the AIPW estimator described in Section 3.3, it is necessary to calculate $\mathbb{E}[U_k | X, \mathbf{Z}]$ for $k = 1, 2, 3$. We begin by focusing on $\mathbf{U}_2$ and $\mathbf{U}_3$. We have $\mathbb{E}[\mathbf{U}_2 | X, \mathbf{Z}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{U}_3 | X, \mathbf{Z}] = \mathbf{U}_3$. Substituting these equalities into (5), estimating π by $\hat{\pi}$, we have the following estimating equations:

$$\begin{aligned} \mathbf{U}_2^\dagger &= \frac{R}{\hat{\pi}} \mathbf{X} \mathbf{Z}^\top (\mathbf{Y} - \mathbf{Z} \boldsymbol{\gamma}) \\ \mathbf{U}_3^\dagger &= \left(\frac{R}{\hat{\pi}} - \frac{(R - \hat{\pi})}{\hat{\pi}} \right) \mathbf{U}_3 = \mathbf{U}_3 \end{aligned}$$

Because $\mathbf{U}_3^\dagger = \mathbf{U}_3$, the propensity score parameters $\boldsymbol{\alpha}$ are estimated as they would be in fully observed data; from a logistic regression model fitted on the whole sample. These estimated parameters are used to obtain the estimated propensity scores, $\hat{e}$. $\mathbf{U}_2^\dagger$ determines that the outcome model for Y_1 is fitted within the fully observed sub-sample after inversely weighting for missingness of Y (weighting by $R/\hat{\pi}$) to obtain estimates of $\boldsymbol{\gamma}$. These estimates give the predicted potential outcomes, $\hat{Y}_1 = \mathbf{Z} \hat{\boldsymbol{\gamma}}$.

We turn finally to U_1 . We have

$$\mathbb{E}[U_1 | \mathbf{X}, \mathbf{Z}] = \frac{\mathbf{X} \mathbb{E}[Y_1 | \mathbf{Z}] - (X - e) \mathbf{Z} \boldsymbol{\gamma}}{e} - \mu_1 = \mathbf{Z} \boldsymbol{\gamma} - \mu_1$$

Substituting this into (5), plugging in estimates of π , e and $\boldsymbol{\gamma}$ as described earlier, we have the following estimating equation:

$$U_1^\dagger = \frac{R}{\hat{\pi}} \left\{ \frac{YX - (X - \hat{e}) \mathbf{Z} \hat{\boldsymbol{\gamma}}}{\hat{e}} \right\} - \frac{(R - \hat{\pi})}{\hat{\pi}} \mathbf{Z} \hat{\boldsymbol{\gamma}} - \mu_1 = \frac{RX}{\hat{\pi} \hat{e}} (Y - \mathbf{Z} \hat{\boldsymbol{\gamma}}) + \mathbf{Z} \hat{\boldsymbol{\gamma}} - \mu_1$$

Solving $\sum_i U_{1i}^\dagger = 0$ gives the following estimate of μ_1 :

$$\hat{\mu}_1^{\text{aipw-y}} = \frac{1}{n} \sum_{i=1}^n \hat{Y}_{1i} - \frac{1}{n} \sum_{i=1}^n w_i^\dagger (\hat{Y}_{1i} - Y_i), \text{ where } w_i^\dagger = \frac{R_i X_i}{\hat{\pi}_i \hat{e}_i} \quad (6)$$

This estimator is very similar in form to (3). The first term is the mean predicted potential outcome (under $X = 1$) taken over the whole sample, and the second term estimates the sample mean error in the prediction $(\hat{Y}_1 - Y_1)$. Now, however, this error is unobserved both for subjects who are unexposed and for subjects whose outcome is missing, so the weighting, by $RX/(\hat{\pi}\hat{e})$, re-weights the exposed fully observed sub-sample to reflect the confounder distribution in the whole sample. By the double-robustness property discussed in Section 3.3, $\hat{\mu}_1^{\text{aipw-y}}$ will be consistent if (i) the full data estimating equation U_1 provides a consistent estimator of μ_1 and (ii) either the π -model or $\mathbb{E}[U_1 | \mathbf{X}, \mathbf{Z}]$ is correct. For (i), $\hat{\mu}_1^{\text{aipw}}$ is consistent if either the y -model or the e -model is correct (Section 3.3). For (ii), the calculation of $\mathbb{E}[U_1 | \mathbf{X}, \mathbf{Z}]$ depended only on the y -model being correct. Thus, $\hat{\mu}_1^{\text{aipw-y}}$ will be a consistent estimator of μ_1 if either the y -model or both the e -model and the π -model are correctly specified.

Standard errors for $\hat{\mu}_1^{\text{aipw-y}}$ can be estimated as described in the Web Appendix.

3.4.1. Connection between our augmented inverse probability weighted estimator Equation (6) and other approaches. An alternative approach to this missing outcome data problem would be to view the missingness of Y and being unexposed both as types of missingness, because both prevent Y_1 from being observed. Then, Y_1 is observed if both $R = 1$ and $X = 1$. We could define a joint missingness indicator, $R^* = RX$. In the ‘full data’ (where no data are missing and everyone is exposed to X), we would estimate μ_1 by using the estimating equation defined by $U = Y_1 - \mu_1$. The probability of being a complete case, using the new definition of complete case, would be $\Pr(R^* = 1 | \mathbf{Z}) = \Pr(R = 1 | \mathbf{Z}, X) \Pr(X = 1 | \mathbf{Z}) = \pi e$. Substituting these into Equation (5) gives the same estimator, $\hat{\mu}_1^{\text{aipw-y}}$. However, this ‘joint missingness’ approach cannot be easily generalized to other situations, such as those with missing exposures or confounders, if the missingness depends on outcome after conditioning on the other fully observed variables. When this occurs, the outcome model fitted on the complete cases will be biased even if correctly specified, so the robustness property is lost.

We noted earlier that Davidian *et al* [16] have proposed a doubly robust estimator of a causal exposure effect in the presence of missing outcome data in a pre-test–post-test scenario where additional information is available on variables intermediate between exposure and outcome. We now consider the extension of our approach (Section 3.4) to this scenario and demonstrate that our approach produces the estimator proposed by Davidian *et al*. Suppose the fully observed component of the data $\mathbf{V}$ contains, in addition to X and $\mathbf{Z}$, post-exposure variables $\mathbf{M}$. In fully observed data, $\mathbf{M}$ would not be included in the

model for Y_1 because these variables may be at least partly a consequence of the exposure X . Further, conditioning on $\mathbf{M}$ may introduce ‘collider’ bias if unmeasured confounding of the $\mathbf{M} - Y$ association exists [25]. Thus, $\mathbf{M}$ would not be added to the predictors in the regression model represented by $\mathbf{U}_2$. Conversely, when applying Equation (5), we would include the fully observed $\mathbf{M}$ as part of the conditioning set when estimating $\pi = \Pr(R = 1|\mathbf{V})$ and when calculating $\mathbb{E}[U_1|\mathbf{V}]$. Applying Equation (5) to estimating equations involving $\mathbf{U}_2$ and $\mathbf{U}_3$ gives $\mathbf{U}_2^*$ and $\mathbf{U}_3^*$, which are equal to $\mathbf{U}_2^\dagger$ and $\mathbf{U}_3^\dagger$ earlier but with π allowed to depend also on $\mathbf{M}$. However, applying Equation (5) to U_1 gives U_1^* , which is not the same as $U_1^\dagger$. Let $Y_1^* = \mathbb{E}[Y|X = 1, \mathbf{Z}, \mathbf{M}]$. Then,

$$\mathbb{E}[U_1|X, \mathbf{Z}, \mathbf{M}] = \frac{Y_1^*X - (X - e)\mathbf{Z}\boldsymbol{\gamma}}{e} - \mu_1$$

Substituting this into (5) to obtain U_1^* and solving $\sum_i U_{1i}^* = 0$ give

$$\hat{\mu}_1^{\text{aipw-}y*} = \frac{1}{n} \sum_{i=1}^n \left\{ \hat{Y}_{1i} - \frac{X_i}{\hat{e}_i} (\hat{Y}_{1i} - \hat{Y}_{1i}^*) \right\} - \frac{1}{n} \sum_{i=1}^n w_i^* (\hat{Y}_{1i}^* - Y_i), \quad w_i^* = \frac{R_i X_i}{\hat{\pi}_i \hat{e}_i} \quad (7)$$

which can be re-arranged to give the estimator presented by Davidian *et al* [16]. As in Equation (6), the first term is the mean predicted potential outcome Y_1 in the whole sample, and the second term is a ‘bias’ correction. The difference between the two estimators lies in the way the first term is calculated. Equation (7) incorporates the additionally available data $\mathbf{M}$ to improve prediction of Y_1 .

To distinguish between the two models for the outcome, we call the new model for Y involving $\mathbf{M}$ the ‘y-imputation model’. The estimator $\hat{\mu}_1^{\text{aipw-}y*}$ is consistent for μ_1 if (i) the y-model or the e -model is correct (the condition required for U_1 to give a consistent estimator of μ_1) and (ii) either the π -model or $\mathbb{E}[U_1|X, \mathbf{Z}, \mathbf{M}]$ is correct. This latter expectation required the y-imputation model to be correct. Thus, $\hat{\mu}_1^{\text{aipw-}y*}$ will be a consistent estimator of μ_1 if (i) either the y-model or the e -model is correct and (ii) either the π -model or the y-imputation model is correct.

The first two models both deal with the issue of confounding (y-model and e -model). The second two deal with the issue of missing data (y-imputation model and π -model). The y-model and y-imputation model are both used in an imputation-like manner. The y-imputation model is used to ‘fill in’ missing outcomes, and the y-model is used to ‘fill in’ potential outcomes Y_1 for unexposed subjects (for whom Y_1 is unobserved). The e -model and π -model, conversely, are both used as inverse weights, so this estimator combines the imputation and inverse-weighting approaches in such a way that as long as one of these two approaches to confounding and one of these two approaches to missingness are correct, the resulting estimator will be consistent.

In the remainder of this paper, we return to the simpler situation where post-exposure variables $\mathbf{M}$ are not available. In this simpler case, the y-model and y-imputation model are identical; thus, the condition earlier simplifies to the condition that either the y-model or both the e -model and the π -model must be correct, which is the condition we arrived at previously, and estimator (7) reduces to estimator (6).

3.5. Augmented inverse probability weighted estimator when the exposure X is partially missing

With a partially missing exposure, the fully observed subset of data is $\mathbf{V} = (\mathbf{Z}, Y)$. We let $R = 1$ if X is observed and 0 otherwise, and $\pi = \Pr(R = 1|\mathbf{Z}, Y)$. As in the previous section, we assume that $\log\{\pi/(1 - \pi)\} = \mathbf{V}\boldsymbol{\xi}$ with $\boldsymbol{\xi} = (\xi_0, \xi_1, \dots, \xi_{p+1})^\top$. Fitting this logistic regression model gives the estimates $\hat{\pi}$. Letting $X^* = \Pr(X = 1|\mathbf{Z}, Y)$, which can be estimated using a logistic regression of X on $\mathbf{Z}$ and Y , we have

$$\begin{aligned} \mathbb{E}[U_1|\mathbf{Z}, Y] &= \mathbb{E}\left[\frac{YX - (X - e)\mathbf{Z}\boldsymbol{\gamma}}{e} - \mu_1 \mid \mathbf{Z}, Y\right] = \frac{YX^* - (X^* - e)\mathbf{Z}\boldsymbol{\gamma}}{e} - \mu_1 \\ \mathbb{E}[U_2|\mathbf{Z}, Y] &= \mathbb{E}[X\mathbf{Z}^\top(Y - \mathbf{Z}\boldsymbol{\gamma})|\mathbf{Z}, Y] = X^*\mathbf{Z}^\top(Y - \mathbf{Z}\boldsymbol{\gamma}) \\ \mathbb{E}[U_3|\mathbf{Z}, Y] &= \mathbb{E}\left[\mathbf{Z}^\top\left(X - \frac{\exp(\mathbf{Z}\boldsymbol{\alpha})}{1 + \exp(\mathbf{Z}\boldsymbol{\alpha})}\right) \mid \mathbf{Z}, Y\right] = \mathbf{Z}^\top(X^* - e) \end{aligned}$$

Substituting these into (5), replacing π and X^* by sample estimates $\hat{\pi}$ and $\hat{X}^*$ and re-arranging, we obtain the following equations:

$$\begin{aligned} U_1^{\dagger\dagger} &= \frac{R}{\hat{\pi}} \left\{ \frac{YX - (X - e)\mathbf{Z}\boldsymbol{\gamma}}{e} \right\} - \frac{(R - \hat{\pi})}{\hat{\pi}} \left\{ \frac{Y\hat{X}^* - (\hat{X}^* - e)\mathbf{Z}\boldsymbol{\gamma}}{e} \right\} - \mu_1 \\ U_2^{\dagger\dagger} &= \left(\hat{X}^* - \frac{R}{\hat{\pi}} (\hat{X}^* - X) \right) \mathbf{Z}^\top (Y - \mathbf{Z}\boldsymbol{\gamma}) \\ U_3^{\dagger\dagger} &= \mathbf{Z}^\top \left\{ \left(\hat{X}^* - \frac{R}{\hat{\pi}} (\hat{X}^* - X) \right) - e \right\} \end{aligned}$$

The propensity score parameters are estimated by solving the modified estimating equation $\sum_i U_{3i}^{\dagger\dagger} = 0$. These are then used to estimate the propensity scores, $\hat{e}$. The equation $U_3^{\dagger\dagger}$ is similar to a logistic score function but with the partially missing X replaced by an estimate of X , namely $\hat{X}^* - (\hat{X}^* - X) R/\hat{\pi}$. The first term in this estimate is $\hat{X}^*$, obtained from a model for X given $\mathbf{Z}$ and Y (the x -model). The second term is a ‘bias’ adjustment, correcting for misspecification of the x -model using the missingness model. The parameters $\boldsymbol{\gamma}$, used to obtain the predicted potential outcomes $\hat{Y}_1$, are estimated by solving $\sum_i U_{2i}^{\dagger\dagger} = 0$. This is similar to the score function from a linear regression model, fitted on the exposed subgroup ($X = 1$), with X replaced by the estimate described earlier. Although these estimating equations do not correspond to standard models, they can be solved in many standard statistical packages (e.g. using the `gmm` command in `Stata`) [26]. Solving $\sum_i U_{1i}^{\dagger\dagger} = 0$ gives the doubly robust estimator

$$\hat{\mu}_1^{\text{aipw}_x} = \frac{1}{n} \sum_{i=1}^n \hat{Y}_{1i} - \frac{1}{n} \sum_{i=1}^n w_i^{\dagger\dagger} (\hat{Y}_{1i} - Y_i), \text{ where } w_i^{\dagger\dagger} = \frac{\hat{X}_i^* - (\hat{X}_i^* - X_i) R_i / \hat{\pi}_i}{\hat{e}_i} \quad (8)$$

The estimator $\hat{\mu}_1^{\text{aipw}_x}$ is again very similar in form to (3). As previously, the second term estimates the mean error in the prediction $(\hat{Y}_1 - Y_1)$. Because this is observed for exposed subjects only, to estimate the error in the whole sample, we must weight by X/e , but this is not possible because X is partially missing. Therefore, the weights are created from the estimate of X , discussed earlier: $\hat{X}^* - (\hat{X}^* - X) R/\hat{\pi}$.

Because the expectation $\mathbb{E}[U_1 | \mathbf{Z}, Y]$ required only the x -model to be correct, the estimator $\hat{\mu}_1^{\text{aipw}_x}$ will be a consistent estimator for μ_1 if (i) either the y -model or the e -model is correct and (ii) either the π -model or the x -model is correct. The first two models deal with confounding, and the last two deal with missing data. As before, this estimator combines the imputation and inverse-weighting approaches in such a way that as long as one of these two approaches to confounding and one of these two approaches to missingness is correct, the resulting estimator will be consistent.

Standard errors for $\hat{\mu}_1^{\text{aipw}_x}$ can be estimated as described in the Web Appendix.

3.6. Augmented inverse probability weighted estimator when a confounder is partially missing—algorithm for an approximate estimator

When a confounder is partially missing, the expectations in (5) can be difficult to calculate explicitly, such as $\mathbb{E}[(\exp(\mathbf{Z}\boldsymbol{\alpha})(1 + \exp(\mathbf{Z}\boldsymbol{\alpha}))^{-1})^{-1} | X, Y]$, as a function of X , Y , the unknown $\boldsymbol{\alpha}$ and unknown parameters of the distribution $f(\mathbf{Z} | X, Y)$. We have, therefore, used an approximation to obtain the doubly robust estimator in this situation, which we describe for a general partially missing variable. As earlier, the full data is $\mathbf{W} = (\mathbf{V}, Q)$, with Q denoting the partially missing variable. To obtain an estimate $\hat{\theta}$ from Equation (5), we must estimate both $\pi = \Pr(R = 1 | \mathbf{V})$ and $\mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta}) | \mathbf{V}]$. The former can be easily estimated from the data using, for example, a logistic regression model of R on $\mathbf{V}$. We estimate the latter by

$$\mathbb{E}[\mathbf{U}(\mathbf{W}; \boldsymbol{\theta}) | \mathbf{V}] = \int \mathbf{U}(Q, \mathbf{V}; \boldsymbol{\theta}) dF(Q | \mathbf{V}) \approx \frac{1}{M} \sum_{m=1}^M \mathbf{U}(\tilde{Q}_m, \mathbf{V}; \boldsymbol{\theta})$$

where $\{\tilde{Q}_m\}_{m=1}^M$ are M observations independently drawn from the distribution of $Q | \mathbf{V}$. We have assumed that the data are MAR, which implies that $f(Q|\mathbf{V}) = f(Q|\mathbf{V}, R = 1)$. Therefore, we can estimate the distribution of $Q | \mathbf{V}$ by using the complete case subjects only.

The estimating equation of (5) can be re-written using this approximation to give

$$\sum_{i=1}^n \left(\frac{R_i}{\pi_i} \mathbf{U}(W_i; \theta) - \frac{(R_i - \pi_i)}{\pi_i} \frac{1}{M} \sum_{m=1}^M \mathbf{U}(\tilde{Q}_{mi}, \mathbf{V}_i; \theta) \right) \approx 0 \quad (9)$$

After substituting in estimates of π , this estimating equation can be solved to obtain an estimate for the unknown parameter θ . In our scenario, $\theta^\top = (\mu_1^{\text{aipw}}, \boldsymbol{\gamma}^\top, \boldsymbol{\alpha}^\top)$, and $\mathbf{U}^\top = (U_1, \mathbf{U}_2^\top, \mathbf{U}_3^\top)$ as defined in Section 2.5.

We now describe a simple algorithm involving data duplication to solve estimating Equation (9) that can be applied using any standard statistical software that allows negative weights. The steps involved in the algorithm are as follows:

- (1) Estimate the probability of being a complete case, $\hat{\pi}$, for each subject.
- (2) Estimate the distribution of $Q | \mathbf{V}$ from the complete case data (subjects with $R = 1$). In practice, this will typically involve fitting a standard regression model for Q on $\mathbf{V}$. For example, suppose the partially missing variable was a continuous variable Z_1 . We might fit the standard linear regression model $Z_1 = \mathbf{V}\boldsymbol{\nu} + \epsilon$, with $\epsilon \sim N(0, \sigma^2)$.
- (3) Create M duplicates of the dataset. Delete all observed values of the partially missing variable Q in these duplicated datasets and ‘fill in’ the variable with random draws from the estimated distribution of $Q | \mathbf{V}$, so, for example, with the partially missing Z_1 in the previous step, our ‘filled-in’ value of Z_1 for subject i in dataset m would be $\mathbf{V}\hat{\boldsymbol{\nu}} + \epsilon_{mi}$, where ϵ_{mi} is drawn from the $N(0, \hat{\sigma}^2)$ distribution.
- (4) Stack the M imputed datasets in a ‘long’ format, adding the complete case dataset—the original dataset after removing subjects with Q missing—to the top. The variable Q now consists of one set of the observed values (for complete cases) and M sets of imputed values (for the whole sample). We call the combined dataset—the M duplicated datasets and the complete case dataset— $\tilde{\mathbf{W}}$.
- (5) Assign a weight of $w_{0i} = 1/\hat{\pi}_i$ to the row of data for subject i in the complete case dataset.
- (6) Assign a weight of $w_{mi} = -(R_i - \hat{\pi}_i) / (\hat{\pi}_i M)$ to the row of data for subject i in duplicated dataset m , for $m = 1, \dots, M$.
- (7) Analyse the combined data, $\tilde{\mathbf{W}}$, as you would analyse the fully observed data but weighting the analyses by the weights described. Here, this involves fitting two standard (weighted) regression models: (i) a logistic regression model of X on $\mathbf{Z}$ to obtain the estimated propensity scores, $\hat{e}$, and (ii) a linear regression model for outcome Y on $\mathbf{Z}$ among exposed subjects (those with $X = 1$) to predict the potential outcomes, $\hat{Y}_1$. Finally, the estimator for μ_1 is obtained by taking the weighted sample mean of $\{YX - (X - \hat{e})\hat{Y}_1\}/\hat{e}$.

The software used to fit the two weighted regression models in the last step must allow negative weights. This can be achieved in the statistical package *Stata*, for example, using importance weights [26]. However, we are not aware of equivalent facilities for allowing negative weights in *SAS* [27] or *R* [28]; therefore, some additional programming would be needed to solve the required estimating equations in these packages. We emphasize that these negative weights are used simply as a trick to allow calculation of the approximate estimating Equation (9) by using standard software, rather than the result of a complex weighting scheme.

We do not estimate standard errors of estimators using this procedure; we rely on bootstrap confidence intervals as described in the next section.

This algorithm uses the estimated distribution of $Q | \mathbf{V}$ to approximate the expectation in (5). Other than the error created by approximating the expectation by the summation over the duplicated datasets, the resulting estimator will be consistent if (i) either the y -model or the e -model is correct (conditions for consistency for μ_1 in fully observed data) and (ii) either the π -model or the model for $Q | \mathbf{V}$ is correct. As for the previous two scenarios, this AIPW estimator combines the imputation and inverse-weighting approaches in such a way that as long as one of these two approaches to confounding and one of these two approaches to missingness is correct, the resulting estimator will be consistent.

4. Application: the B-Aware trial

In this section, we apply the methods developed in Section 3 to address a non-randomized comparison in data taken from a randomized trial in anaesthesia. In this example, a confounder is partially missing; thus, we use the procedure described in Section 3.6 to approximate the appropriate AIPW estimator.

General anaesthesia can be maintained by inhaled volatile anaesthetics, such as sevoflurane, or by intravenously administered drugs (total intravenous anaesthetic; TIVA), such as propofol. Previous studies have suggested that TIVA is associated with lower post-operative anxiety and minor post-operative complications. The B-Aware trial [29] is a multicentre, randomized controlled trial investigating the effect of bispectral index (BIS) monitoring—a method for measuring of depth of anaesthesia—in preventing awareness during general anaesthesia in patients at high risk of awareness (e.g. high-risk cardiac surgery, severe end-stage lung disease, past history of awareness and known or suspected heavy alcohol intake). We use data taken from this trial to estimate the effect of TIVA, a non-randomized exposure, on post-operative anxiety.

In the original trial, 2463 surgical patients aged 18 years or older, who were at high risk of awareness, were randomized to BIS monitoring (target range 40–60) or routine care. In all other respects, peri-operative care was at the discretion of the anesthesiologist. The anxiety subscale of the Hospital Anxiety and Depression (HAD) scale [30], an ordinal scale from 0 to 21 with higher values indicating a more anxious state, was used to measure pre-operative and 30 days post-operative anxiety. Baseline characteristics including ASA status (assessment of patient's pre-operative physical state), weight, gender and age were measured. The trial took place between September 2000 and December 2002.

For the current analysis, we have taken a random sample of approximately 80% of the patients allocated to the BIS arm, leaving $n = 942$ patients. This analysis is intended to illustrate the methods and should not be used to inform decisions regarding use of TIVA. In our sub-sample of data, pre-operative anxiety was not measured for 131 patients (14%). All other relevant variables are fully observed.

4.1. Statistical methods

We first analyse the data using a complete case approach, where only subjects with measured pre-operative anxiety are included. We perform four different complete case analyses: (i) linear regression, (ii) G-computation, (iii) propensity score analysis (IPW) and (iv) complete case AIPW. We then analyse the data using two approaches to dealing with missing data: multiple imputation followed by methods (i)–(iii) and applying the approach outlined in this paper—using the AIPW estimator obtained via the approximation described in Section 3.6.

The outcome of interest is post-operative anxiety, and the binary exposure is TIVA. Age, sex, weight, ASA status and the partially missing pre-operative anxiety are potential confounding variables (hereafter called ‘the confounders’).

We first describe the complete case analyses. For (i), we fitted a linear regression model for post-operative anxiety including the confounders and TIVA as independent variables in the model and used the estimated TIVA regression coefficient as the estimate of TIVA effect. For (ii), the G-computation estimator, we fitted two linear regression models for post-operative anxiety, one within each TIVA group, including all confounders. We used the two models to predict post-operative anxiety under both exposure conditions (TIVA/no TIVA) for all complete case subjects. We then estimated the effect of TIVA by the difference between the complete case sample means of these two predicted potential outcomes. For (iii), the IPW estimator, we estimated the propensity score by fitting a logistic regression model for TIVA on the confounders, using only the complete cases. This fitted logistic regression model was used to predict the probability of receiving TIVA—the propensity score—for each complete case subject. Participants who received TIVA were allocated a weight of the inverse of their estimated propensity score. Participants who did not receive TIVA were allocated a weight of the inverse of 1 minus their estimated propensity score. We obtained the IPW estimator by fitting a linear regression model for post-operative anxiety including only TIVA, to the complete case sample, using these weights as probability weights and taking the estimated regression coefficient for TIVA as the IPW estimate of effect. Finally, for (iv), the complete case AIPW estimator, the two potential outcomes under TIVA/no TIVA were predicted for every subject in the complete case sample as described earlier for the G-computation estimator, and the propensity score was estimated as for the IPW estimator. Estimators of the mean outcomes under TIVA and no TIVA were obtained by calculating the sum given in Equation (3) and an analogous

expression for the sample mean under no TIVA. The AIPW estimate of TIVA effect was obtained as the difference between these two means.

The conventional standard errors and 95% confidence intervals (CIs) are reported for the linear regression estimate (i). Because standard statistical software does not contain routines to produce correct standard errors for multi-step estimators, bootstrapping was used to obtain 95% confidence intervals for estimators (ii)–(iv). Specifically, bootstrap percentile confidence intervals were calculated using 500 bootstrap samples.

We now describe our two approaches to dealing with the missing confounder data. For analyses using multiple imputation, the imputation model was a linear regression model for pre-operative anxiety including all confounders and the outcome, post-operative anxiety. Fifty imputed datasets were generated. Each imputed dataset was analysed using methods (i)–(iii) as described earlier, and the resulting estimates and standard errors were combined across the imputed datasets using Rubin's rules [1].

Our second approach to the missing data is to obtain an AIPW estimator as described in this paper. Because the partially missing variable is a confounder, our AIPW estimator has no closed-form solution, so the procedure described in Section 3.6 was used to obtain an approximation to the estimator. $M = 50$ data replications were used for this procedure. To obtain the AIPW estimator, four models are required: (i) a model for missingness of pre-operative anxiety, (ii) a model for pre-operative anxiety, (iii) a propensity score model (model predicting assignment to TIVA) and (iv) a model for post-operative anxiety. For (i), we fitted a logistic regression model where the dependent variable was an indicator for pre-operative anxiety being observed and which included, as independent variables, the confounders and TIVA. For (ii), we fitted a linear regression model for pre-operative anxiety including TIVA, the confounders and post-operative anxiety. The remaining two models included the same independent variables as described for the complete case G-computation and IPW estimators. To obtain our AIPW estimator, we follow the seven steps in the algorithm described in Section 3.6. Step 1 involves predicting the probability of having observed pre-operative anxiety from model (i). Step 2 fits model (ii). Step 3 uses the fitted model (ii) to create M sets of imputed pre-operative anxiety values for the whole sample. Steps 4–6 arrange the data in a format suitable for analysis and create analysis weights. Step 7 involves fitting models (iii) and (iv) in the weighted dataset. Then, the AIPW estimator is obtained by replacing Equation (3) by a weighted mean. Bootstrap percentile confidence intervals were calculated for the AIPW estimator using 500 bootstrap samples.

4.2. Results

Table I shows the baseline characteristics of our sub-sample of data from the B-Aware trial by TIVA status and by whether or not pre-operative anxiety was measured. A total of 531 (56%) of the sample

Table I. Baseline characteristics of our sub-sample of the B-Aware trial, by TIVA status and by missingness status of pre-operative anxiety.

Characteristic	Exposure (TIVA) status					Pre-operative anxiety (confounder)				
	TIVA (n = 531)		No TIVA (n = 411)		$p^{\dagger}$	Missing (n = 811)		Observed (n = 131)		$p^{\dagger}$
TIVA, n (%)						68	(51.9)	463	(57.1)	0.27
Female, n (%)	218	(41.1)	149	(36.3)	0.13	43	(32.8)	324	(40.0)	0.12
ASA status, n (%)										
No functional impairment	173	(32.6)	53	(12.9)	<0.001	25	(19.1)	201	(24.8)	<0.001
Some impairment	239	(45.0)	188	(45.7)		43	(32.8)	384	(47.4)	
Severe disease	119	(22.4)	170	(41.4)		63	(48.1)	226	(27.9)	
Age in years, mean (SD)	56.0	(17.5)	61.1	(14.2)	<0.001	53.8	(17.8)	59.0	(16.0)	<0.001
Weight in kg, median (IQR)	70	(60, 84)	71	(62, 82)	0.77	70	(60, 80)	70	(61, 84)	0.17
Pre-operative anxiety, mean (SD)	6.9	(4.6)	6.6	(4.6)	0.33					

† p -value from chi-squared test, t -test or Wilcoxon rank sum test, as appropriate.

Table II. Estimates of effect of TIVA on post-operative anxiety.[†]

Method	Estimated mean difference	95% CI	Width of CI
Complete case analyses			
Unadjusted difference	0.974	(0.408, 1.539)	1.131
Linear regression (LR)	0.764	(0.284, 1.244)	0.960
G-computation (G-comp)	0.769	(0.273, 1.221)	0.948
IPW	0.774	(0.248, 1.211)	0.963
AIPW	0.773	(0.402, 1.123)	0.721
Missing data approaches ($M = 50^{\dagger\dagger}$)			
Multiple imputation with LR	0.737	(0.286, 1.188)	0.902
Multiple imputation with G-comp	0.738	(0.298, 1.178)	0.880
Multiple imputation with IPW	0.739	(0.305, 1.174)	0.869
AIPW	0.763	(0.341, 1.182)	0.841

[†]Post-operative anxiety (HAD scale: 0–21), with higher values indicating a more anxious state.

^{††}The number of imputed datasets (multiple imputation) or data replications (AIPW).

received TIVA. Those who received TIVA were in a better physical state (ASA status) prior to the operation, were younger and a higher proportion were women. Those who are missing pre-operative anxiety were in a worse physical state prior to the operation, were younger and a higher proportion were men.

The mean (SD) post-operative anxiety was 3.8 (3.6) in the no TIVA group and 4.7 (4.3) in the TIVA group. All estimates of effect of TIVA in Table II are in the same direction, indicating an increase of post-operative anxiety associated with TIVA. However, the magnitudes of the point estimates are all very small compared with the range of the outcome scale, so none indicate clinically significant effects. The complete case analyses tended to give estimates of effect further from the null. In both the complete case analysis and the missing data approaches, standard linear regression gave the estimates closest to the null, and the AIPW estimates were among the furthest from the null.

Given the design of this trial, it seems likely that missingness depends only on baseline characteristics; thus, we might expect the complete case estimates to be unbiased. Interestingly, the AIPW estimator accounting for the missing data was very close to the complete case estimates, but the estimates obtained via multiple imputation were less so, perhaps indicating that the imputation model used for the multiple imputation does not fit well. In comparison with the complete case estimates, the AIPW estimate accounting for the missing data has a narrower confidence interval than all but the complete case AIPW estimator. In allowing for uncertainty about the missingness mechanism, some precision has been lost in the AIPW estimator.

5. Simulation study

We conducted a simulation study on the basis of the B-Aware example in the previous section, investigating three separate scenarios. In each scenario, a single variable was partially missing, and all other variables were fully observed. The partially missing variable in the three scenarios were (i) the outcome, post-operative anxiety, (ii) the exposure of interest, TIVA, and (iii) one of the confounders, pre-operative anxiety. We focus on estimation of the average causal effect, $\Delta = \mu_1 - \mu_0$, with μ_1 estimated as described in Section 3 and μ_0 estimated in a similar way.

5.1. The simulation scenario

Our simulation scenario is based on the B-Aware example. The measured baseline characteristics are $Z_1 = \text{sex}$, $Z_2 = \text{sex-specific obesity indicator}$, $Z_3 = \text{age (years)}$, $Z_4 = \text{disease severity underlying ASA status}$ and $Z_5 = \text{pre-operative anxiety}$. The exposure of interest, X , was TIVA, and the outcome, Y , was post-operative anxiety. The variable Z_4 is taken to be the underlying continuous quantity from which the ASA categories are derived. The assumed distributions, given in the following, give distributions for these variables similar to the real data. However, for simplicity, we simulated

anxiety from a normal distribution, so these variables are not restricted to the same range as the original variables.

Sex	$Z_1 \sim \text{Bernoulli}(50\%);$	ASA status	$Z_4 \sim \text{Normal}(\mu_{z4}, 0.5^2)$
Obesity	$Z_2 \sim \text{Bernoulli}(30\%);$	Pre-op anxiety	$Z_5 \sim \text{Normal}(\mu_{z5}, 2.675^2)$
Age	$Z_3 \sim \text{Uniform}[30, 70];$	TIVA	$X \sim \text{Bernoulli}(e)$
		Post-op anxiety	$Y \sim N(\mu_y, 1.875^2)$

$$\begin{aligned}\mu_{z4} &= 2.7 - 0.4Z_1 - 0.25Z_2 - Z_1Z_2 - 0.005Z_3 \\ \mu_{z5} &= 8.415 + 0.535Z_1 + 0.535Z_2 + 0.02675Z_3 - 0.02675Z_4 + 0.002675Z_4^2 \\ \ln\left(\frac{e}{1-e}\right) &= -3.969 - \ln(1.75)Z_1 - \ln(2.5)Z_2 + \ln(1.05)Z_3 + \ln(2)Z_4 + \ln(1.16)Z_5 \\ \mu_y &= X - 1.225Z_1 + 0.0000625Z_3 + 1.875Z_4 + \exp(Z_5/10)\end{aligned}$$

Approximately 65% of the sample are exposed. The propensity score distribution is negatively skewed. In a typical sample of $n = 5000$, the median (min, max) of the propensity score was 0.52 (0.04, 0.95) in the unexposed (no TIVA) group and 0.73 (0.08, 0.97) in the exposed (TIVA) group. The true average causal effect of the treatment, in our simulation study, is $\Delta = 1$: TIVA increases post-operative anxiety by one unit.

We have chosen extreme missingness mechanisms to more clearly illustrate the robustness properties of the AIPW estimators. For the partially missing outcome, Y , and partially missing confounder Z_5 , the complete case ($R = 1$) probability is defined by

$$\text{logit}\{\Pr(R = 1 \mid Z_1, Z_2)\} = \ln(5)Z_1 + \ln(1.05)Z_2$$

and for the partially missing TIVA exposure, X , the complete case probability is given by

$$\text{logit}\{\Pr(R = 1 \mid Z_5, Y)\} = -2.93 + \ln(1.16)Z_5 + \ln(1.48)Y$$

In all missing data scenarios, approximately 30% of subjects have missing values in the outcome, exposure or confounder.

To estimate the causal effect in each scenario, several models must be specified. For each model required, we compare results using the correct model and a particular incorrect model. The correct model for X given Y and Z_1, Z_2, Z_3, Z_4, Z_5 (the x -model) can be determined explicitly from the distributions given earlier, with a small amount of algebraic manipulation. For the scenario where a confounder is partially missing, however, the exact distribution of Z_5 conditional on Y, X, Z_1, Z_2, Z_3, Z_4 (the z -model) is required, but this is analytically intractable. The ‘correct’ model was therefore approximated by the closest fitting linear regression model identified using a large simulated dataset. As for the missing data mechanisms, we pick fairly extreme model misspecification to clearly illustrate the properties of our AIPW estimators. For the y -models and e -models, the incorrect model omits the variable ASA status, Z_4 . The incorrect z -model omits both ASA status, Z_4 , and post-operative anxiety, Y . The incorrect x -model omits both ASA status, Z_4 , and TIVA, X . When pre-operative or post-operative anxiety are partially missing, the incorrect π -model omits sex, Z_1 . When TIVA is partially missing, the incorrect π -model omits post-operative anxiety, Y .

5.2. The analysis

For each of the three scenarios (Y, X and Z_5 partially missing), 5000 datasets each containing $n = 1000$ subjects were simulated. The average causal exposure effect, Δ , was estimated by the appropriate AIPW estimator, either using Equations (6) and (8) with the analogous equation for estimating μ_0 or, when the confounder is partially missing, using the procedure of Section 3.6 with $M = 50$ replications of the data to approximate the AIPW estimator. The AIPW estimators were obtained under all possible combinations of correct and incorrect specifications of the models required. Asymptotic confidence intervals for the AIPW estimators defined by Equations (6) and (8) were obtained using the sandwich variance estimates given in the Web Appendix, with coverage estimated as the fraction of these containing the true value ($\Delta = 1$). Bootstrap percentile confidence intervals, using 500 bootstrap samples, are presented for

estimates obtained using the approximation of Section 3.6, on the basis of 500 simulations only because of computational considerations.

For comparison, Δ was also estimated using two complete case estimators. The first was a correctly specified linear regression model for the outcome including X, Z_1, Z_3, Z_4 and $\exp(Z_5/10)$ fitted in the complete case data. The second complete case estimator was an IPW estimator obtained as the X coefficient from a weighted linear regression model for Y on X , weighting exposed subjects by $1/\hat{e}$, the inverse of the propensity score estimated from a correctly specified model, and weighting unexposed subjects by $1/(1 - \hat{e})$. The conventional 95% confidence intervals are presented for the complete case outcome regression estimator, and the correction to the standard error proposed by Lunceford and Davidian was used to obtain 95% confidence intervals for the IPW estimator [14].

The mean of the estimates, the empirical standard error (SE, the standard deviation of estimates across simulations), the bias (the absolute difference between the mean estimate and the true value of 1) as a percentage of the standard error ($100 \times \text{bias}/\text{SE}$), the mean-squared error ($\text{MSE} = \text{bias}^2 + \text{SE}^2$) and the coverage of the 95% confidence intervals are presented.

To assess the accuracy of the approximate procedure of Section 3.6, we estimated Δ using both an explicit formula for the estimator and the approximate procedure in the two scenarios where both approaches are possible (those where the outcome or exposure are partially missing). We considered a range of number of dataset replications (M) to investigate the sensitivity of the estimates to the number of replications used, $M = 5, 10, 25, 50$. We simulated 500 datasets each containing $n = 1000$.

5.3. Simulation results

When the outcome Y is partially missing, we expect the two complete case estimators to be consistent. The AIPW estimator should be consistent if either the y -model is correct or if both the e -model and the π -model are correct. The simulation results, shown in Table III, show that this is indeed the case. Interestingly, although the complete case IPW estimator is consistent if the e -model is correctly specified, this consistency is lost by the AIPW estimator if the y -model and π -model are both incorrectly specified. The lowest standard error is seen using the complete case outcome regression model, despite the smaller sample size due to incomplete cases being discarded. Of the consistent estimators, the standard error is highest when the y -model is incorrectly specified.

When the exposure X is partially missing, the missingness depends on the outcome Y , which creates bias in the complete case estimators. The AIPW estimators should be consistent if (i) either the y -model or the e -model is correct and (ii) either the x -model or the π -model is correct, which is what we see in Table IV. The coverages of the 95% confidence intervals are fairly high for scenarios with a consistent estimator where the y -model is incorrect. We conjecture that this is because the sample size is not sufficient in these cases to attain the asymptotic distribution of the estimator for Δ because with larger

Table III. Properties of the AIPW estimator of Δ (the TIVA effect on post-operative anxiety) from Equation (6) and an analogous expression for μ_0 using 5000 simulations and $n = 1000$ when the outcome Y , post-operative anxiety, is partially missing.

Model specification			Consistent?	Mean	Empirical	% Bias/SE	MSE	% Coverage of 95% CI
y -	e -	π -			SE			
✓	—	—	✓	1.00	0.12	0.5	0.01	94.6
—	✓	—	✓	1.00	0.15	1.7	0.02	93.7
✓	✓	✓	✓	1.00	0.14	0.6	0.02	93.9
×	✓	✓	✓	1.00	0.16	2.1	0.03	93.9
✓	×	✓	✓	1.00	0.15	1.0	0.02	93.8
✓	✓	×	✓	1.00	0.14	0.3	0.02	94.0
×	×	✓	×	1.38	0.16	237.3	0.17	31.7
×	✓	×	×	0.93	0.16	43.0	0.03	92.6
✓	×	×	✓	1.00	0.14	0.8	0.02	93.9
×	×	×	×	1.34	0.16	214.0	0.14	40.7

Ticks and crosses indicate which models were correctly (tick) and incorrectly (cross) specified. A tick in the 'Consistent?' column indicates scenarios where the doubly robust property should result in a consistent estimator. For comparison, the top two lines show two complete case estimators: the estimated regression coefficient from a correctly specified outcome regression model and an IPW estimator using a correctly specified propensity score model.

Table IV. Properties of the AIPW estimator of Δ (the TIVA effect on post-operative anxiety) from Equation (8) and an analogous expression for μ_0 using 5000 simulations and $n = 1000$ when the exposure X , TIVA, is partially missing.

Model specification				Consistent?	Mean	Empirical SE	% Bias/SE	MSE	% Coverage of 95% CI
y -	e -	x -	π -						
✓	—	—	—	×	0.91	0.16	55.8	0.03	91.6
—	✓	—	—	×	0.93	0.20	33.1	0.05	93.6
✓	✓	✓	✓	✓	1.00	0.20	1.8	0.04	94.6
×	✓	✓	✓	✓	1.01	0.21	5.7	0.04	98.5
✓	×	✓	✓	✓	1.00	0.20	0.2	0.04	94.5
✓	✓	×	✓	✓	1.02	0.35	5.1	0.12	95.0
✓	✓	✓	×	✓	1.00	0.19	1.2	0.04	95.1
×	×	✓	✓	×	1.39	0.23	170.3	0.20	58.9
×	✓	×	✓	✓	1.03	0.37	8.5	0.14	98.9
×	✓	✓	×	✓	1.00	0.20	1.2	0.04	99.3
✓	×	×	✓	✓	1.00	0.22	0.3	0.05	94.7
✓	×	✓	×	✓	1.00	0.19	1.5	0.04	94.8
✓	✓	×	×	×	0.91	0.18	50.6	0.04	93.6
×	×	×	✓	×	1.39	0.25	153.0	0.22	67.5
×	×	✓	×	×	1.39	0.21	184.1	0.19	60.2
×	✓	×	×	×	0.92	0.19	43.0	0.04	99.4
✓	×	×	×	×	0.90	0.18	56.0	0.04	92.9
×	×	×	×	×	1.24	0.20	120.0	0.10	85.1

Ticks and crosses indicate which models were correctly (tick) and incorrectly (cross) specified. A tick in the 'Consistent?' column indicates scenarios where the doubly robust property should result in a consistent estimator. For comparison, the top two lines show two complete case estimators: the estimated regression coefficient from a correctly specified outcome regression model and an IPW estimator using a correctly specified propensity score model.

Table V. Properties of the AIPW estimator of Δ (the TIVA effect on post-operative anxiety) using the approximation of Section 3.6 with $M = 50$ data duplications, 5000 simulations and $n = 1000$ when a confounder, pre-operative anxiety, is partially missing.

Model specification				Consistent?	Mean	Empirical SE	% Bias/SE	MSE	% Coverage of 95% CI
y -	e -	z -	π -						
✓	—	—	—	✓	1.00	0.17	0.4	0.03	95.0
—	✓	—	—	✓	1.03	0.20	0.8	0.04	95.8
✓	✓	✓	✓	✓	1.00	0.16	0.5	0.03	93.4
×	✓	✓	✓	✓	1.00	0.17	2.8	0.03	96.4
✓	×	✓	✓	✓	1.00	0.17	0.3	0.03	95.2
✓	✓	×	✓	✓	1.00	0.17	1.0	0.03	93.3
✓	✓	✓	×	✓	1.00	0.16	1.2	0.03	94.2
×	×	✓	✓	×	1.38	0.18	211.8	0.18	42.5
×	✓	×	✓	✓	1.01	0.17	3.4	0.03	95.0
×	✓	✓	×	✓	1.00	0.17	1.6	0.03	94.5
✓	×	×	✓	✓	1.00	0.17	0.8	0.03	94.2
✓	×	✓	×	✓	1.00	0.16	1.3	0.03	94.5
✓	✓	×	×	×	1.00	0.16	0.7	0.03	94.3
×	×	×	✓	×	1.38	0.18	211.5	0.18	40.4
×	×	✓	×	×	1.38	0.18	215.1	0.18	40.5
×	✓	×	×	×	1.01	0.17	3.9	0.03	94.3
✓	×	×	×	×	1.00	0.16	2.2	0.03	94.2
×	×	×	×	×	1.38	0.18	215.7	0.18	37.7

Ticks and crosses indicate which models were correctly (tick) and incorrectly (cross) specified. A tick in the 'Consistent?' column indicates scenarios where the doubly robust property should result in a consistent estimator. For comparison, the top two lines show two complete case estimators: the estimated regression coefficient from a correctly specified outcome regression model and an IPW estimator using a correctly specified propensity score model.

Table VI. Comparison of AIPW estimates of Δ (the TIVA effect on post-operative anxiety) for scenarios with Y or X partially missing, calculated using analytic expressions (Equations (6) and (8) and analogous expressions for μ_0) and using the approximation of Section 3.6 with a range of data replications (M).

Partially missing variable	Method	M	Combination of models							
			(i)		(ii)		(iii)		(iv)	
			Mean	SE	Mean	SE	Mean	SE	Mean	SE
Outcome Y	Exact	N/A	1.00	0.20	1.00	0.20	1.00	0.19	1.35	0.21
	Approximate	5	1.00	0.20	1.00	0.21	1.00	0.20	1.34	0.22
		10	1.00	0.20	1.00	0.21	1.00	0.19	1.35	0.22
		25	1.00	0.20	1.00	0.20	1.00	0.19	1.35	0.22
		50	1.00	0.20	1.00	0.20	1.00	0.19	1.35	0.21
Exposure X	Exact	N/A	1.01	0.20	1.01	0.22	0.92	0.19	1.24	0.20
	Approximate	5	1.01	0.21	1.01	0.23	0.92	0.20	1.24	0.20
		10	1.01	0.21	1.01	0.23	0.92	0.19	1.24	0.20
		25	1.01	0.20	1.01	0.22	0.92	0.19	1.24	0.20
		50	1.01	0.20	1.01	0.22	0.92	0.19	1.24	0.20

A total of 500 simulations were performed with $n = 1000$. Four combinations of models are considered: (i) all models correct, (ii) y -model and π -model correct only, (iii) e -model correct only and (iv) all models incorrect.

sample sizes these coverages reduce to around 95% (results not shown). For moderate sample sizes, such as this, bootstrapping may be an attractive alternative.

When the confounder Z_5 is partially missing, we expect the two complete case estimators to be consistent because the confounder being missing does not depend on the outcome. The AIPW estimators should be consistent if (i) either the y -model or the e -model is correct and (ii) either the z -model or the π -model is correct. The results in Table V agree with this in most cases. However, we see no evidence of inconsistency in some situations where consistency is not guaranteed. This is explained by the fact that our misspecified z -model was, unlike the other misspecified models, not badly misspecified and the confounder Z_5 has little confounding effect on the exposure effect. For consistent estimators, the bootstrap confidence intervals give approximately 95% coverage.

Table VI compares estimates of Δ obtained using the explicit formulae and the approximation using data duplication, for scenarios where Y or X are partially missing. For this example, the estimates are very similar even with five replications of the dataset. With much larger fractions of missing data (e.g. 70%), the accuracy of the approximation with 5–10 replications was lower for this scenario, although 50 replications still gave very good approximations (data not shown).

6. Discussion

Confounding and missing data are both likely to occur when observational data are used to estimate the causal effect of an exposure. In dealing with each of these problems, we face the additional challenge of not knowing the true mechanisms by which the confounding and missing data arise. Standard statistical solutions to the problems are therefore likely to create errors because of model misspecification. In this paper, we have presented estimators that account for both confounding and missing data and that have an element of robustness to the misspecification of the models required. We have derived an explicit expression for the estimator for scenarios where the outcome is partially missing and discussed the connection between our estimator and the one proposed by Davidian *et al* [16]. We have extended this approach to scenarios where the exposure of interest is partially missing and have proposed a novel estimator that is robust to model misspecification for both confounding and missing data. When a confounder is missing, explicit calculation is complex, but we have proposed a novel but simple data duplication method, which can provide close approximations to the estimator. We demonstrated that using 5–10 dataset replications often gives accurate estimators. This parallels standard considerations for multiple imputation, where 5–10 imputed datasets are often used. These parallels could be further explored in future research. Our approximation, however, is computationally demanding, particularly for the bootstrapped confidence intervals; thus, this method may be infeasible for very large datasets.

Our simulation study demonstrated that our AIPW estimators have the robustness properties expected. An important feature of these properties is that the addition of a misspecified model may result in a worse estimate. For example, when the outcome is partially missing, an inverse probability weighted estimate involving only a model for the propensity score is consistent. Our AIPW estimator, however, can be inconsistent even if the propensity score model is correct. We therefore arrive at a similar conclusion to Kang and Schafer: ‘at least in some settings, two wrong models are not better than one’ [4].

We have considered only scenarios where a single variable has missing data. More commonly, many variables are partially missing. Future research will consider the extent to which these methods can be extended to arbitrary missing data patterns. It is likely that methods applied for multiple imputation, such as the use of data augmentation using Markov chain Monte Carlo procedures to approximate the distribution of the missing data [31], will be useful in extending the ‘imputation’ aspect of our AIPW estimators to more general missing data patterns.

Acknowledgements

We are grateful to Professor Paul Myles for providing us with the B-Aware data. We thank three anonymous reviewers, whose comments have greatly improved the manuscript. This work was partly supported by a grant from the National Health and Medical Research Council of Australia.

References

1. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
2. Li L, Shen C, Li X, Robins J. On weighting approaches for missing data. *Statistical Methods in Medical Research* 2011. DOI: 10.1177/0962280211403597. Advance online access.
3. Seaman SR, White I. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research* 2011. DOI: 10.1177/0962280210395740. Advance online access.
4. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 2007; **22**:523–539. DOI: 10.1214/07-STS227.
5. Robins JM, Rotnitzky A, Zhao L. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89**:846–866.
6. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 1995; **90**:122–129.
7. Rotnitzky A, Robins JM. Analysis of semi-parametric regression models with nonignorable non-response. *Statistics in Medicine* 1997; **16**:81–102.
8. Scharfstein DO, Rotnitzky A, Robins J. Adjusting for nonignorable drop-out using semiparametric nonresponse model (with discussion). *Journal of the American Statistical Association* 1999; **94**:1096–1146.
9. Carpenter J, Kenward M, Vansteelandt S. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society, Series (A)* 2006; **169**:571–584.
10. Tsiatis AA. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer: New York, 2006.
11. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; **61**:962–972. DOI: 10.1111/j.1541-0420.2005.00377.x.
12. Holland P. Statistics and causal inference. *Journal of the American Statistical Association* 1986; **81**:945–960.
13. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55. DOI: 10.1093/biomet/70.1.41.
14. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 2004; **23**:2937–2960. DOI: 10.1002/sim.1903.
15. Austin P, Mamdani M. A comparison of propensity score methods: a case-study estimating the effectiveness of post-ami statin use. *Statistics in Medicine* 2006; **25**:2084–2106. DOI: 10.1002/sim.2328.
16. Davidian M, Tsiatis AA, Leon S. Semiparametric estimation of treatment effect in a pretest-posttest study with missing data. *Statistical Science* 2005; **20**:261–301. DOI: 10.1214/088342305000000151.
17. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology* 2009; **20**:3–5. DOI: 10.1097/EDE.0b013e31818ef366.
18. Rubin DB. Comment on ‘Randomization analysis of experimental data in the Fisher randomization test’ by Basu. *Journal of the American Statistical Association* 1980; **75**:591–593.
19. Westreich D, Cole SR. Invited commentary: positivity in practice. *American Journal of Epidemiology* 2010; **171**:674–677. DOI: 10.1093/aje/kwp436.
20. Dawid A. Conditional independence in statistical-theory. *Journal of the Royal Statistical Society B—Methodological* 1979; **41**:1–31.
21. Snowden JM, Rose S, Mortimer KM. Implementation of G-Computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology* 2011; **173**:731–738. DOI: 10.1093/aje/kwq472.
22. Vansteelandt S, Keiding N. Invited commentary: G-Computation—lost in translation? *American Journal of Epidemiology* 2011; **173**:739–742. DOI: 10.1093/aje/kwq474.
23. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine* 2010; **29**:2920–2931. DOI: 10.1002/sim.3944.

24. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research* 2011. DOI: 10.1177/0962280210394469. Advance online access.
25. Cole SR, Hernan MA. Fallibility in estimating direct effects. *International Journal of Epidemiology* 2002; **31**:163–165. DOI: 10.1093/ije/31.1.163.
26. Stata Statistical Software: Release 11. StataCorp LP: College Station, TX, USA, 2009.
27. SAS. SAS Institute Inc.: Cary, NC, USA.
28. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2006. <http://www.R-project.org>, ISBN 3-900051-07-0.
29. Myles P, Leslie K, McNeil J, Forbes A, Chan M, for the B-Aware trial group. Bispectral index monitoring to prevent awareness during anaesthesia: the B-Aware randomised controlled trial. *Lancet* 2004; **363**:1757–1763. DOI: 10.1016/S0140-6736(04)16300-9.
30. Zigmond A, Snaith R. The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica* 1983; **67**:361–370.
31. Schafer JL. *Analysis of Incomplete Multivariate Data*. Chapman & Hall: London, 1997.