The final, definitive version of this paper has been published in Statistical Methods in Medical Research (2017), 26(1): 248–267 DOI: http://dx.doi.org/10.1177/0962280214544251. Published by SAGE Publishing, All rights reserved.

Statistical methods for incomplete data: Some results on model misspecification

MICHAEL McISAAC

Department of Public Health Science,

Queen's University, Kingston, ON, K7L 3N6, Canada

E-mail: mcisaacm@queensu.ca

RICHARD J. COOK

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

Summary

Inverse probability weighted estimating equations and multiple imputation are two of the most studied frameworks for dealing with incomplete data in clinical and epidemiological research. We examine the limiting behaviour of estimators arising from inverse probability weighted estimating equations, augmented inverse probability weighted estimating equations and multiple imputation when the requisite auxiliary models are misspecified.We compute limiting values for settings involving binary responses and covariates and illustrate the effects of model misspecification using simulations based on data from a breast cancer clinical trial. We demonstrate that, even when both auxiliary models are misspecified, the asymptotic biases of double-robust augmented inverse probability weighted estimators are often smaller than the asymptotic biases of estimators arising from complete-case analyses, inverse probability weighting or multiple imputation. We further demonstrate that use of inverse probability weighting or multiple imputation with slightly misspecified auxiliary models can actually result in greater asymptotic bias than the use of naïve, complete case analyses. These asymptotic results are shown to be consistent with empirical results from simulation studies.

Keywords: asymptotic bias, asymptotic variance, augmented inverse probability weighting, double robust, incomplete data, inverse probability weighting, model misspecification, multiple imputation

1 INTRODUCTION

Failure to collect intended data in clinical and epidemiological research can seriously compromise the integrity of a study by rendering standard complete-case estimators inconsistent [12]. *Ad hoc*

approaches for dealing with incomplete data such as non-responder imputation or lastobservation-carried-forward are generally not recommended since they only lead to consistent estimators under strong implicit assumptions [7, 11, 15], and typically give conservative variance estimates. More refined procedures based on inverse probability weighted estimating equations or multiple imputation rely on auxiliary models to exploit information available in subjects with incomplete data. The auxiliary models in these two frameworks rely on quite different assumptions but, subject to their correct specification, consistent estimators may be obtained when data are *missing at random* [12]. These approaches can also be used to conduct sensitivity analyses in setting where the data are thought to be *missing not at random* [11, 23, 25].

Multiple imputation [12] involves augmenting the available data to create several complete pseudodatasets. Each of these pseudo-datasets is made complete by randomly drawing from an imputation model whenever information is missing. There has been much discussion and research regarding the development of suitable imputation models, and care must be taken to ensure that variance estimation is valid [25, 26, 31].

Inverse probability weighted estimating equations [13, 17, 21] involve restricting attention to individuals with complete data but achieve consistent estimation by weighting each contribution according to the inverse probability that the data are complete. The so-called selection models are specified to determine the weights and ensure suitable adjustment for the "bias" sample obtained by restricting to individuals with complete data. Augmented inverse probability weighted estimating equations extend the inverse probability weighted approach to increase the efficiency and robustness of the resultant estimator [21, 29]. With augmented inverse probability weighted analyses, consistent estimators result if one or both of the auxiliary models are correctly specified; the term "double-robustness" was coined to reflect this [20].

The efficiencies of weighting and multiple imputation methods have been compared in a variety of settings when the auxiliary model assumptions are correct [3, 18, 25, 27], but limited empirical work has been directed at comparing such methods when the auxiliary model assumptions are incorrect. Bang and Robins [1] and Kang and Schafer [9] both conducted simulation studies to compare the empirical performance of double-robust (DR) and outcome-regression estimators of a mean. Bang and Robins highlight the additional robustness of the DR estimator which "offe the analyst two chances to make nearly correct inference" [1], while Kang and Schafer argued that "in at least some settings, two wrong models are not better than one" [9]. Kang and Schafer suggested that the simulation study conducted by Bang and Robins was unduly favourable toward their DR methods, while Robins and Wang [22] remarked that "[Kang and Schafer's] chosen data-generating distribution was as if optimized" for the outcome-regression estimator. In this paper, we give an accessible presentation of the various approaches to estimation with incomplete data and examine the asymptotic properties of these resulting estimators when one or both of the auxiliary models are misspecified. We focus on the estimation of regression parameters and consider a common outcomeregression model to examine the effects of misspecification in analyses based on multiple imputation and augmented inverse probability weighted estimating equations. We also examine the empirical behaviour of these estimators through a simulation study with parameter values chosen to reflect the setting of a breast cancer clinical trial. Throughout this paper, misspecification is considered in the auxiliary models through the omission of a common confounder, or through the omission of an interaction term; in this way, we hope to allay the concerns raised by some authors when simulation studies allowed different auxiliary models to adjust for different confounders [9].

The remainder of this paper is organized as follows. In Section 2, we define inverse probability weighted and augmented inverse probability weighted estimating equations and describe the procedure of multiple imputation for incomplete response data. We also demonstrate in Section 2 how the asymptotic properties of the resulting estimators can be derived using the results of Robins *et al.* [21], Robins and Wang [22], Pierce [16] and Lawless *et al.* [10], and we review the simplifications

that occur under correct model specification. In Section 3, we focus on the simple case in which all variables are binary and derive the asymptotic bias and variance of these estimators when model assumptions are incorrectly specified. In Section 4, we demonstrate the empirical properties of the estimators under misspecification by simulating incomplete data consistent with a breast cancer clinical trial. Concluding remarks are made in Section 5.

2 Estimators and their Limiting Behaviour

Consider a random sample of N individuals yielding data $\{(Y_i, X_i), i = 1, \ldots, N\}$ where Y_i is the univariate response and X_i is a vector of explanatory covariates for individual i. Suppose that interest lies in estimating the $p \times 1$ vector of regression parameters for the conditional mean model of Y given X, $\mu(X; \alpha)$. An estimator of α can be found as the solution to the estimating equation

$$0 = \sum_{i=1}^{N} U_i(\alpha) \ = \sum_{i=1}^{N} h(X_i)[Y_i - \mu(X_i; \alpha)], \tag{1}$$

where $h(X)$ is a known $p \times 1$ function of X such that $E[h(X)\partial\mu(X_i; \alpha)/\partial\alpha']$ is non-singular [21]. If the conditional distribution of the response is in the exponential family, the canonical link is used to relate the mean $\mu(X_i; \alpha)$ to the linear predictor $X_i'\alpha$, and $h(X_i)$ is specified as X_i', then (1) is the score function and the root of this equation is the maximum likelihood estimator for α.

More generally, under mild regularity conditions which we assume henceforth [21, 22], the solution to the estimating equation in (1), denoted $\widehat{\alpha}$, is a $\sqrt{N}$-consistent estimator of the parameter of interest α_0 satisfying $E[U_i(\alpha)] = 0$, with

$$N^{1/2}(\widehat{\alpha} - \alpha_0) \xrightarrow{D} MVN(0, \mathcal{A}_0^{-1}\mathcal{B}_0[\mathcal{A}_0^{-1}]'), \tag{2}$$

where $\mathcal{A}_0^{-1}\mathcal{B}_0[\mathcal{A}_0^{-1}]'$ is called the asymptotic variance of $\widehat{\alpha}$ and

$$\mathcal{A}_0 = -E[\partial U_i(\alpha)/\partial\alpha']_{\alpha=\alpha_0} \text{ and } \mathcal{B}_0 = E[U_i(\alpha)U_i'(\alpha)]_{\alpha=\alpha_0} .$$

2.1 Analysis with Incomplete Data

If responses are only observed for a subset of individuals in the sample, we let R_i be the indicator that Y_i is observed. Estimation of α based on (1) would require solving $0 = \sum_{i=1}^{N} \big[R_iU_i(\alpha) + (1 - R_i)U_i(\alpha)\big]$, but this clearly cannot be done since $U_i(\alpha) = h(X_i)[Y_i - \mu(X_i; \alpha)]$ is unknown when $R_i = 0$.

A natural estimator when data are incomplete is obtained by restricting attention to individuals who provide complete information. That is, we could estimate α_0 with the complete-case (CC) estimator $\widehat{\alpha}_{\text{cc}}$ obtained by solving the estimating equation

$$0 = \sum_{i=1}^{N} R_i \cdot U_i(\alpha) \ . \tag{3}$$

The limiting behaviour of this estimator can be derived analogously to (2) and it can be shown that $\widehat{\alpha}_{\text{cc}}$ will consistently estimate α_{cc}, the solution to $E[R_iU_i(\alpha)] = 0$ [24]. Thus

$$N^{1/2}(\widehat{\alpha}_{\text{cc}} - \alpha_{\text{cc}}) \xrightarrow{D} MVN(0, \mathcal{A}_{\text{cc}}^{-1}\mathcal{B}_{\text{cc}}[\mathcal{A}_{\text{cc}}^{-1}]'),$$

where

$$\mathcal{A}_{\text{cc}} = -E[R_i \ \partial U_i(\alpha)/\partial\alpha']_{\alpha=\alpha_{\text{cc}}} \text{ and } \mathcal{B}_{\text{cc}} = E[R_iU_i(\alpha)U_i'(\alpha)]_{\alpha=\alpha_{\text{cc}}} .$$

Here α_{cc} is equivalent to α_0 if $E[R_i U_i(\alpha)] = E[U_i(\alpha)]$, which in turn occurs if $Y \perp R|X$ (i.e. if Y and R are conditionally independent given X). Thus, complete-case analyses will consistently estimate α_0 only in the special case where, for a given X, the subsets of individuals that are completely observed are representative of the original sample; i.e. the missing data mechanism is *missing at random* (MAR) [12] in the presence of X. We will refer to the difference between α_{cc} and α_0 as the *asymptotic bias* of the CC estimator.

More sophisticated methods for accommodating missing data are possible if auxiliary information is available and suitable assumptions are made. We suppose in what follows that there exists an auxiliary covariate vector V which is known for all individuals and which is associated with both the response and the missingness indicator in such a way that $Y \perp R|X, V$; that is conditioning on V, in addition to X, renders the missingness mechanism MAR. When the objective is to fit simple descriptive response models, there may be several covariates known to be associated with the response that are not contained in X. Any such covariates that are also associated with missingness could be represented in V. In smoking prevention studies, for example, social model risk scores give useful information about children's peers and risk of smoking, but this also reflects risk of noncompliance and study withdrawal [5]. We explore the use of such variables in what follows.

2.2 Multiple Imputation

Imputation allows for use of the complete-data estimating equation (1) by replacing missing responses with imputed values; if one could replace missing Y_i values with suitable Y_i^{imp}, then the complete-data estimator could be approximated by solving $0 = \sum_{i=1}^{N} \big[R_i \cdot h(X_i)[Y_i - \mu(X_i;\alpha)] + (1 - R_i) \cdot h(X_i)[Y_i^{imp} - \mu(X_i;\alpha)]\big]$. Two challenges arise with this approach: (i) one must find appropriate values for imputation in order to avoid introducing bias, and (ii) treating the imputed values as known will result in underestimation of the true variability in the estimator. The first challenge can be addressed by relying on implicit models to define a measure of similarity between individuals and replacing missing responses with observed responses from "similar" individuals (e.g. hot-deck imputation [30] or the approximate Bayesian Bootstrap [12]), or through *parametric* imputation by simulating missing responses using random draws from an explicit imputation model. The second challenge can be addressed by simulating missing responses multiple times (say, M times) to create M complete pseudo-datasets and examining the variability between the estimators obtained from the multiple "complete" datasets. Thus correct specification of the imputation model ensures consistent estimators, while imputing multiple times enables estimation of the variability over different imputed samples. Naive methods of single imputation, such as non-responder imputation or last-observation carried forward, are commonly employed [15], but only result in consistent estimators if strong implicit assumptions hold and typically result in incorrect variance estimates [7, 14]. We therefore focus attention on a parametric multiple imputation procedure and examine asymptotic biases that result if the parametric modelling assumptions do not hold.

A common framework for this type of imputation is *proper* parametric multiple imputation, which has a Bayesian flavour in that the parameter indexing the imputation model is itself randomly drawn from a "posterior density" of the parameter given the observed data [12, 30]. Commonly used formulae for the asymptotic variance of estimators obtained under proper multiple imputation are given by Rubin [25]. We consider *improper* multiple imputation here, however, in which missing responses are drawn from an imputation model based on the maximum likelihood estimate of the parameters in the imputation model. This type of multiple imputation procedure is advocated by Wang and Robins [30] who show that the resulting estimator has a strictly smaller asymptotic variance than the one obtained by proper imputation when models are correctly specified and M is finite; as $M \to \infty$, these estimators become asymptotically equivalent. [30]

In order to facilitate comparisons with the augmented inverse probability weighted estimators, we

consider a non-iterative estimator based on a conditional imputation model $g(Y|X,V,R;\eta)$ [22] that is known up to the $q \times 1$ parameter η indexing $m(X,V;\eta)$, the model for the conditional mean of Y given the available covariates (as in the mean score imputation of Clayton *et al.* [6]); model of this sort arises in the outcome-regression approach employed in Bang and Robins [1] and Kang and Schafer [9]. If $Y \perp R|X,V$, then a consistent estimator of η, the parameter indexing the imputation model, can be obtained as the solution to the complete-case estimating equation

$$0 = \sum_{i=1}^{N} W_i(\eta) = \sum_{i=1}^{N} R_i \cdot h_{\text{mi}}(X_i,V_i)[Y_i - m(X_i,V_i;\eta)], \tag{4}$$

where it is assumed that (i) the $q \times 1$ function $h_{\text{mi}}(X_i,V_i)$ is a defined analogously to $h(X_i)$ in (1), (ii) $\widehat{\eta}$ converges to a limit η_{mi} and (iii) there exists an appropriate finite variance influence function of $\widehat{\eta}$ with finite variance.

A multiple imputation estimator based on M imputations can be found by solving

$$0 = \sum_{i=1}^{N} U_i^{\text{mi}}(\alpha,\widehat{\eta}) = \sum_{i=1}^{N} M^{-1} \sum_{j=1}^{M} h(X_i)[Y_{ij}^{imp}(\widehat{\eta}) - \mu(X_i;\alpha)]$$

where $Y_{ij}^{imp}(\widehat{\eta})$ are drawn independently from $g(Y|X,V,R;\widehat{\eta})$ if $R_i = 0$ and $Y_{ij}^{imp}(\widehat{\eta}) = Y_i$ if $R_i = 1$ [22]. Under mild regularity conditions, the resultant estimator $\widehat{\alpha}_{\text{mi}}$ has the property that, for α_{mi} solving $E\big[h(X_i)[Y_{ij}^{imp}(\widehat{\eta}) - \mu(X_i;\alpha)]\big] = 0$,

$$N^{1/2}(\widehat{\alpha}_{\text{mi}} - \alpha_{\text{mi}}) \xrightarrow{D} MVN(0, \mathcal{A}_{\text{mi}}^{-1}\mathcal{B}_{\text{mi}}[\mathcal{A}_{\text{mi}}^{-1}]'),$$

where,

$$\mathcal{A}_{\text{mi}} = E[-\partial U_i^{\text{mi}}(\alpha,\eta_{\text{mi}})/\partial\alpha']_{\alpha=\alpha_{\text{mi}}} = E[-\partial U_i(\alpha)/\partial\alpha']_{\alpha=\alpha_{\text{mi}}}$$

and, temporarily suppressing the dependence on $(\alpha_{\text{mi}},\eta_{\text{mi}})$,

$$\mathcal{B}_{\text{mi}} = E[U_i^{\text{mi}}U_i^{\text{mi}\prime}] - \kappa_{\text{mi}}E[W_iU_i^{\text{mi}\prime}] - E[U_i^{\text{mi}}W_i']\kappa_{\text{mi}}' + \kappa_{\text{mi}}E[W_iW_i']\kappa_{\text{mi}}'$$

with $\kappa_{\text{mi}} = E[U_i^{\text{mi}}S_{\text{mis}}']E[\partial W_i/\partial\eta']^{-1}$ and $S_{\text{mis}} = \partial \log g(Y|X,V,R;\eta)/\partial\eta$. In the aforementioned expressions, expectations are taken with respect to the imputation distribution $[g(Y|X,V,R=0;\eta_{\text{mi}})]^{1-R}$ $[f_0(Y|X,V,R=1)]^R$ $f_0(X,V,R)$, where $f_0(\cdot)$ represents the true joint density [22]. These asymptotic results hold even if the imputation model is incorrectly specified or incompatible with the response model [22]. Misspecification of the imputation model will, however, directly impact the asymptotic bias of the resultant estimator $(\alpha_{\text{mi}} - \alpha_0)$; this imputation approach is asymptotically unbiased when the imputation model is correctly specified.

2.3 Inverse Probability Weighting

The auxiliary data V can alternatively be used to help make an analysis based solely on the individuals with complete data more suitable. This is done by reweighting these observations so the completely-observed pseudo-sample is representative of the original sample. So in this framework, we focus on individuals with $R_i = 1$, but instead of solving the complete-case estimating equation (3) which only yields a consistent estimator if $Y \perp R|X$, we solve the Horvitz-Thompson-style estimating equation

$$0 = \sum_{i=1}^{N} U_i^{\text{ipw}}(\alpha,\delta) = \sum_{i=1}^{N} \frac{R_i}{\pi(X_i,V_i;\delta)} \cdot U_i(\alpha),$$

where $\pi(X_i, V_i; \delta)$ is a model for the "selection" probability $P(R_i = 1|X_i, V_i; \delta)$ [21]; this gives a consistent estimator of α if $Y \perp R|X, V$. We call this the inverse probability weighted (IPW) estimating equation and refer to the resulting estimator, $\widehat{\alpha}_{\text{ipw}}$, as the IPW estimator.

The unknown $s \times 1$ selection parameter δ can be replaced by $\widehat{\delta}$, the solution to an appropriate estimating equation

$$0 = \sum_{i=1}^{N} S_i(\delta) = \sum_{i=1}^{N} h_{\text{ipw}}(X_i, V_i)[R_i - \pi(X_i, V_i; \delta)],$$

where we assume that the known $s \times 1$ function $h_{\text{ipw}}(X_i, V_i)$ is defined analogously to $h(X_i)$ and that $\widehat{\delta}$ converges to a limit which we denote δ_{ipw}. Logistic regression is often used to estimate δ since R is binary.

Provided that $\pi(X_i, V_i; \delta)$ is bounded away from 0, this IPW estimator will consistently estimate α_{ipw}, the root of the function $E[U_i^{\text{ipw}}(\alpha, \delta_{\text{ipw}})]$ (i.e. the value of α such that $E[U_i^{\text{ipw}}(\alpha, \delta_{\text{ipw}})] = 0$), and

$$N^{1/2}(\widehat{\alpha}_{\text{ipw}} - \alpha_{\text{ipw}}) \xrightarrow{D} MVN(0, \mathcal{A}_{\text{ipw}}^{-1}\mathcal{B}_{\text{ipw}}[\mathcal{A}_{\text{ipw}}^{-1}]'),$$

where, as shown in Appendix A,

$$\mathcal{A}_{\text{ipw}} = -E[\partial U_i^{\text{ipw}}(\alpha, \delta)/\partial\alpha']_{\alpha=\alpha_{\text{ipw}};\ \delta=\delta_{\text{ipw}}}$$

and, again suppressing the dependence on $(\alpha_{\text{ipw}}, \delta_{\text{ipw}})$

$$\mathcal{B}_{\text{ipw}} = E[U_i^{\text{ipw}}U_i^{\text{ipw}\prime}] - \kappa_{\text{ipw}}E[S_iU_i^{\text{ipw}\prime}] - E[U_i^{\text{ipw}}S_i']\kappa_{\text{ipw}}' + \kappa_{\text{ipw}}E[S_iS_i']\kappa_{\text{ipw}}'$$

with $\kappa_{\text{ipw}} = E[\partial U_i^{\text{ipw}}/\partial\delta']E[\partial S_i/\partial\delta']^{-1}$.

If S_i is the score function based on the true model (i.e. if δ is modelled correctly and $\widehat{\delta}$ consistently estimates the true δ_0), then $E[R_i\pi^{-1}U_i(\alpha)] = E[U_i(\alpha)]$, so $\alpha_{\text{ipw}} = \alpha_0$ and $\mathcal{A}_{\text{ipw}} = \mathcal{A}_0$. Furthermore, in this case, $E[\partial S_i(\delta)/\partial\delta'] = -E[S_i(\delta)S_i'(\delta)]$ and $E[\partial U_i^{\text{ipw}}(\alpha, \delta)/\partial\delta'] = -E[U_i^{\text{ipw}}(\alpha, \delta)S_i'(\delta)]$ by the "generalized information equality" of Pierce [16], and the asymptotic variance of this IPW estimator

$$\mathcal{A}_0^{-1}\big\{E[U_i^{\text{ipw}}U_i^{\text{ipw}\prime}] - E[U_i^{\text{ipw}}S_i']E[S_iS_i']^{-1}E[U_i^{\text{ipw}}S_i']'\big\}[\mathcal{A}_0^{-1}]'.$$

We also note here that if the true δ_0 was known instead of estimated, then the asymptotic variance would be $\mathcal{A}_0^{-1}E[U_i^{\text{ipw}}U_i^{\text{ipw}\prime}][\mathcal{A}_0^{-1}]'$ [10, 21].

2.4 Augmented Inverse Probability Weighting

Robins, Rotnitzky, and Zhao [21] showed that the IPW estimating equation could be augmented to better exploit the partial information available from individuals with incomplete data. As a result, an augmented inverse probability weighted (AIPW) estimating equation of the form

$$0 = \sum_{i=1}^{N} U_i^{\text{aipw}}(\alpha, \widehat{\delta}, \widehat{\eta}) = \sum_{i=1}^{N} R_i\pi^{-1}(X_i, V_i; \widehat{\delta})U_i(\alpha) - [R_i - \pi(X_i, V_i; \widehat{\delta})]\pi^{-1}(X_i, V_i; \widehat{\delta})\phi(X_i, V_i; \alpha, \widehat{\eta})$$

can be asymptotically more efficient than the IPW estimator. In the absence of further auxiliary covariates, the optimal choice for the augmentation function can be asymptotically more efficient than the IPW estimator. $\phi(\cdot)$ is $E[U(\alpha)|X, V]$ [21, 29, 32]. In practice, this optimal augmentation term can be approximated by specifying an appropriate conditionalmean model $m(X_i, V_i; \eta)$, as in the imputation approach. We will denote the resultant augmentation term by $U_i(\alpha, \eta)$, where by (1),

$$U_i(\alpha, \eta) = h(X_i)[m(X_i, V_i; \eta) - \mu(X_i; \alpha)].$$

For the remainder of this paper, we will focus on this AIPW approach which requires specification of both an "imputation" model, $m(X_i, V_i; \eta)$, and a "selection" model, $\pi(X_i, V_i; \delta)$. The resulting AIPW estimator, $\widehat{\alpha}_{\text{aipw}}$ consistently estimates α_{aipw}, the root of $E[U_i^{\text{aipw}}(\alpha, \delta_{\text{ipw}}, \eta_{\text{mi}})]$ and the estimating function can equivalently be written as [9]

$$0 = \sum_{i=1}^{N} U_i^{\text{aipw}}(\alpha, \widehat{\delta}, \widehat{\eta}) = \sum_{i=1}^{N} U_i(\alpha, \widehat{\eta}) + R_i \pi(X_i, V_i; \widehat{\delta})^{-1}[U_i(\alpha) - U_i(\alpha, \widehat{\eta})]. \tag{5}$$

The estimator $\widehat{\alpha}_{\text{aipw}}$ satisfies

$$N^{1/2}(\widehat{\alpha}_{\text{aipw}} - \alpha_{\text{aipw}}) \xrightarrow{D} MVN(0, \mathcal{A}_{\text{aipw}}^{-1} \mathcal{B}_{\text{aipw}} [\mathcal{A}_{\text{aipw}}^{-1}]'),$$

where, as shown in Appendix A,

$$\mathcal{A}_{\text{aipw}} = -E[\partial U_i(\alpha)/\partial \alpha']_{\alpha = \alpha_{\text{aipw}}}$$

and, suppressing the dependence on $(\alpha_{\text{aipw}}, \delta_{\text{aipw}})$,

$$\begin{aligned}
\mathcal{B}_{\text{ipw}} =& E[U_i^{\text{aipw}} U_i^{\text{aipw}\prime}] + \kappa_{\text{ipw}}^* E[S_i W_i'][\kappa_{\text{mi}}^*]' + \kappa_{\text{mi}}^* E[W_i S_i'][\kappa_{\text{ipw}}^*]' \\
&- \kappa_{\text{ipw}}^* E[S_i U_i^{\text{aipw}\prime}] - E[U_i^{\text{aipw}} S_i'][\kappa_{\text{ipw}}^*]' + \kappa_{\text{ipw}}^* E[S_i S_i'][\kappa_{\text{ipw}}^*]' \\
&- \kappa_{\text{mi}}^* E[W_i U_i^{\text{aipw}\prime}] - E[U_i^{\text{aipw}} W_i'][\kappa_{\text{mi}}^*]' + \kappa_{\text{mi}}^* E[W_i W_i'][\kappa_{\text{mi}}^*]',
\end{aligned}$$

where $\kappa_{\text{ipw}}^* = E[\partial U_i^{\text{aipw}}/\partial \delta'] E[\partial S_i/\partial \delta']^{-1}$, and $\kappa_{\text{mi}}^* = E[\partial U_i^{\text{aipw}}/\partial \eta'] E[\partial S_i/\partial \delta']^{-1}$.

The AIPW approach is "double robust" in the sense that the estimator will be asymptotically unbiased for data that are MAR if either auxiliary model is correctly specified. This is easy to see from the form of (5) by noting that correct specification of the so-called imputation model leads to $U_i(\alpha, \eta_0) = E_{Y|R,X,V}[U_i(\alpha)]$ so $E[U_i^{\text{aipw}}(\alpha_0, \delta_{\text{ipw}}, \eta_0)] = E[U_i(\alpha_0)] = 0$. Moreover, with correct specification of the selection model, $E_{R|Y,X,V}\{R_i\}\pi(X_i, V_i; \delta_0)^{-1} = 1$, so $E[U_i^{\text{aipw}}(\alpha_0, \delta_0, \eta_{\text{mi}})] = E_{Y,X,V}[U_i(\alpha_0)] = 0$; in this case, the asymptotic variance matrix can be simplified through application of the generalized information equality:

$$E[\partial U_i^{\text{aipw}}/\partial \delta'] = -E[U_i^{\text{aipw}} S_i'] \qquad \text{and} \qquad E[\partial S_i/\partial \delta'] = E[S_i S_i']$$

(see Appendix A).

3 Asymptotic Behaviour of Estimators Under Misspecified Auxiliary Models

We now consider simple violations to the model assumptions necessary for achieving asymptotically unbiased estimators through inverse probability weighting and multiple imputation. Suppose V can render the response and missingness indicator conditionally independent – i.e. $Y \not\perp R|X$, but $Y \perp R|X, V$. Availability of V in this case results in a MAR mechanism and asymptotically unbiased estimation is possible through the methods of analysis described in the previous section if the necessary models are correctly specified. We consider violations of model assumptions that arise from neglecting to accommodate the interaction terms between V and X in the models.

Suppose that Y, X, V and R are scalar binary variables and arise according to the models

$$E[Y|X, V; \eta] = \text{expit}(\eta_1 + \eta_x X + \eta_v V + \eta_{xv} XV), \tag{6}$$

and

$$P(R=1|Y,X,V;\delta)=\text{expit}(\delta_1+\delta_x X+\delta_v V+\delta_{xv}XV), \tag{7}$$

where X and V are independent. Further suppose the model of interest for the conditional mean response is

$$\mu(X;\alpha)=\text{expit}(\alpha_1+\alpha_x X),$$

and we consider the optimally efficient complete data estimating function $U_i(\alpha)=[1,X_i]'[Y_i-\mu(X;\alpha)]$. Note that the true value of α_0 can be recovered from η_0 and $E[V]$ by exploiting the fact that here

$$E[Y|X;\alpha_0]=E_V\{E[Y|X,V;\eta_0]\}. \tag{8}$$

We consider the limiting behaviour of estimators that attempt to accommodate the missing data using the incorrectly specified imputation model

$$m(X,V;\eta^*)=\text{expit}(\eta_1^*+\eta_x^* X+\eta_v^* V),$$

and selection model

$$\pi(X,V;\delta^*)=\text{expit}(\delta_1^*+\delta_x^* X+\delta_v^* V),$$

where the maximum likelihood estimators arise by solving the estimating equations

$$0=\sum_{i=1}^{N}S_i(\delta^*)=\sum_{i=1}^{N}[1,X_i,V_i]'[R_i-\pi(X_i,V_i;\delta^*)], \tag{9}$$

$$0=\sum_{i=1}^{N}W_i(\eta^*)=\sum_{i=1}^{N}R_i\cdot[1,X_i,V_i]'[Y_i-m(X_i,V_i;\eta^*)]. \tag{10}$$

Multiple imputation then involves replacing missing values of Y with simulated Bernoulli-distribute data with conditional mean $m(X,V;\widehat{\eta})$, i.e. we use $g(Y|X,V,R;\widehat{\eta})=\left[m(X,V;\widehat{\eta})^Y\cdot(1-m(X,V;\widehat{\eta}))^{1-Y}\right]^{1-R}$.

For further insight to the asymptotic biases resulting from model misspecification, consider the asymptotic biases that result from a specified parameter set where $\eta_x=\delta_x=0$, $E[Y]=0.5$, $E[X]=0.5$, $E[V]=0.5$, and $E[R]=0.5$, and we consider the effect of a range of values for η_v, δ_v, η_{xv}, and δ_{xv}. The asymptotic bias of estimators of the log odds ratios (α_x) using the above methods can be found in Figure 1; the asymptotic bias of the complete-case (CC) estimator is also included for comparison with the multiple imputation (MI), inverse probability weighted (IPW) and augmented inverse probability weighted (AIPW) estimators. Coverage probabilities for nominal 95% confidence intervals are given in Figure 2. These coverage probabilities are calculated using the limiting distributions (see Section 2) and are functions of both the asymptotic bias and variance of the estimators; the corresponding empirical coverage probabilities are explored in Section 4.

The panels in Figures 1 and 2 display information for the same four cases: in panel 1 (top left), only the selection model is misspecified (we set $\delta_{xv}\neq 0$ and $\eta_{xv}=0$); in panel 2 (top right), only the imputation model is misspecified (we set $\delta_{xv}=0$ and $\eta_{xv}\neq 0$); in panels 3 and 4 (bottom row), both the imputation model and the missingness model are misspecified (we set both $\delta_{xv}\neq 0$ and $\eta_{xv}\neq 0$); the choice of δ_v differs between panels 3 (bottom left) and 4 (bottom right).

Not surprisingly, all estimators were consistent when the required model assumptions were satisfied (see Figure 1). When $\eta_{xv}=0$, the imputation model was correctly specified and the MI estimator was asymptotically unbiased (panel 1); when $\delta_{xv}=0$, the selection model was correctly specified

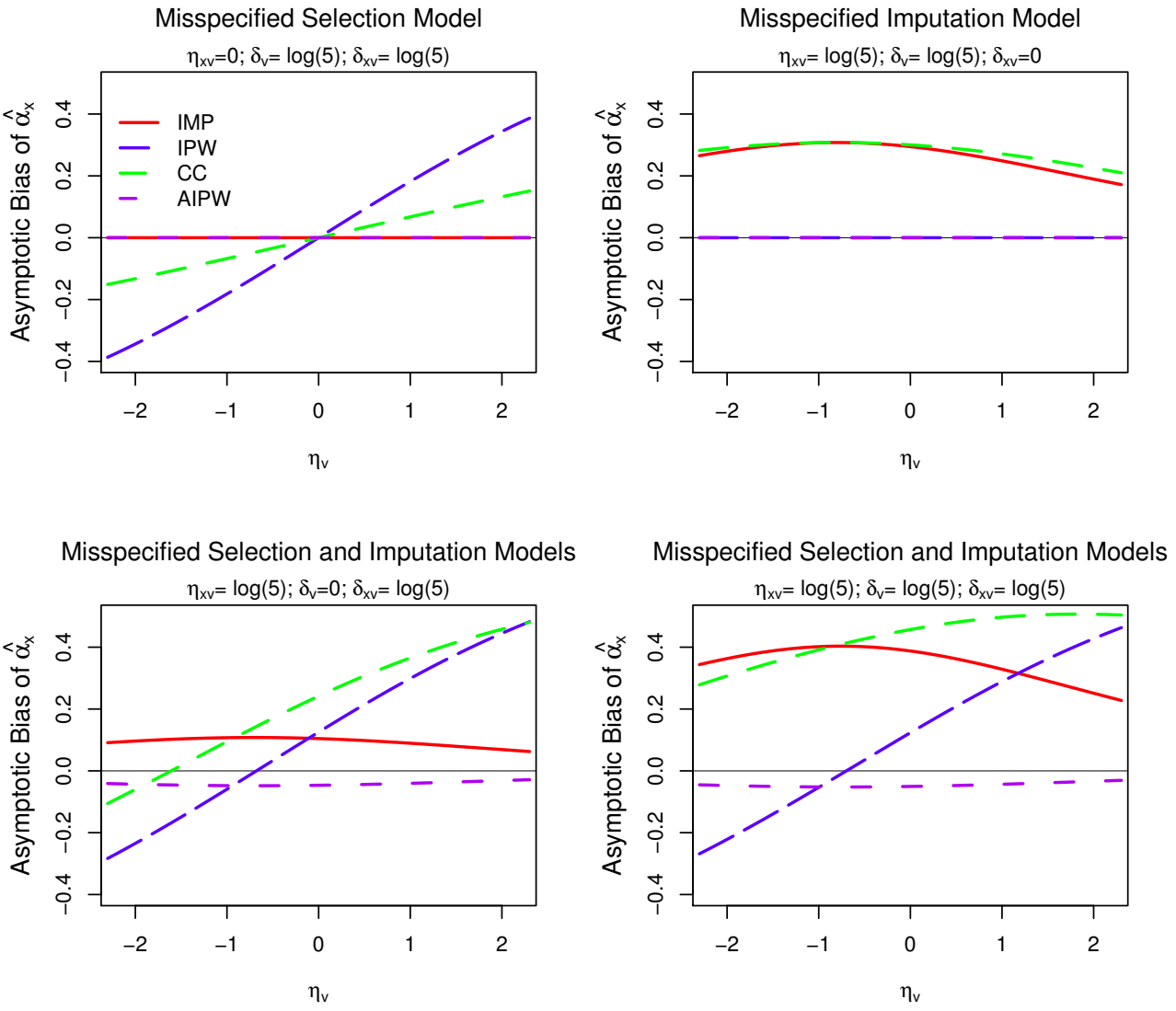

Figure 1: Asymptotic bias: difference between the truth and the limiting value of estimators of log odds ratios from analyses with potentially misspecified models when the response is incomplete; the four panels allow for a range of values of η_x, δ_v, η_{xv} and δ_{xv}, while taking $\eta_x = \delta_x = 0$, $E[Y] = 0.5$, $E[X] = 0.5$, $E[V] = 0.5$, and $E[R] = 0.5$.

and the IPW estimator was asymptotically unbiased (panel 2); when either $\eta_{xv} = 0$ or $\delta_{xv} = 0$, the AIPW estimator was asymptotically unbiased (panels 1 and 2). The double robustness of the AIPW analysis is observable here since, unlike the other methods of analysis, the AIPW estimators were asymptotically unbiased unless both models were incorrectly specified. When both models were incorrectly specified, no method of analysis was universally best; however, the AIPW estimators generally demonstrated the least asymptotic bias (panels 3 and 4).

Interestingly, the AIPW estimator often had the largest asymptotic variance of the considered methods when models were misspecified; however, this apparent asymptotic inefficiency is offset by the greater robustness to model misspecification. This can be seen by examining Figure 2 which displays the probability that the true value of α_x will be contained in a nominal 95% confidence interval for α_x based on a sample of $N = 1000$ individuals, given by $(\widehat{\alpha}_x \pm 1.96\sqrt{asvar(\widehat{\alpha}_x)/1000})$, where $asvar(\widehat{\alpha}_x)$ represents the asymptotic variance derived in Section 2.

Here the AIPW analysis is the only approach which results in confidence intervals that are consistently near the nominal 95% level for all considered parameter sets (see Figure 2). It is also important

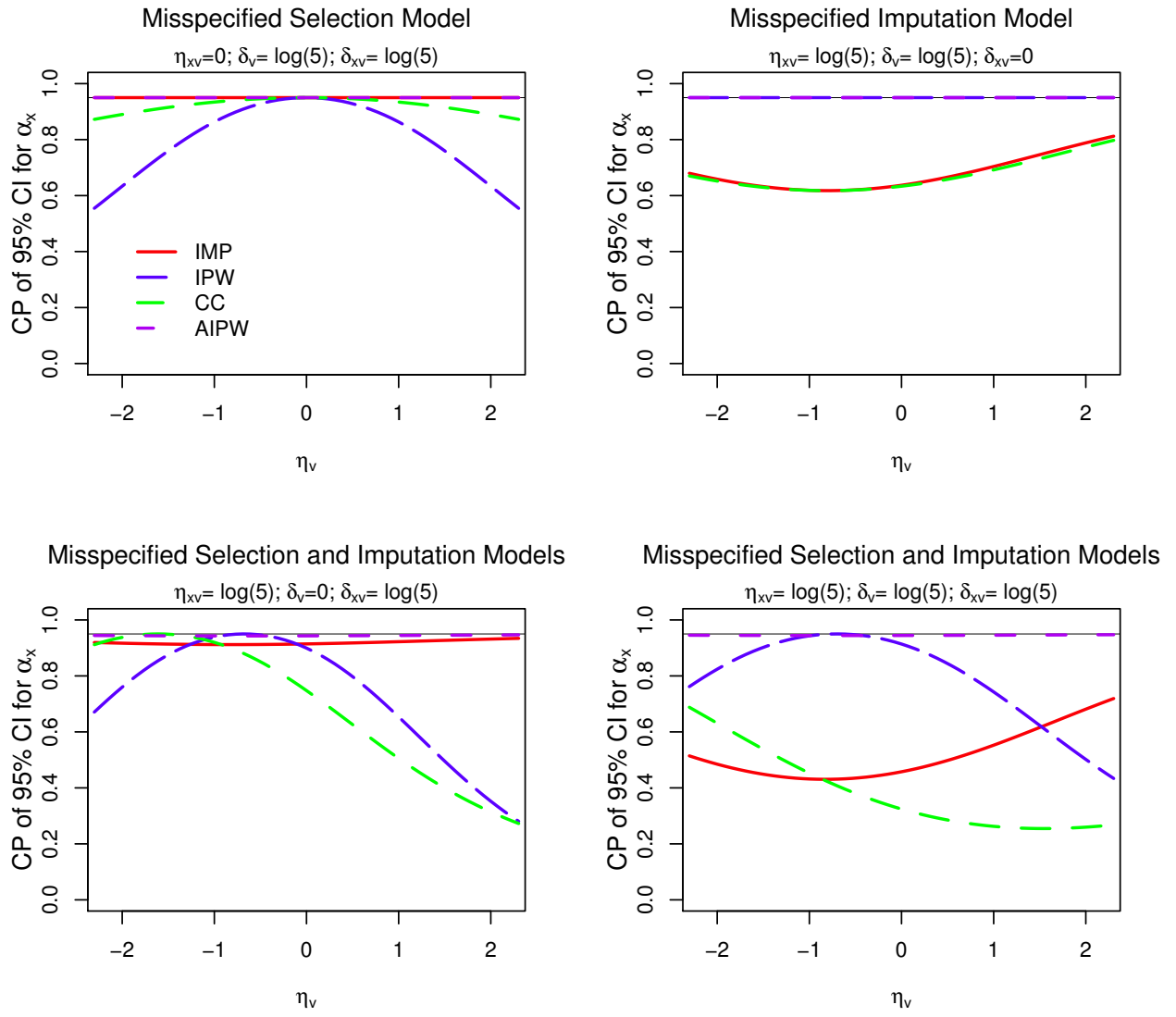

Figure 2: Coverage: probability that nominal 95% confidence intervals of estimators of α_x from samples of $N = 1000$ individuals contain the true value; the four panels are based on calculations from large-sample distributions and allow for a range of values of η_v, δ_v, η_{xv} and δ_{xv}, while taking $\eta_x = \delta_x = 0$, $E[Y] = 0.5$, $E[X] = 0.5$, $E[V] = 0.5$, and $E[R] = 0.5$.

to note that the coverage probability for the IPW and MI estimators are at times worse than those from the CC estimator. We have considered coverage probabilities of nominal 95% confidence intervals based on samples with $N = 1000$ individuals. Larger sample sizes will result in narrower confidence intervals, but since these intervals will be centered on the estimators incorrectlimiting value, the coverage of corresponding confidence intervals will be lower with larger sample sizes.

The general trend seen in Figures 1 and 2 was not unique to the case of independence between X and V; Figure 3 shows that, when both auxiliary models are misspecified, the AIPW analysis generally resulted in smaller asymptotic bias and better coverage probabilities than the CC, IPW and MI methods regardless of the odds ratio $OR(X, V) = P(X = 1|V = 1)P(X = 0|V = 0)/\{P(X = 1|V = 0)P(X = 0|V = 1)\}$ characterizing the association between X and V. Figure 3 Figure 3 is presented for the case that $\eta_v = 1$, but similar results were seen for other choices (not presented).

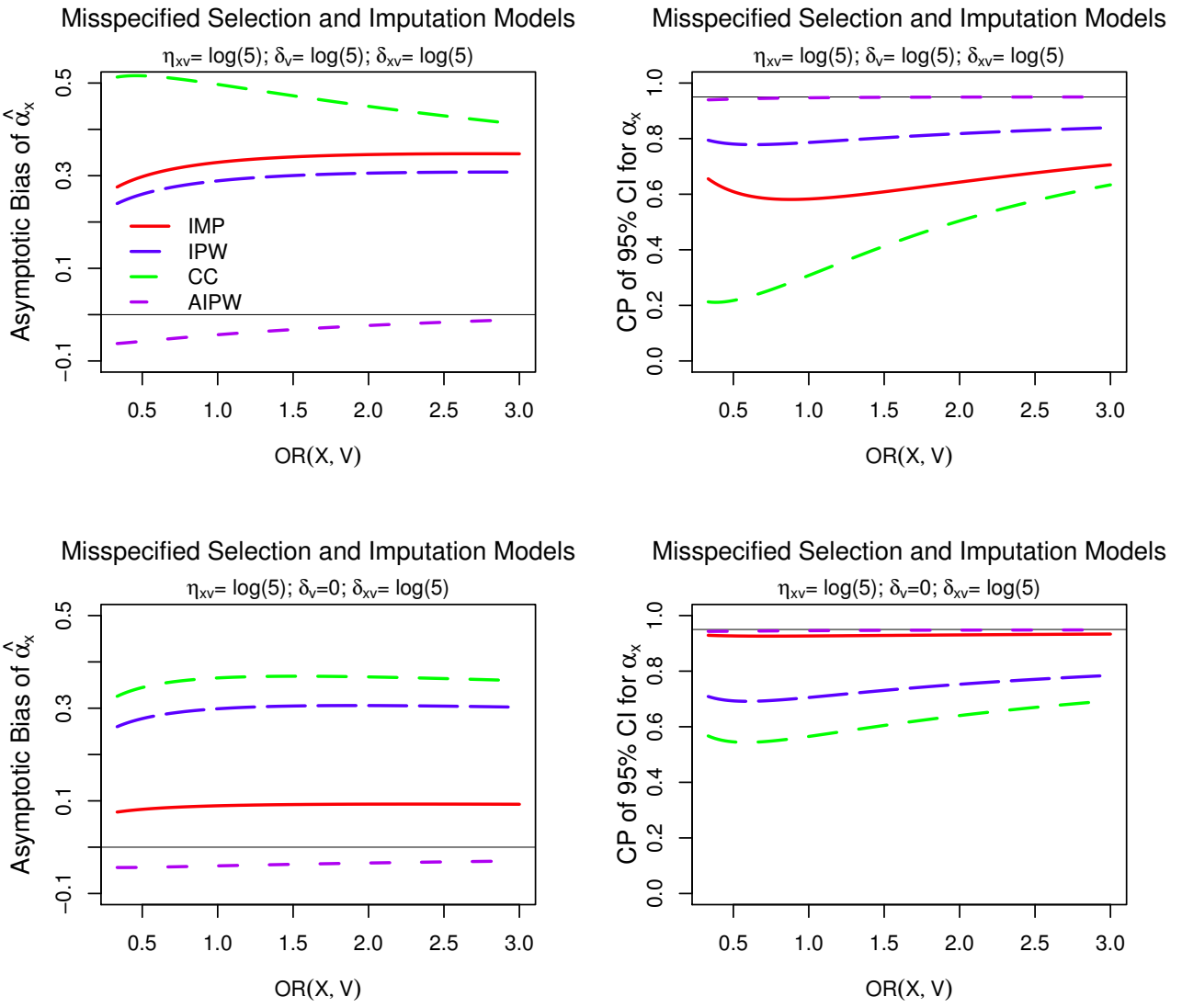

Figure 3: Asymptotic bias (left panels) and coverage probability of nominal 95% confidence intervals based on $N = 1000$ individuals (right panels) of estimators of α_x when X and V are not necessarily independent; a range of odds ratios relating X and V is considered. Both auxiliary models are misspecified by ignoring the non-zero parameters $\eta_{xv} = \log(5)$ and $\delta_{xv} = \log(5)$; we consider both $\delta_v = \log(5)$ (top panels) and $\delta_v = 0$ (bottom panels) and set $\eta_v = 1$, $\eta_x = \delta_x = 0$, $E[Y] = 0.5$, $E[X] = 0.5$, $E[V] = 0.5$, and $E[R] = 0.5$.

4 Empirical Behaviour of Estimators

Here we illustrate the use of the various methods and examine the empirical properties through simulation studies based on an application to a recent trial of breast cancer patients with skeletal metastases [8]. The response Y is an indicator of a skeletal complication (event) during the first year following randomization to either monthly IV infusions of a bisphosphonate therapy pamidronate ($X = 1$) or a placebo control ($X = 0$). We consider the auxiliary variable V as indicating if the patient had an elevated pain score at study entry, a marker of the extent of skeletal metastases. In this trial, complete information on these binary variables was available for 214 individuals, and analyses of these data were used to determine the parameters for use in our simulations.

In the simulation study, data for $N = 1000$ individuals were generated according to model (6) with the parameters chosen to be consistent with data from the breast cancer trial. Indicators of missingness were generated according to the selection model (7) with the selection parameters taken

to be $\delta_{\text{sim1}} = [\delta_{11}, 0, \log 5, \log 5]'$ or $\delta_{\text{sim2}} = [\delta_{12}, 0, 0, \log 5]'$, where the intercept terms were chosen so that $E[R] = 0.50$, representing the situation where responses were unavailable for 50% of the patients. These simulated data were then analysed by fitting misspecified models involving η^* and δ^* which ignored the interaction term between X and V as described in the previous section. This process of simulating missingness and analysing the resulting dataset was repeated 2000 times and the empirical properties of the estimators of α_x were recorded so they could be compared with the asymptotic results.

A similar simulation was conducted for the situation in which an additional covariate, V_2, was associated with both Y and R; misspecification of the auxiliary models here comes from omission of V_2 rather than omission of an interaction term. The response was generated according to the models $E[Y|X, V, V_2; \gamma] = \text{expit}(\gamma_1 + \gamma_x X + \gamma_v V + \gamma_{v2} V_2)$ and $P(R = 1|Y, X, V, V_2; \delta) = \text{expit}(\delta_1 + \delta_x X + \delta_v V + \delta_{v2} V_2)$ where the η is derived from corresponding analysis of the breast cancer trial data with V_2 taken to be an indicator of advanced age at disease onset; we again consider the selection models δ_{sim1} and δ_{sim2}.

Tables 1 and 2 present the asymptotic bias of the estimators under model misspecication (i.e. we present the difference between the limiting value of the estimator, α_x^*, and the true value, α_x) as well as the observed bias of the estimators (the difference between the mean estimate, $\overline{\widehat{\alpha}}_x$, and the true value). We also report the large-sample standard errors averaged over all simulated datasets (SE) and the empirical standard error (ESE), defined as the square root of the sample variance of the point estimates over all simulations. The coverage probabilities (CP) of nominal 95% confidence intervals are reported based on large-sample theory (as in Section 3), as well as the empirical coverage probability (ECP), defined as the proportion of simulations in which nominal 95% confidence intervals contained the truth. We note that these confidence intervals would have the nominal 95% coverage if all auxiliary models were correctly specified; more generally, however, for a given sample size, the coverage of these intervals will decrease as the asymptotic bias increases.

Table 1 contains results from the simulation setting where the auxiliary models are misspecified due to omission of the interaction between X and V. Table 2 contains similar results for the simulation setting where the auxiliary models are misspecified due to omission of the covariate V_2 in the main-effect models.

Table 1: The asymptotic and empirical properties of estimators of α_x when using misspecified auxiliary models that omit the interaction between X and V.

	$(\alpha_x^* - \alpha_x)$	$(\overline{\widehat{\alpha}}_x - \alpha_x)$	SE	ESE	CP	ECP
			δ_{sim1}			
CC	-0.122	-0.132	0.198	0.197	90.5	90.6
MI	-0.209	-0.219	0.192	0.191	80.4	80.7
IPW	0.186	0.177	0.217	0.218	86.2	85.7
AIPW	0.024	0.014	0.215	0.216	94.9	95.2
			δ_{sim2}			
CC	0.135	0.134	0.190	0.189	89.0	89.5
MI	-0.048	-0.048	0.186	0.184	94.2	94.7
IPW	0.230	0.231	0.192	0.191	77.5	76.7
AIPW	0.029	0.030	0.189	0.187	94.8	94.6

Note: Here $\eta = [0.071, -0.439, 1.898, -1.007]'$, $E[X] = 0.509$, $E[V] = 0.430$, and $OR(X, V) = 1$.

Table 2: The asymptotic and empirical properties of estimators of α_x when using misspecified auxiliary models that omit the covariate V_2.

	$(\alpha_x^* - \alpha_x)$	$(\overline{\widehat{\alpha}}_x - \alpha_x)$	SE	ESE	CP	ECP
			δ_{sim1}			
CC	0.016	0.014	0.188	0.193	94.9	94.4
MI	-0.029	-0.031	0.184	0.190	94.7	94.1
IPW	0.050	0.049	0.194	0.198	94.2	94.0
AIPW	-0.003	-0.005	0.190	0.194	95.0	94.3
			δ_{sim2}			
CC	0.061	0.056	0.184	0.186	93.7	94.0
MI	0.002	-0.004	0.181	0.183	95.0	95.0
IPW	0.058	0.053	0.184	0.187	93.8	94.2
AIPW	-0.001	-0.007	0.180	0.182	95.0	94.6

Note: Here $\gamma = [0.696, -0.767, 1.303, -1.007]'$, $E[X|V, V_2] = \text{expit}(0.247 - 0.768V - 0.154V_2 + 1.021VV_2)$, $E[V|V_2] = \text{expit}(-0.111 - 0.350V_2)$, $E[V_2] = 0.495$.

The empirical results tracked the asymptotic calculations very closely in terms of bias, variance and coverage probability. Here again we see that a complete-case analysis can lead to substantial bias and poor coverage when data are not missing at random (Table 1). Furthermore, it can also be seen that use of inverse probability weighted estimating equations and multiple imputation can lead to even greater bias and worse coverage than the CC estimator when the weighting and imputation models are misspecfied (Tables 1 and 2). However, in all cases the augmented inverse probability weighted estimating equations, which exploit both of these misspecified models, led to an estimator with small bias and good coverage. As would be expected, additional simulations (not presented) showed that the magnitude of the biases decreased with decreasing levels of missingness (i.e. all estimators were closer to the truth for lower values of $E[R]$). However, the relative sizes of biases among these estimators were similar regardless of the amount of missing data.

5 General Remarks

Commonly used, naïve methods for analysing incomplete data (e.g. complete-case analysis, non-responder imputation, last-observation carried forward imputation) yield consistent estimators only in very special circumstances. Weighted estimating equations and model-based multiple imputation approaches can be more generally appropriate, but they can require further explicit modellingassumptions. Greater understanding of the underlying causes of missingness in a given study will provide analysts with a greater hope of making correct modelling decisions. However, it is very difficult in practice to ensure that modelling assumptions made to account for missing data are correct. We have demonstrated that using misspecified models to adjust for response-biased observed data can result in increased rather than decreased bias, and this problem is especially problematic when the rate of missingness is large. The double robustness property of the augmented inverse probability weighted estimator makes this method appealing since it allows for two chances to get the model right. Bang and Robins [1] suggested that this double robustness property will be advantageous even when both models are slightly misspecified; however, Kang and Schafer [9] demonstrated empirically that this is not always the case. In the settings we explored, the double robustness property of the augmented inverse probability weighted estimating equations was evident and we found that this estimator had

relatively small asymptotic and empirical biases when both models were incorrectly specied, despite the fact that use of the misspecfied selection model in inverse-weighted estimating equations or the misspecied imputation model in multiple imputation could actually result in larger asymptotic biases than use of a naïve complete-case analysis. As discussed in Kang and Schafer [9] and in the published comments to that paper [22], there are situations in which estimators arising from augmented inverse probability weighted estimating equations can have poor empirical properties and other approaches may be preferable. In particular, augmented inverse probablity weighted estimators may have poor empirical properties when the weights are highly variable. We did not observe such a problem in the settings considered in our simulation studies, although the double-robust approach may have benefited from the stability of the weights arising from the categorical nature of our data. If variability in the weights is a concern, stabilized weights [19] or an enhanced propensity score model [2] may be implemented to improve performance of the double-robust estimator.

In this paper and its appendices, we derived explicit forms for the limiting values of certain estimators and have shown that there is no universally least-biased approach to handling incomplete data when necessary model assumptions are wrong. Therefore, it is important to consider carefully the models that are specified to accommodate missingness, to ensure that these modelling decisions are tenable, and to carry out sensitivity analyses exploring the robustness of conclusions to changes in the missing data model [14, 15, 16, 22].

A Derivations on the Asymptotic Behaviour for Estimators

A.1 Inverse Probability Weighting

If $\theta = (\alpha', \delta')$, the inverse probability weighted estimator is found by solving the estimating equation

$$0 = \sum_{i=1}^{N} T_i(\theta) = \sum_{i=1}^{N} \begin{pmatrix} U_i^{\text{ipw}}(\alpha, \delta) \\ S_i(\delta) \end{pmatrix}$$

for $\widehat{\theta} = (\widehat{\alpha}', \widehat{\delta}')'$. Under mild regularity conditions

$$N^{1/2}\left(\widehat{\theta}_{\text{ipw}} - \theta_{\text{ipw}}\right) \xrightarrow{D} MVN(0, E[-\partial T_i/\partial\theta']^{-1} E[T_i T_i'][E[-\partial T_i/\partial\theta']^{-1}]'),$$

where these expressions are evaluated at θ_{ipw} [2, 10, 21, 28].

Now, suppressing the dependence on θ_{ipw}

$$\begin{aligned} E[-\partial T_i/\partial\theta']^{-1} &= -E\begin{bmatrix} \partial U_i^{\text{ipw}}/\partial\alpha' & \partial U_i^{\text{ipw}}/\partial\delta' \\ \partial S_i/\partial\alpha' & \partial S_i/\partial\delta' \end{bmatrix}^{-1} \\ &= -\begin{bmatrix} -E[\partial U_i^{\text{ipw}}/\partial\alpha']^{-1} & E[\partial U_i^{\text{ipw}}/\partial\alpha']^{-1} E[\partial U_i^{\text{ipw}}/\partial\delta'] E[\partial S_i/\partial\delta']^{-1} \\ 0 & -E[\partial S_i/\partial\delta']^{-1} \end{bmatrix} \end{aligned}$$

and

$$E[T_i T_i'] = E\begin{bmatrix} U_i^{\text{ipw}} U_i^{\text{ipw}\prime} & U_i^{\text{ipw}} S_i' \\ S_i U_i^{\text{ipw}\prime} & S_i S_i' \end{bmatrix},$$

and the desired asymptotic variance for $\widehat{\alpha}_{\text{ipw}}$ can be found by extracting the $p \times p$ upper left sub-matrix of the asymptotic variance of $\widehat{\theta}$.

Let $f(Z)$ represent the true density function for the data $Z = (Y, X, V, R)$. If the estimating function for δ is correctly specified as the score function for the true log-likelihood so that $S_i = \partial \log f(Z)/\partial\delta$, then

$$0 = E[U_i^{\text{ipw}}(\theta_{\text{ipw}})] = \int U_i^{\text{ipw}}(\theta_{\text{ipw}}) f(Z) dZ.$$

So,

$$\begin{aligned}
0 &= \int \left\{ \partial [U_i^{\text{ipw}}(\theta_{\text{ipw}}) f(Z)]/\partial\delta' \right\} dZ \\
&= \int \left\{ \partial [U_i^{\text{ipw}}(\theta_{\text{ipw}})]/\partial\delta' \cdot f(Z) + U_i^{\text{ipw}}(\theta_{\text{ipw}}) \cdot \partial [f(Z)]/\partial\delta' \right\} dZ \\
&= \int \left\{ \partial U_i^{\text{ipw}}(\theta_{\text{ipw}})/\partial\delta' \cdot f(Z) \right\} dZ + \int \left\{ U_i^{\text{ipw}}(\theta_{\text{ipw}}) \cdot \partial [\log f(Z)]/\partial\delta' \cdot f(Z) \right\} dZ \\
&= E[\partial U_i^{\text{ipw}}(\theta_{\text{ipw}})/\partial\delta'] + E[U_i^{\text{ipw}}(\theta_{\text{ipw}}) S_i'(\theta_{\text{ipw}})], \qquad \text{(A.1)}
\end{aligned}$$

which establishes the generalized information equality $E[\partial U_i^{\text{ipw}}(\theta_{\text{ipw}})/\partial\delta'] = -E[U_i^{\text{ipw}}(\theta_{\text{ipw}}) S_i'(\theta_{\text{ipw}})]$.

A.2 Augmented Inverse Probability Weighting

The augmented inverse probability weighted estimator is found by solving the estimating equation for $\widehat{\Omega} = (\widehat{\alpha}', \widehat{\eta}', \widehat{\delta}')'$

$$0 = \sum_{i=1}^{N} T_i(\Omega) = \sum_{i=1}^{N} \begin{pmatrix} U_i^{\text{aipw}}(\alpha, \delta, \eta) \\ W_i(\eta) \\ S_i(\delta) \end{pmatrix},$$

and under mild regularity conditions and suppressing dependence on the parameter Ω_{aipw} as before

$$N^{1/2} \left(\widehat{\Omega}_{\text{aipw}} - \Omega_{\text{aipw}} \right) \xrightarrow{D} MVN(0, E[-\partial T_i/\partial\Omega']^{-1} E[T_i T_i'][E[-\partial T_i/\partial\Omega']^{-1}]'),$$

evaluated at Ω_{aipw} [10, 21, 28]. Here, $E[-\partial T_i/\partial\Omega']^{-1}$ is

$$\begin{aligned}
&- E \begin{bmatrix} \partial U_i^{\text{aipw}}/\partial\alpha' & \partial U_i^{\text{aipw}}/\partial\eta' & \partial U_i^{\text{aipw}}/\partial\delta' \\ 0 & \partial W_i/\partial\eta' & 0 \\ 0 & 0 & \partial S_i/\partial\delta' \end{bmatrix}^{-1} \\
&= - \begin{bmatrix} -E[\partial U_i^{\text{aipw}}/\partial\alpha']^{-1} & E[\partial U_i^{\text{aipw}}/\partial\alpha']^{-1}\kappa_{\text{mi}}^* & E[\partial U_i^{\text{aipw}}/\partial\alpha']^{-1}\kappa_{\text{ipw}}^* \\ 0 & & -E[\partial S_i/\partial\delta']^{-1} \end{bmatrix}
\end{aligned}$$

and

$$E[T_i T_i'] = E \begin{bmatrix} U_i^{\text{aipw}} U_i^{\text{aipw}\prime} & U_i^{\text{aipw}} W_i' & U_i^{\text{aipw}} S_i' \\ W_i U_i^{\text{aipw}\prime} & W_i W_i' & W_i S_i' \\ S_i U_i^{\text{aipw}\prime} & S_i W_i' & S_i S_i' \end{bmatrix},$$

where $\kappa_{\text{ipw}}^* = E[\partial U_i^{\text{aipw}}/\partial\delta'] E[\partial S_i/\partial\delta']^{-1}$ and $\kappa_{\text{mi}}^* = E[\partial U_i^{\text{aipw}}/\partial\eta'] E[\partial W_i/\partial\eta']^{-1}$; the desired asymptotic variance for $\widehat{\alpha}_{\text{aipw}}$ can be found by extracting the $p \times p$ upper left sub-matrix of the asymptotic variance of $\widehat{\Omega}$ and by noting that $E[\partial U_i^{\text{aipw}}/\partial\alpha'] = E[\partial U_i(\alpha)/\partial\alpha']$ since $E[\partial U_i(\alpha, \eta)/\partial\alpha'] = E[\partial U_i(\alpha)/\partial\alpha']$.

By following similar arguments to those used to establish (A.1) (i.e. by replacing $U_i^{\text{ipw}}(\theta_{\text{ipw}})$ with $U_i^{\text{aipw}}(\theta_{\text{aipw}})$ or $S_i(\theta_{\text{aipw}})$), it can be shown that if $S_i = \partial \log f(Z)/\partial\delta$, then

$$E[\partial U_i^{\text{aipw}}/\partial\delta'] = -E[U_i^{\text{aipw}} S_i'] \qquad \text{and} \qquad E[\partial S_i/\partial\delta'] = E[S_i S_i'].$$

B EXPLICIT FORMS OF LIMITING VALUES OF ESTIMATORS

Here, we consider the explicit forms of the limiting values of estimators, and therefore the asymptotic biases, arising in the problem introduced in Section 3. We note again that, as in (8), under correct model specification, the conditional mean model will satisfy

$$\mu(X;\alpha_0) = E_V\{E[Y|X,V;\eta_0]\}$$

and we derive here the limiting value of the conditional mean model under the model misspecification described in Section 3.

B.1 COMPLETE-CASE ANALYSIS

Here, $\widehat{\alpha}_{\rm cc}$, the root of the estimating equation $\sum_{i=1}^{N} R_i\big[Y_i - \mu(X_i;\alpha)\big]\,[1,\ X_i]'$, will not consistently estimate α_0 since $(Y \not\perp R)|X$. In fact, the limiting value of the estimator of the conditional mean is

$$\mu(X;\alpha_{\rm cc}) = E[Y|X,R=1;\eta_0,\delta_0] = \frac{E_V\{E[Y|X,V;\eta_0]P(R=1|X,V;\delta_0)\}}{E_V\{P(R=1|X,V;\delta_0)\}}, \qquad \text{(B.1)}$$

since $X \perp V$ and $Y \perp R|X,V$. Note that this estimator is asymptotically unbiased if $Y \perp R|X$, which occurs if $Y \perp V|X$ or $R \perp V|X$ (that is, if here $\eta_v = \eta_{xv} = 0$ or $\delta_v = \delta_{xv} = 0$), as then (B.1) reduces to (8).

B.2 ESTIMATORS FROM MULTIPLE IMPUTATION UNDER MISSPECIFICATION

We are supposing that the imputation model is misspecified as in (10). Such an estimator consistently estimates the $\eta_{\rm mi}$ which solves

$$\begin{aligned} 0 &= E\Big\{ R_i\big[Y_i - m(X_i,V_i;\eta_{\rm mi})\big]\,[1,\ X_i,\ V_i]' \Big\} \\ &= E_{XV}\Big\{ P(R_i=1|X_i,V_i;\delta_0)\big[E[Y_i|X_i,V_i;\eta_0] - m(X_i,V_i;\eta_{\rm mi})\big]\,[1,\ X_i,\ V_i]' \Big\}. \end{aligned}$$

Thus, the limiting imputation estimator $\alpha_{\rm mi}$ solves

$$\begin{aligned} 0 =& E\{U_i^{\rm mi}(\alpha,\eta_{\rm mi})\} \\ =& E\big\{ \big[R_i\big[Y_i - \mu(X_i;\alpha)\big] + (1-R_i)\big[Y_{ij}^{imp}(\eta_{\rm mi}) - \mu(X_i;\alpha)\big]\big]\,[1,\ X_i]' \big\} \\ =& E\big\{ \big[\big[Y_{ij}^{imp}(\eta_{\rm mi}) - \mu(X_i;\alpha)\big] + R_i\big[Y_i - Y_{ij}^{imp}(\eta_{\rm mi})\big]\big]\,[1,\ X_i]' \big\} \\ =& E_{XV}\big\{ \big[m(X_i,V_i;\eta_{\rm mi}) + P(R_i=1|X_i,V_i;\delta_0)\big[E[Y_i|X_i,V_i;\eta_0] - m(X_i,V_i;\eta_{\rm mi})\big] - \mu(X_i;\alpha)\big]\,[1,\ X_i]' \big\} \\ =& E_{XV}\big\{ \big[m(X_i,V_i;\eta_{\rm mi}) - \mu(X_i;\alpha)\big]\,[1,\ X_i]' \big\} \end{aligned}$$

and so $\alpha_{\rm mi}$ satisfies

$$\mu(X;\alpha_{\rm mi}) = E_V\{m(X,V;\eta_{\rm mi})\}. \qquad \text{(B.2)}$$

Note that an expected conditional mean (ECM) or outcome-regression estimator estimates $E[Y|X;\alpha]$ through estimation of $E_V\{m(X,V;\eta)\}$ [4], so the misspecified imputation approach described here has the same asymptotic bias as a misspecified ECM estimator. If the imputation model was correctly specified so that $m(X,V;\eta_{\rm mi}) = E[Y|X,V;\eta_0]$ (i.e. if here $\eta_{xv} = 0$), then (B.2) reduces to (8).

B.3 Misspecified Inverse Probability Weighted Estimating Equations

Under this misspecification of the selection model, a root of $E[U_i^{\text{ipw}}(\alpha, \delta_{\text{ipw}})]$ is α_{ipw}, such that

$$\mu(X; \alpha_{\text{ipw}}) = \frac{E_V\big\{E[Y|X,V;\eta_0]P(R=1|X,V;\delta_0)/\pi(X,V;\delta_{\text{ipw}})\big\}}{E_V\big\{P(R=1|X,V;\delta_0)/\pi(X,V;\delta_{\text{ipw}})\big\}}, \tag{B.3}$$

since

$$\begin{aligned}
0 &= E_{RYXV}\left\{\frac{R_i}{\pi(X_i,V_i;\delta_{\text{ipw}})}[Y_i - \mu(X_i;\alpha)]\,[1,\ X_i]'\right\} \\
&= E_{YXV}\left\{\frac{P(R_i=1|X_i,V_i;\delta_0)}{\pi(X_i,V_i;\delta_{\text{ipw}})}[Y_i - \mu(X_i;\alpha)]\,[1,\ X_i]'\right\}, \qquad \text{as } (R\perp Y)|X,V \\
&= E_X\left\{\left[E_V\left\{\frac{P(R_i=1|X_i,V_i;\delta_0)}{\pi(X_i,V_i;\delta_{\text{ipw}})}E[Y_i|X_i,V_i;\eta_0]\right\} - E_V\left\{\frac{P(R_i=1|X_i,V_i;\delta_0)}{\pi(X_i,V_i;\delta_{\text{ipw}})}\right\}\mu(X_i;\alpha)\right][1,\ X_i]'\right\}.
\end{aligned}$$

This conditional mean estimator is asymptotically unbiased if the missingness model is correctly specified so $\pi(X,V) = P(R=1|X,V;\delta)$ (i.e. if $\delta_{xv}=0$, so $\delta_{\text{ipw}}=\delta$), since then (B.3) reduces to (8). Note that δ_{ipw}, the value that is being consistently estimated by solving (9), can be found by solving $0 = E\big\{[R_i - \pi(X_i,V_i;\delta)][1,X_i,V_i]'\big\}$. With binary data, this expectation can be easily calculated as a sum over eight distinct types of individual (correpsonding to the 2^3 possibilities of R,X,V), each weighted by the corresponding probability $P(R,X,V)$.

B.4 Augmented Inverse Probability Weighting under Misspecification

The AIPW estimator is consistently estimating α_{aipw} which satisfies

$$\begin{aligned}
0 =& E\{U_i^{\text{aipw}}(\alpha,\delta_{\text{ipw}},\eta_{\text{ipw}})\} \\
=& E\{U_i(\alpha,\eta) + R_i\pi(X_i,V_i;\delta)^{-1}[U_i(\alpha) - U_i(\alpha,\eta)]\} \\
=& E_{RYXV}\big\{\big[[m(X_i,V_i;\eta_{\text{mi}}) - \mu(X_i;\alpha)] + R_i\pi(X_i,V_i;\delta_{\text{ipw}})^{-1}[Y_i - m(X_i,V_i;\eta_{\text{mi}})]\big]\,[1,\ X_i]'\big\} \\
=& E_{XV}\big\{\big[m(X_i,V_i;\eta_{\text{mi}}) + P(R_i=1|X_i,V_i;\delta_0)\pi(X_i,V_i;\delta_{\text{ipw}})^{-1}[E[Y_i|X_i,V_i;\eta_0] - m(X_i,V_i;\eta_{\text{mi}})] \\
&\qquad - \mu(X_i;\alpha)\big]\,[1,\ X_i]'\big\}
\end{aligned}$$

so

$$\mu(X;\alpha_{\text{aipw}}) = E_V\left\{\frac{P(R=1|X,V;\delta_0)}{\pi(X,V;\delta_{\text{ipw}})}[E[Y|X,V;\eta_0] - m(X,V;\eta_{\text{mi}})] + m(X,V;\eta_{\text{mi}})\right\}. \tag{B.4}$$

Note that if either the selection or imputation model was correctly specified, then (B.4) would reduce to (8) and the AIPW estimator would allow for consistent estimation of the true conditional mean response.

Acknowledgements

Richard Cook is a Tier I Canada Research Chair in Statistical Methods for Health Research. The authors thank Novartis for permission to use data from the breast cancer trial for illustration.

FUNDING

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by an Alexander Graham Bell Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council (NSERC) of Canada to Michael McIsaac and grants from NSERC (RGPIN 155849) and the Canadian Institutes of Health Research (FRN 13887) to Richard Cook, Canada Research Chair (Tier 1) - CIHR funded (950-226626)

REFERENCES

[1] Bang, H. and Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972.

[2] Cao, W., Tsiatis, A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734.

[3] Carpenter, J., Kenward, M., and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical, Society A*, 169(3):571–584.

[4] Chen, B. and Cook, R. (2012). Strategies for bias reduction in estimation of marginal means with data missing at random. In Pardalos, P. and Coleman, T.F. amd Xanthopoulos, P., editors, *Optimization and Data Analysis on Biomedical Informatics*, pages 99–115. American Mathematics Society.

[5] Chen, B., Yi, G., and Cook, R. (2010). Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Associaton*, 105:336–353.

[6] Clayton, D., Spielgelhalter, D., Dunn, G., and Pickles, A. (1998). Analysis of longitudinal binary data from multi-phase sampling. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 60(1):71–87.

[7] Cook, R., Zeng, L., and Yi, G. (2004). Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation. *Biometrics*, 60:820–828.

[8] Hortobagyi, G., Theriault, R., Porter, L., Blayney, D., Lipton, A., Sinoff, C., Wheeler, H., Simeone, J., Seaman, J., Knight, R., et al. (1996). Efficacy of pamidronate in reducing skeletal complications in patients with breast cancer and lytic bone metastases. *New England Journal of Medicine*, 335(24):1785–1792.

[9] Kang, J. and Schafer, J. (2007). Demystifying double robustness. *Statistical Science*, 22(4):523–539.

[10] Lawless, J., Kalbfleisch, J., and Wild, C. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 61(2):413–438.

[11] Little, R., D'Agostino, R., Cohen, M., Dickersin, K., Emerson, S., Farrar, J., Frangakis, C., and et al. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367:1355–1360.

[12] Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data, Second Edition*. John Wiley & Sons, New York.

[13] McIsaac, M. and Cook, R. (2014a). Response-dependent two-phase sampling designs for biomarker studies. *Canadian Journal of Statistics*, 42(2):268–284.

[14] McIsaac, M. and Cook, R. (2014b). Statistical models and methods for incomplete data in randomized clinical trials. In van Montfort, K., Oud, J., and Ghidey, W., editors, *Developments In Statistical Evaluation Of Clinical Trials*, pages 1–27. Springer.

[15] McIsaac, M., Cook, R., and Poulin-Costello, M. (2013). Incomplete data in randomized dermatology trials: Consequences and statistical methodology. *Dermatology*, 226(1):19–27.

[16] Pierce, D. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *The Annals of Statistics*, 10:475–478.

[17] Reilly, M. and Pepe, M. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82(2):299–314.

[18] Reilly, M. and Pepe, M. (1997). The relationship between hot-deck multiple imputation and weighted likelihood. *Statistics in Medicine*, 16:5–19.

[19] Robins, J., Hernan, M., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidermiology*, 11(5):550–560.

[20] Robins, J. and Rotnitzky, A. (2001). Comment on "inference for semiparametric models: some questions and an answer," by P. J. Bickel and J. Kwon. *Statistica Sinica*, 11:920–936.

[21] Robins, J., Rotnitzky, A., and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

[22] Robins, J. and Wang, N. (2000). Inference for imputation estimators. *Biometrika*, 87:113–124.

[23] Rotnitzky, A., Robins, J., and Scharfstein, D. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444):1321–1339.

[24] Rotnitzky, A. and Wypij, D. (1994). A note on the bias of estimators with missing data. *Biometrics*, 50(4):1163–1170.

[25] Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

[26] Schafer, J. (1999). Multiple Imputation: A Primer. *Statistical Methods in Medical Research*, 8:3–15.

[27] Seaman, S., White, I., Copas, A., and Li, L. (1999). Combining multiple imputation and inverse-probability weighting. *Biometrics*, 68:129–137.

[28] Stefanski, L. and Boos, D. (2002). The calculus of m-estimation. *The American Statistician*, 56(1):29–38.

[29] Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. Springer Science + Business Media, New York.

[30] Wang, N. and Robins, J. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85:935–948.

[31] Xie, F. and Paik, M. (1997). Multiple imputation methods for the missing covariates in generalized estimating equation. *Biometrics*, 53:1538–1546.

[32] Yu, M. and Nan, B. (2006). A revisit of semiparametric regression models with missing data. *Statistica Sinica*, 16(4):1193.