Received: 11 January 2020 | Revised: 13 February 2021 | Accepted: 24 March 2021

DOI: 10.1002/sim.8978

RESEARCH ARTICLE

Statistics in Medicine WILEY

Evaluation of predictive model performance of an existing model in the presence of missing data

Pin Li[1] | Jeremy M. G. Taylor[1,2] | Daniel E. Spratt[2] | R. Jeffery Karnes[3] | Matthew J. Schipper[1,2]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA

[2]Department of Radiation Oncology, University of Michigan, Ann Arbor, Michigan, USA

[3]Department of Urology, Mayo Clinic, Rochester, Minnesota, USA

Correspondence
Pin Li, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.
Email: pinli@umich.edu

Funding information
U.S. National Institutes of Health, Grant/Award Numbers: CA059827, CA129102

In medical research, the Brier score (BS) and the area under the receiver operating characteristic (ROC) curves (AUC) are two common metrics used to evaluate prediction models of a binary outcome, such as using biomarkers to predict the risk of developing a disease in the future. The assessment of an existing prediction models using data with missing covariate values is challenging. In this article, we propose inverse probability weighted (IPW) and augmented inverse probability weighted (AIPW) estimates of AUC and BS to handle the missing data. An alternative approach uses multiple imputation (MI), which requires a model for the distribution of the missing variable. We evaluated the performance of IPW and AIPW in comparison with MI in simulation studies under missing completely at random, missing at random, and missing not at random scenarios. When there are missing observations in the data, MI and IPW can be used to obtain unbiased estimates of BS and AUC if the imputation model for the missing variable or the model for the missingness is correctly specified. MI is more efficient than IPW. Our simulation results suggest that AIPW can be more efficient than IPW, and also achieves double robustness from miss-specification of either the missingness model or the imputation model. The outcome variable should be included in the model for the missing variable under all scenarios, while it only needs to be included in missingness model if the missingness depends on the outcome. We illustrate these methods using an example from prostate cancer.

KEYWORDS

area under the ROC curve, augmented inverse probability weighting, Brier score, inverse probability weighting, multiple imputation

1 | INTRODUCTION

In clinical research, patient information such as clinical features, diagnostic tests, and biomarkers are often used to help with diagnosis or to provide prognosis of a future outcome for a patient with disease. When the outcome of interest is binary, a typical prediction model will numerically combine the covariates, for example, using a linear combination, to estimate the predicted probability of the binary outcome. The evaluation of an existing prediction model in a different populations is of considerable interest. If a model is to be transportable to other populations, it needs to be validated, which is usually thought of as meaning that it has similar and good performance in other populations. The performance of existing prediction models can be assessed using a variety of metrics, such as the Brier score (BS) to indicate accuracy of the probabilistic predictions, and area under the receiver operating characteristic (ROC) curve (or the concordance statistic) for discriminative ability.[1] Very often, a covariate may be partially missing, that is, the values will be missing for some patients. The assessment of prediction models in data with missing covariate values is a challenge. The context we are considering is that the existing model or models were developed on other datasets, which we call the external data, and are already completely specified. We do not have access to the data used to develop these models. Rather, our goal is to assess the performance of the existing model in an available dataset, which we call the internal data, that has missing values for some covariates and we want to get valid and efficient estimates of the BS and the area under the curve (AUC).

In general there are two types of methods for estimation in the presence of missing data, one is based on multiple imputation (MI) and the other is based on inverse probability weighting (IPW). For MI, a model for the distribution of the missing variable, or variables, needs to be specified. For IPW method, a model for the probability of missingness needs to be specified, which is also called the weight model. For MI, M completed datasets are created and M model performance measures can be estimated from each of the completed dataset and then averaged.[2] Alternatively, an overall measure of model performance can be estimated directly from a simple completed dataset that includes the average of the M predictions for each missing value. As previously recommended,[3] the former is preferred. The analysis of only the observations with complete data is frequently biased, and in a cleverly titled article Janssen et al[4] showed that to impute is generally better than to ignore. Alternatively IPW is a commonly used approach to correct their bias.[5] IPW is also used to adjust for unequal sampling fractions in sample surveys and causal inference.[6,7] Augmented inverse probability weighting (AIPW) has been proposed as an extension of IPW. It is a double-robust method that is robust to the misspecification of either a model for the missingness mechanism or a model for the distribution of the variables with missing values (but not both).[8] Willamson et al[7] present AIPW estimators that account for both confounding in causal inference and missing data. AIPW generally results in improved efficiency compared with IPW, although this is not guaranteed to be the case.

When analyzing data with missing values an important consideration is the missingness mechanism, and the mechanism will impact the properties and merits of different methods. Missing complete at random (MCAR) is when the probability of any variable being missing for a subject does not depend on the value of any of the variables. Generally all methods will work under MCAR. Analysis of the complete cases (CCs) will be unbiased, but are frequently quite inefficient compared with other methods, depending on the amount of missingness. Missing at random (MAR) is when the probability of being missing can depend on other covariates, but only those that are observed. In general MI, IPW, and AIPW are valid under MAR, if models are appropriately specified. CC analysis is frequently biased under MAR. Missing not at random (MNAR) is when the probability of missing depends on the value of variables that are fully observed, including the unobserved value of the variable itself. Generally all methods are biased under MNAR.

A basic question for all the above MI, IPW, and AIPW methods is whether the observed data for the outcome variable should be included in the required imputation models or weight models. This is also related to how the covariate is missing, whether the missingness is completely at random, or depend on other covariates and/or the outcome, or the covariate itself. The argument in favor of including the outcome variable in these models is from the theoretical developments associated with missing data and MI. In general, it is well known that for inference about a quantity of interest it is necessary to include the outcome variable as one of the variables in the imputation model when developing a new prediction model. Omitting the outcome variable can lead to biased estimates.[9] In general notation, if Q is the quantity of interest, and $D = (D_{\text{obs}}, D_{\text{mis}})$ is the data where D_{obs} and D_{mis} denote the observed and missing data, then from a Bayesian perspective, inference about Q is based on its posterior distribution $P(Q|D_{\text{obs}})$. This posterior distribution can be written as $P(Q|D_{\text{obs}}) = \int P(Q|D_{\text{obs}}, D_{\text{mis}})P(D_{\text{mis}}|D_{\text{obs}})pD_{\text{mis}}$, and this applies whether Q is a simple parameter in a model or a more complex function, that is, such as the BS or the AUC. This formula is the recipe for MI and motivates imputation of the missing data using $P(D_{\text{mis}}|D_{\text{obs}})$, followed by inference for Q using the complete data (D_{obs}, D_{mis}), and repeating these steps many times and averaging them. Since in our setting the outcome variable is part of D_{obs}, it is clear that the outcome variable should be used as part of the imputation scheme. In practice, the general recommendation for MI is that the imputation model should include every variable that predicts the incomplete variable, and sometimes the imputation model can contain more variables than will be used in the final analysis.[10]

The intuitive argument against including the outcome variable in the models used for imputation is the belief that there is some circularity. Since we are trying to evaluate how good a model is at predicting outcome, the thinking is that we don't want to use the outcome to help impute the missing covariates, because then we will make the model look better than it really is. However, Moons et al argued that imputation of missing values using all other information will not create information. It only makes use of the strength of associations between predictors and outcomes present in the CCs, to enable valid analyses.[9] The additional intuitive argument against using the outcome variable is that the intended use of these models is in the situation where we want to make a prediction for a single patient and we only have covariates available and the outcome is not known. It is certainly a challenge of how to make a prediction for a single patient if some of the covariates are missing, but this is a different situation than ours of evaluating an existing prediction model using a new dataset.

In this article, we propose IPW and AIPW estimates of AUC and BS to handle the missing data and evaluate their prediction performance in comparison with MI by simulation. We focus on including the outcome or not in the weight models or imputation models. The missing mechanisms could be MCAR, MAR, and MNAR. We consider a variety of existing prediction models including ones that are both consistent with and not consistent with the internal data distribution, and ones that depend on a subset of the covariates. An example from prostate cancer is used as an illustration of the proposed methods.

2 | METHODS

We consider the setting in which we have available an internal dataset of size N, consisting of binary outcome Y and p-dimensional vector of covariates X. Let $R_i = 1$ if there are no missing X values for subject i, else $R_i = 0$ if there are missing values, and R_i's are conditionally independent. Assume there is an existing external model, that requires as input the variables X or a subset of the variables, and produces as output an estimate of the probability that $Y = 1$, denoted by $\hat{p}(Y = 1|X)$. We use notation I to denote distributions associated with the available or internal data, and E to denote the distributions associated with the external data that was used to build the existing model. Let $F_I(X)$ and $F_I(Y|X)$ denote the true probability distribution functions for the internal data. Thus, $F_I(X)$ is the density of X if X is continuous and $F_I(Y=y|X=x) = P_I(Y=y|X=x)$. Let $F_E(X)$ and $F_E(Y|X)$ denote the true distributions for the external data. We would expect some of the X's to be correlated with each other.

The existing model $\hat{p}(Y = 1|X)$ is an approximation to $F_E(Y = 1|X)$, and it is usually a monotonic function of a weighted combination of covariates, denoted as $g(\beta X)$. The estimates of β could be good estimates if, for example, the external dataset is large and good methods were used, or they could be poor estimates if the external dataset is small or poor methods were used. From the internal dataset with sample size N that are sampled from $F_I(X)$ and $F_I(Y|X)$, we can calculate the BS and AUC. The BS is given by

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{p}_i)^2, \tag{1}$$

where $\hat{p}_i = \mathrm{P}(Y = 1|X_i)$ is obtained from the existing model.

The AUC, which is equivalent to the Concordance-index (C-index) for a binary outcome, is estimated using

$$\mathrm{AUC/C\text{-}index} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} I(\beta X_i > \beta X_j)I(Y_i > Y_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} I(Y_i > Y_j)}. \tag{2}$$

An alternative way to estimate the AUC is to first estimate the ROC curve and then calculate the area under it. Let n_1 denote the number of cases, n_0 denote the number of controls, and $n_1 + n_0 = N$. Let X_1 denote the covariates in cases and X_0 denote the covariates in controls. The ROC curve depicts relative trade-offs between true positive rate (TPR) and false positive rate (FPR),

$$\begin{aligned}
\mathrm{TPR}(c) &= \Pr(\beta X_1 \geq c) = \frac{1}{n_1}\sum_{i=1}^{n_1} I(\beta X_i \geq c),\\
\mathrm{FPR}(c) &= \Pr(\beta X_0 \geq c) = \frac{1}{n_0}\sum_{j=n_1+1}^{N} I(\beta X_j \geq c),\\
\mathrm{ROC}(c) &= \mathrm{TPR}(\mathrm{FPR}^{-1}(c)),\\
\mathrm{AUC} &= \int_0^1 \mathrm{ROC}(c)dc.
\end{aligned} \tag{3}$$

The integration of ROC to calculate the AUC is performed numerically. The quantities called BS and AUC given above are estimates of population quantities, which we call $\text{TrueBrier}_I(\hat{p})$ and $\text{TrueAUC}_I(\hat{p})$. Given the distribution $F_I(X)$ and $F_I(Y|X)$, for any existing formula $\hat{p}$ that provides a probability that Y = 1 given X, the true BS is defined as

$$\text{TrueBrier}_I(\hat{p}) = \sum_{Y=0}^{1} \int_X (Y - \hat{p})^2 F_I(Y|X) F_I(X) dX. \tag{4}$$

For covariates in cases X_1 and controls X_0, denote their distributions as $F_I(X_1) = F_I(X|Y = 1)$ and $F_I(X_0) = F_I(X|Y = 0)$, respectively. Then the true AUC is

$$\text{TrueAUC}_I(\hat{p}) = \Pr(\beta X_1 > \beta X_0) = \int_{X_1} \int_{X_0} I(\beta X_1 > \beta X_0) F_I(X_1) F_I(X_0) dX_1 dX_0. \tag{5}$$

Equations (4) and (5) give the true values of BS and AUC for a fixed β. The goal is to get good estimates of these population quantities $\text{TrueAUC}_I(\hat{p})$ and $\text{TrueBrier}_I(\hat{p})$, using the available data in the internal dataset of size N. A good estimate is one that has small bias, low variability and is robust to misspecification of any models that are used in the estimation procedure.

Also note from Equation (4) that the true value depends on both $F_I(Y|X)$ and $F_I(X)$, and similarly for Equation (5). This makes it clear that even if the existing prediction model for Y given X is correct for the internal distribution, it will not usually lead to the same AUC and BS because these depend on the X distribution as well. In practice it would seem likely that the internal and external distributions of the X's do differ.

In real data analysis with large sample size, missing data are a common occurrence. Suppose our dataset has missing values for some covariates of X, and the missingness may be MCAR, MAR, or MNAR. The practical question we are trying to address is how to get a good estimate of $\text{TrueAUC}_I(\hat{p})$ and $\text{TrueBrier}_I(\hat{p})$ from the available dataset with missing covariates. The best conceivable estimates are the ones that would have been obtained using Equations (1) to (3) if there had been no missing data.

2.1 | CC analysis

Using only CCs (ie, $R_i = 1$) the simplest estimates are

$$\text{BS}_{\text{CC}} = \frac{\sum_{i=1}^{N} (Y_i - \hat{p}_i)^2 R_i}{\sum_{i=1}^{N} R_i}, \tag{6}$$

$$\text{C-index}_{\text{CC}} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} I(\beta X_i > \beta X_j) I(Y_i > Y_j) R_i R_j}{\sum_{i=1}^{N} \sum_{j=1}^{N} I(Y_i > Y_j) R_i R_j}. \tag{7}$$

For AUC,

$$\text{TPR}_{\text{CC}}(c) = \frac{\sum_{i=1}^{n_1} I(\beta X_i \geq c) R_i}{\sum_{i=1}^{n_1} R_i},$$

$$\text{FPR}_{\text{CC}}(c) = \frac{\sum_{j=1}^{n_0} I(\beta X_j \geq c) R_j}{\sum_{j=1}^{n_0} R_j}. \tag{8}$$

However, these estimates may be biased in MAR and MNAR settings and may lack efficiency in MCAR situations.

2.2 | Multiple imputation

When there is partially missing in X, we can do MI to impute the missing values based on the available data, then average the predicted BS and AUC from the multiple imputed datasets using Rubin's rule. The first step is to impute the missing values by drawing a value of X_{mis} from a model either for $F(X_{\text{mis}}|X_{\text{obs}}, Y)$ or for $F(X_{\text{mis}}|X_{\text{obs}})$, and then apply the external model on the imputed complete data to get the predictions of Y and calculate BS and AUC. The models used for imputation are typically linear regression for continuous X_{mis}, logistic regression for binary X_{mis}, polytomous regression for unordered categorical X_{mis}, and proportional odds model for ordered categorical X_{mis}, although more complicated models could be used. After repeating the first step for M times (we use $M = 5$), the average of the estimates of BS and AUC from the multiple imputed datasets gives the final single point estimate. When there is more than one covariate with missing values, a chained equation approach is used to impute the missing values sequentially.[10] The program `mice()` in `R` is used to implement the MIs and the different models mentioned above can be built with different options.

2.3 | Inverse probability weighting

IPW weights the CCs in the calculation of the quantity of interest. The weight (W_i) is the inverse probability of the observation being complete ($R_i = 1$) under different assumptions, so either $W_i = 1/\Pr(R_i = 1|X_i, Y_i)$ or $W_i = 1/\Pr(R_i = 1|X_i)$. We use logistic regression to build the model of either $\Pr(R_i = 1|X_i, Y_i)$ or $\Pr(R_i = 1|X_i)$ conditional on the fully observed covariates and outcome to get the estimates of the weight. Then

$$\text{BS}_{\text{IPW}} = \frac{\sum_{i=1}^{N} (Y_i - \hat{p}_i)^2 R_i W_i}{\sum_{i=1}^{N} R_i W_i}, \tag{9}$$

$$\text{C-index}_{\text{IPW}} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} I(\beta X_i > \beta X_j) I(Y_i > Y_j) R_i W_i R_j W_j}{\sum_{i=1}^{N} \sum_{j=1}^{N} I(Y_i > Y_j) R_i W_i R_j W_j}. \tag{10}$$

For AUC,

$$\begin{aligned} \text{TPR}_{\text{IPW}}(c) &= \frac{\sum_{i=1}^{n_1} I(\beta X_i \geq c) R_i W_i}{\sum_{i=1}^{n_1} R_i W_i}, \\ \text{FPR}_{\text{IPW}}(c) &= \frac{\sum_{j=1}^{n_0} I(\beta X_j \geq c) R_j W_j}{\sum_{j=1}^{n_0} R_j W_j}. \end{aligned} \tag{11}$$

With the TPR_{IPW} and FPR_{IPW}, ROC_{IPW}, and AUC_{IPW} can be calculated following (3).

2.4 | Augmented inverse probability weighting

The IPW method only uses the CCs, and ignores the subjects with missing data. One way to improve it is to include information from subjects with missing data, which is called AIPW. For ease of notation we describe the method in the situation of only one covariate having missing values. In the Appendix, we describe how to apply it when multiple covariates have missing values. First we build a model for the covariate with missing values on all the other covariates, that is, $F(X_{\text{mis}}|X_{\text{obs}}, Y)$ or $F(X_{\text{mis}}|X_{\text{obs}})$, to get the predicted mean X^*_{mis}, which is $E(X_{\text{mis}}|X_{\text{obs}}, Y)$ or $E(X_{\text{mis}}|X_{\text{obs}})$. This is a single imputation of the mean and is different from MI which incorporates random variation. The X^*_{mis} is created for that variable for all subjects and is different from MI which only imputes missing values. Then applying the external model to the dataset with X replaced by $X^* = (X^*_{\text{mis}}, X_{\text{obs}})$ gives $\hat{p}_i{}^*$. Combining this model with a model for the weight, the proposed AIPW estimator of the BS is

$$\text{BS}_{\text{AIPW}} = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{p}_i)^2 R_i W_i + (Y_i - \hat{p}_i{}^*)^2 (1 - R_i W_i). \tag{12}$$

A subject with complete data has $R_i = 1$, and contributes $(Y_i - \hat{p}_i)^2 W_i + (Y_i - \hat{p}_i{}^*)^2 (1 - W_i)$. A subject with missing values has $R_i = 0$ and contributes $(Y_i - \hat{p}_i{}^*)^2$. Because all the subjects with complete data or missing values are evaluated, the denominator is N.

For the C-index,

$$\text{C-index}_{\text{AIPW}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} I(Y_i > Y_j)\left\{I(\beta X_i > \beta X_j)R_iW_iR_jW_j + I(\beta X_i^* > \beta X_j^*)(1 - R_iW_iR_jW_j)\right\}}{\sum_{i=1}^{N}\sum_{j=1}^{N} I(Y_i > Y_j)}. \quad (13)$$

A pair of cases and controls X_i, X_j that are both complete has $R_i = 1, R_j = 1$, and contributes $I(\beta X_i > \beta X_j)W_iW_j + I(\beta X_i^* > \beta X_j^*)(1 - W_iW_j)$. Otherwise, a pair of cases and controls that has missing value, that is, $R_i = 0$ and/or $R_j = 0$ contributes $I(\beta X_i^* > \beta X_j^*)$.

For the area under the ROC curve method of calculating the AUC,

$$\text{TPR}_{\text{AIPW}}(c) = \frac{1}{n_1}\sum_{i=1}^{n_1} I(\beta X_i \geq c)R_iW_i + I(\beta X_i^* \geq c)(1 - R_iW_i),$$
$$\text{FPR}_{\text{AIPW}}(c) = \frac{1}{n_0}\sum_{j=1}^{n_0} I(\beta X_j \geq c)R_jW_j + I(\beta X_j^* \geq c)(1 - R_jW_j). \quad (14)$$

A subject with complete data has $R_i = 1$, and contributes $I(\beta X_i \geq c)W_i + I(\beta X_i^* \geq c)(1 - W_i)$. A subject with missing value has $R_i = 0$ and contributes $I(\beta X_i^* \geq c)$. With the TPR_{AIPW} and FPR_{AIPW}, ROC_{AIPW} and AUC_{AIPW} can be calculated following (3).

2.5 | Consistency of IPW and AIPW estimators

Considering the C-index using the IPW method. Let

$$U_{ij}(\theta, \gamma_1) = \theta I(Y_i > Y_j)R_iW_iR_jW_j - I(Y_i > Y_j)I(\beta X_i > \beta X_j)R_iW_iR_jW_j,$$

where W_i depends on the weight model which has parameters γ_1. Let $U_N(\theta, \gamma_1) = 0.5N^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[U_{ij}(\theta, \gamma_1) + U_{ji}(\theta, \gamma_1)\right]$, then it is straight forward to show that $\text{C-index}_{\text{IPW}}$ is the solution of $U_N(\theta, \gamma_1) = 0$. Let $U_E = E(U_N) = 0.5E\left[U_{ij}(\theta, \gamma_1) + U_{ji}(\theta, \gamma_1)\right]$. Let γ_1^* be the large sample limit of $\hat{\gamma}_1$ using the weight model $\Pr(R = 1|X_{\text{obs}}, Y; \gamma_1)$. When the weight model is correctly specified, that is, $\Pr(R = 1|X_{\text{obs}}, Y; \gamma_1^*) = \Pr(R = 1|X_{\text{obs}}, Y)$, and R_i's are conditionally independent, then $E(R_iW_iR_jW_j) = 1$, and it is clear that $U_E(\theta, \gamma_1^*) = 0$. Because $U_N(\theta, \gamma_1)$ converges uniformly to $U_E(\theta, \gamma_1)$, $\text{C-index}_{\text{IPW}}$ is a consistent estimator.

The proof for AIPW estimators is similar. Here, we mimic the proof in Long et al,[11] and first demonstrate double robustness for a slightly different estimator, which we label $\text{C-index}_{\text{AIPW*}}$ with

$$\text{C-index}_{\text{AIPW*}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} I(Y_i > Y_j)\left\{I(\beta X_i > \beta X_j)R_iW_iR_jW_j + E[I(\beta X_i > \beta X_j)](1 - R_iW_iR_jW_j)\right\}}{\sum_{i=1}^{N}\sum_{j=1}^{N} I(Y_i > Y_j)}$$

Let

$$V_{ij}(\theta, \gamma_1, \gamma_2) = \theta I(Y_i > Y_j) - I(Y_i > Y_j)\left\{I(\beta X_i > \beta X_j)R_iW_iR_jW_j + E[I(\beta X_i > \beta X_j)](1 - R_iW_iR_jW_j)\right\}$$

where W_i depend on weight model $\Pr(R = 1|X_{\text{obs}}, Y; \gamma_1)$ with parameters γ_1 and in $E[I(\beta X_i > \beta X_j)]$ the expectation is with respect to the distribution of the missing covariates and depends on the model $F(X_{\text{mis}}|X_{\text{obs}}, Y; \gamma_2)$ which has parameters γ_2. Let $V_N(\theta, \gamma_1, \gamma_2) = 0.5N^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[V_{ij}(\theta, \gamma_1, \gamma_2) + V_{ji}(\theta, \gamma_1, \gamma_2)\right]$, then it is straightforward to see that $\text{C-index}_{\text{AIPW*}}$ is the value of θ that solves $V_N(\theta, \gamma_1, \gamma_2) = 0$. Let $V_E = E(V_N) = 0.5E\left[V_{ij}(\theta, \gamma_1, \gamma_2) + V_{ji}(\theta, \gamma_1, \gamma_2)\right]$. It is easy to see that $V_N(\theta, \gamma_1, \gamma_2)$ converges uniformly to $V_E(\theta, \gamma_1, \gamma_2)$, thus the solution to $V_N(\theta, \gamma_1, \gamma_2) = 0$ converges to the solution of $V_E(\theta, \gamma_1, \gamma_2) = 0$.

Let γ_1^* be the probability limits of γ_1 using the weight model $\Pr(R = 1|X_{\text{obs}}, Y; \gamma_1)$. When the weight model is correctly specified, that is, $\Pr(R = 1|X_{\text{obs}}, Y; \gamma_1^*) = \Pr(R = 1|X_{\text{obs}}, Y)$, and R_i's are conditionally independent, then $E(R_iW_iR_jW_j) = 1$. Let γ_2^* be the probability limits of γ_2 using the model for the missing covariates $F(X_{\text{mis}}|X_{\text{obs}}, Y; \gamma_2)$. When the model is correctly specified, that is, $E(X_{\text{mis}}|X_{\text{obs}}, Y; \gamma_2^*) = E(X_{\text{mis}}|X_{\text{obs}}, Y)$, then $E\left\{I(Y_i > Y_j)\{E[I(\beta X_i > \beta X_j)] - I(\beta X_i > \beta X_j)\}\right\} = 0$.

When either working model is correctly specified, it is clear that $V_E(\theta, \gamma_1, \gamma_2) = 0$, and that the θ that solves $V_E(\theta, \gamma_1, \gamma_2) = 0$ is the true AUC. Because V_N converges uniformly to V_E, C-index$_{\text{AIPW}^*}$ is a consistent estimator.

The estimator we describe in Section 2.4, C-index$_{\text{AIPW}}$ is an approximation to C-index$_{\text{AIPW}^*}$, in which instead of calculating the conditional expectation $E[I(\beta X_i > \beta X_j)]$, we propose to use $I(\beta X_i^* > \beta X_j^*)$.

The proof of consistency is similar for BS and is shown in the Appendix.

3 | SIMULATION STUDIES

In this section, we present results of numerical studies to investigate the performance of the proposed methods under different settings. We consider three covariates and denote them as X_1, X_2, X_3. We consider situations where the given external model is based on all of X_1, X_2, and X_3, and situations where it is only based on X_1 and X_2. The true distribution for the internal data, $F_I(Y|X)$, is defined as

$$\text{logit}(Pr(Y = 1)) = 0.25 + 0.7X_1 + 0.6X_2 - 0.5X_3$$

The internal data are sampled from the above model. X_1, X_2, X_3 are sampled from $N(0, 1)$ and about 40% to 50% of X_1 is missing. The covariates can be independent, or correlated with $\text{cor}(X_1, X_3) = -0.5$. Four different external models are evaluated using the "internal" data; (M_1) the true model with X_1, X_2, and X_3; (M_2) the best model based on just X_1 and X_2; (M_3) a poor model based on X_1, X_2, and X_3 with wrong coefficients; and (M_4) an incorrect intercept model.

The simulation is conducted as follows:

(a) For M_1, we use the true coefficients, $M_1 = (0.25, 0.70, 0.60, -0.50)$. For M_2, we obtain the coefficients for the external model by generating a dataset of 100 000 observations from the true model, and fitting a logistic model based on X_1 and X_2. For independent covariates, $M_2 = (0.25, 0.67, 0.58, 0)$. For $\text{cor}(X_1, X_3) = -0.5, M_2 = (0.25, 0.91, 0.58, 0)$. It is noted that with independent covariates, the estimated coefficients are biased toward the null compared with the true model.[12] With correlated X_1, X_3 and X_3 is omitted, the estimates of the coefficients for X_1, X_2 are biased in opposite directions in the reduced model. For M_3, we obtain the coefficients by generating an external dataset with sample size 50. For independent covariates, $M_3 = (0.26, 0.66, 0.90, 0.39)$, and for correlated covariates, $M_3 = (0.53, -0.40, 0.88, -0.75)$. With such small sample size, the estimated coefficients are not close to the true values. For M_4, we set different prevalence's for the external data and internal data, and $M_4 = (1.00, 0.70, 0.60, -0.50)$.
(b) Based on the distributions $F_I(X), F_I(Y|X)$, get the true AUC and BS for each of M_1, M_2, M_3, and M_4 using their coefficients and Equations (4) and (5). We label these as the true target values.
(c) Sample internal data with $N = 1000$, and evaluate the external models M_1, M_2, M_3, M_4 on the internal data. Use different methods to handle the missing covariates in the internal data to estimate AUC and BS, repeat 1000 times to get the mean and standard deviation, and compare with each other and with the true target value calculated in (b).

We consider four different missingness mechanisms. For MCAR, the missing of X_1 is random with probability 0.4, that is, $\Pr(X_1 \text{ is missing}) = 0.4$. For MAR($X_2, X_3$), the missing of X_1 depends on other covariates X_2, X_3 with about 45% missing, $\Pr(X_1 \text{ is missing}) = \text{expit}(-0.5 + 2X_2 - 2X_3)$. For MAR($X_2, Y$), the missing of X_1 depends on both covariate X_2 and outcome Y with about 50% missing, $\Pr(X_1 \text{ is missing}) = \text{expit}(-0.5 + 2X_2 + Y)$. For MNAR, the missing of X_1 depends on the value of X_1 with about 45% missing, $\Pr(X_1 \text{ is missing}) = \text{expit}(-0.5 + 3X_1)$.

As listed in Table 1, we compared the validation of external models on the full internal data without missing values (Full), on CCs only, IPW with the weight model excluding outcome Y (IPW1) or including outcome Y (IPW2), MI with the imputation model excluding outcome Y (MI1) or including outcome Y (MI2). When calculating AUC by AIPW, the two methods, which are based on the C-index and the area under the ROC curve, respectively, gave similar results in terms of bias and efficiency with 40% to 50% missing of X_1, thus we show the results for the C-index using a weight model that excludes the outcome Y (AIPW1, AIPW3) or includes the outcome Y (AIPW2, AIPW4) and using an imputation model that excludes the outcome Y (AIPW1, AIPW2) or includes the outcome Y (AIPW3, AIPW4). For the IPW and AIPW methods the weight models are regarded as misspecified in the MAR(X_2, Y) situation if they don't include Y, that is, IPW1, AIPW1, and AIPW3, and all IPW and AIPW weight models are misspecified in the MNAR situation.

TABLE 1 List of methods for comparison

True target	True value based on internal data distribution
Full	Data without missing
CC	Complete cases analysis
IPW1[a]	Weight model uses X
IPW2	Weight model uses X and Y
MI1[b]	Imputation model uses X
MI2	Imputation model uses X and Y
AIPW1[a,b]	Weight model uses X, imputation model uses X
AIPW2[b]	Weight model uses X and Y, imputation model uses X
AIPW3[a]	Weight model uses X, imputation model uses X and Y
AIPW4	Weight model uses X and Y, imputation model uses X and Y

Abbreviation: CC, complete case.
[a]Methods for which the weight model is misspecified under MAR(X,Y).
[b]Methods for which the imputation model is misspecified.

In this simulation, `mice()` in `R` with linear regression using bootstrap is used to implement MI for the missing continuous covariates. `glm()` with logistic link was used to build weight models and `lm()` was used to calculate the predicted X_1^* in the AIPW method.

3.1 | Simulation results

Figures 1 and 2 show the simulation results of AUC and BS for existing model M_1 with independent covariates under MCAR, MAR(X_2, X_3), MAR(X_2, Y), and MNAR(X_1). The left column shows the bias of the various methods. As expected the full data analysis does achieve the true target AUC and BS. However, the CC analysis is unbiased only in the MCAR setting. MI with Y (MI2) is unbiased under MCAR and MAR, but without Y (MI1) the bias is more than 10% for both AUC and BS. All the IPW and AIPW methods are unbiased under MCAR and MAR(X_2, X_3), regardless of whether Y is included or not. Under MAR(X_2, Y) when Y is related to the missingness, the only unbiased IPW method (IPW2) is the one including Y, which indicates the importance of correct specification of the weight model. For AIPW2 and AIPW4, when the weight model includes Y, the results are unbiased. Without Y in the weight model, AIPW3 includes Y in the imputing model, and the results are unbiased too. However, when both weight model and imputing model exclude Y, as in AIPW1, the results are biased, especially for AUC. For the double robustness of AIPW, as least one of the weight model and imputing model need to be correctly specified. Under MNAR for which the missingness depend on X_1, all the methods are biased.

The right column shows the relative SD of the methods comparing with full data estimation. As expected all values are equal to 1.0 or larger. The variance of IPW is always the largest, since it only weights the CCs. The variance of AIPW is between IPW and MI, and is much smaller than IPW under MAR.

For the model M_2 with omitted covariate X_3, under all scenarios, the reduced model M_2 has lower AUC and higher BS compared with true target values for model M_1. This is to be expected since omitting an important covariate will generally lead to an inferior external model. As shown in Figures 3 and 4 the full model results do achieve the target true value for M_2, and they represent the best that could be achieved for M_2. The relative performance of the various MI, IPW, and AIPW methods for the handling the missing data compared with the full model results are quite similar to those shown in Figures 1 and 2, both for bias and SD.

We also considered using a poor external model M_3 with wrong coefficients. The results are shown in Figures 5 and 6. Again in comparison with full data analysis, the MI2, IPW2, AIPW2, and AIPW4 appear to give no bias, except in the MNAR case. The variability of the MI2 method is the smallest.

For the scenario when external data and internal data have different prevalence, we consider an existing model with the intercept = 1 while the other coefficients are the same as the true model. The changed intercept in M_4 has no influence on the AUC compared with the true value, since changing the intercept does not change the discrimination ability.

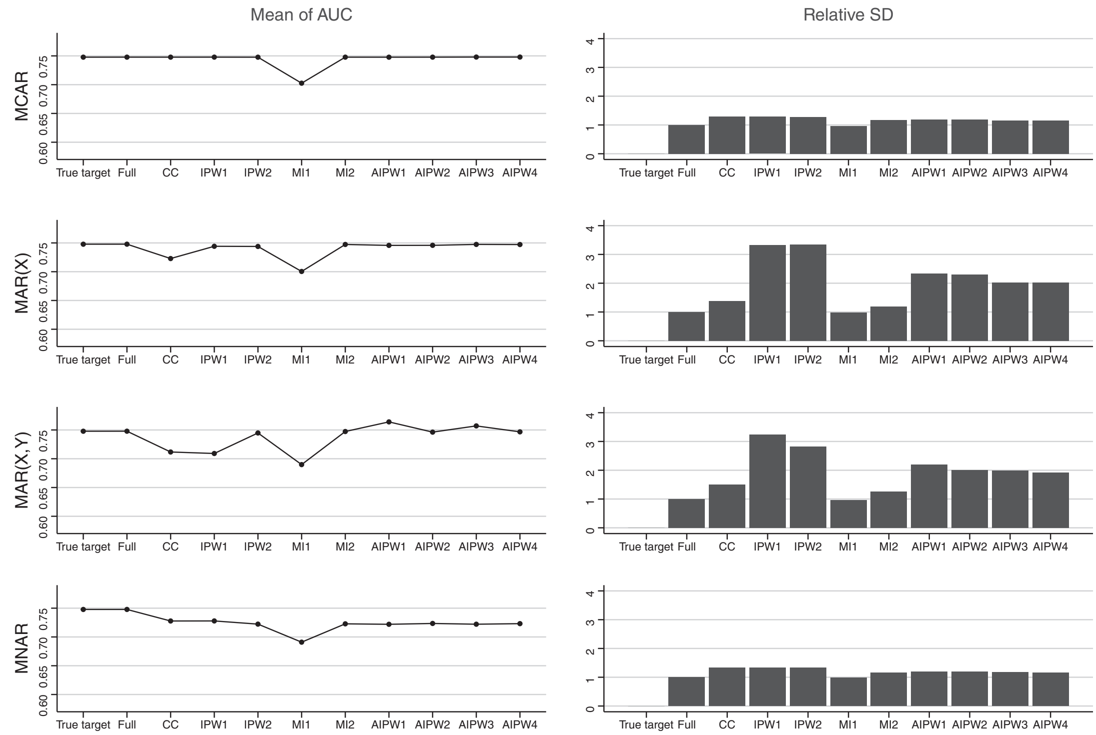

FIGURE 1 Simulation results of mean and relative SD of AUC for existing model M_1: correct model. Column left denotes mean AUC. Column right denotes SD relative to full data analysis. The four rows are different missingness mechanisms. AUC, area under the curve

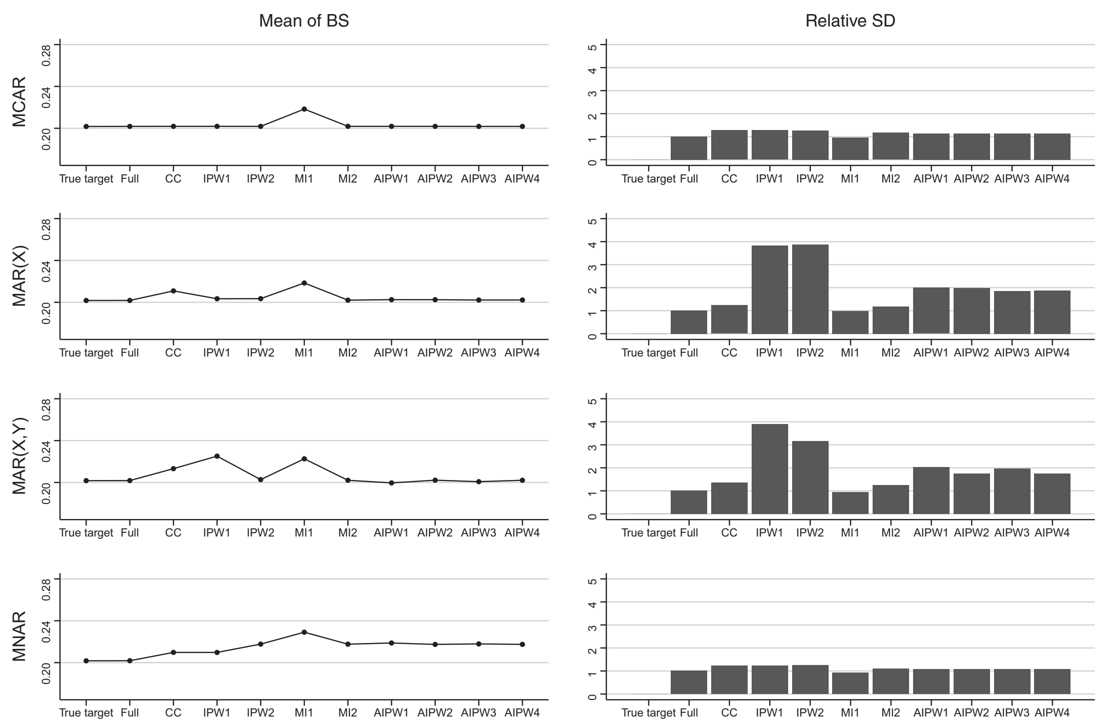

FIGURE 2 Simulation results of mean and relative SD of Brier score for existing model M_1: correct model. Column left denotes mean BS. Column right denotes SD relative to full data analysis. The four rows are different missingness mechanisms. BS, Brier score

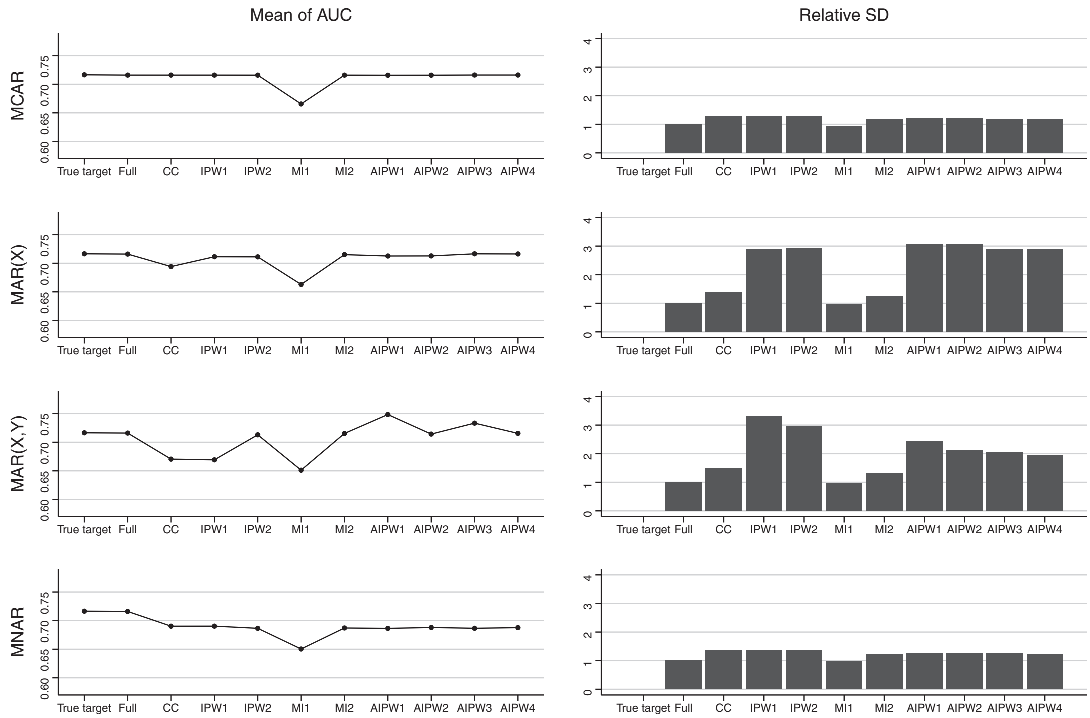

FIGURE 3 Simulation results of mean and relative SD of AUC for existing model M_2: best model based on just X_1, X_2. Column left denotes mean AUC. Column right denotes SD relative to full data analysis. The four rows are different missingness mechanisms. AUC, area under the curve

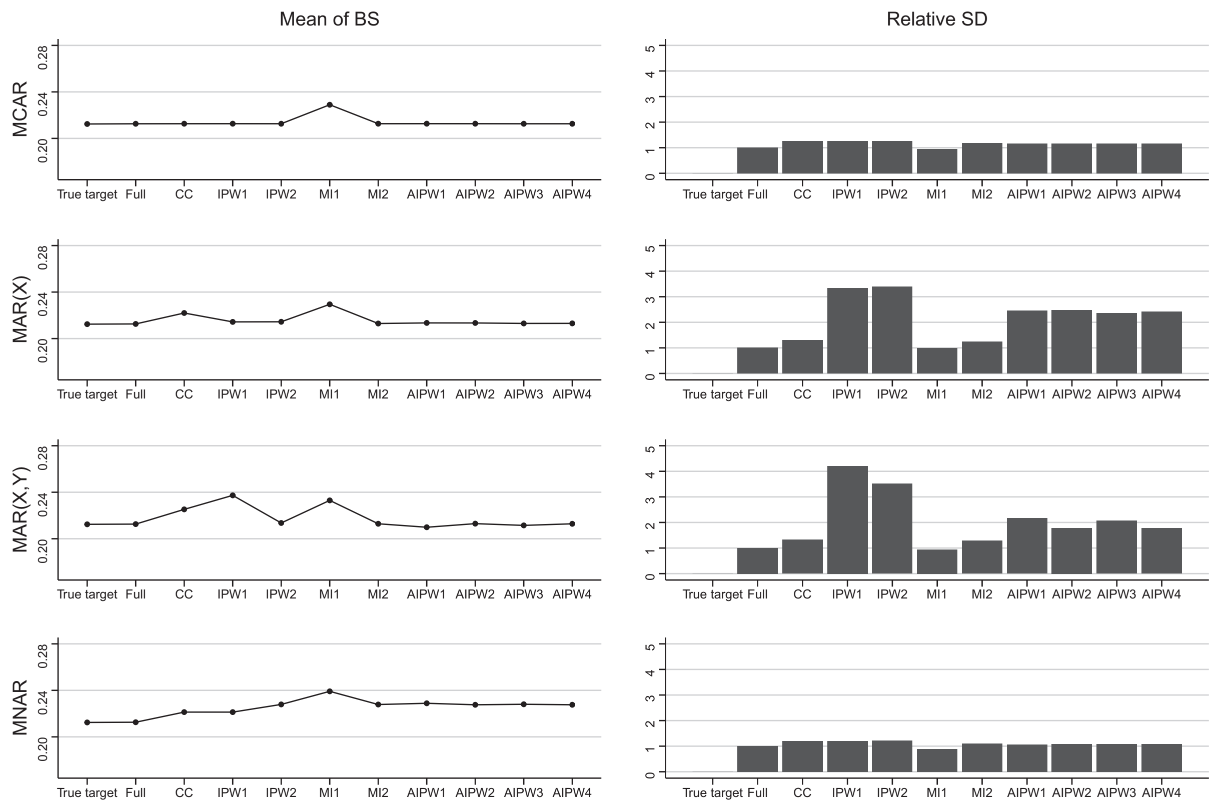

FIGURE 4 Simulation results of mean and relative SD of BS for existing model M_2: best model based on just X_1, X_2. Column left denotes mean BS. Column right denotes SD relative to full data analysis. The four rows are different missingness mechanisms. BS, Brier score

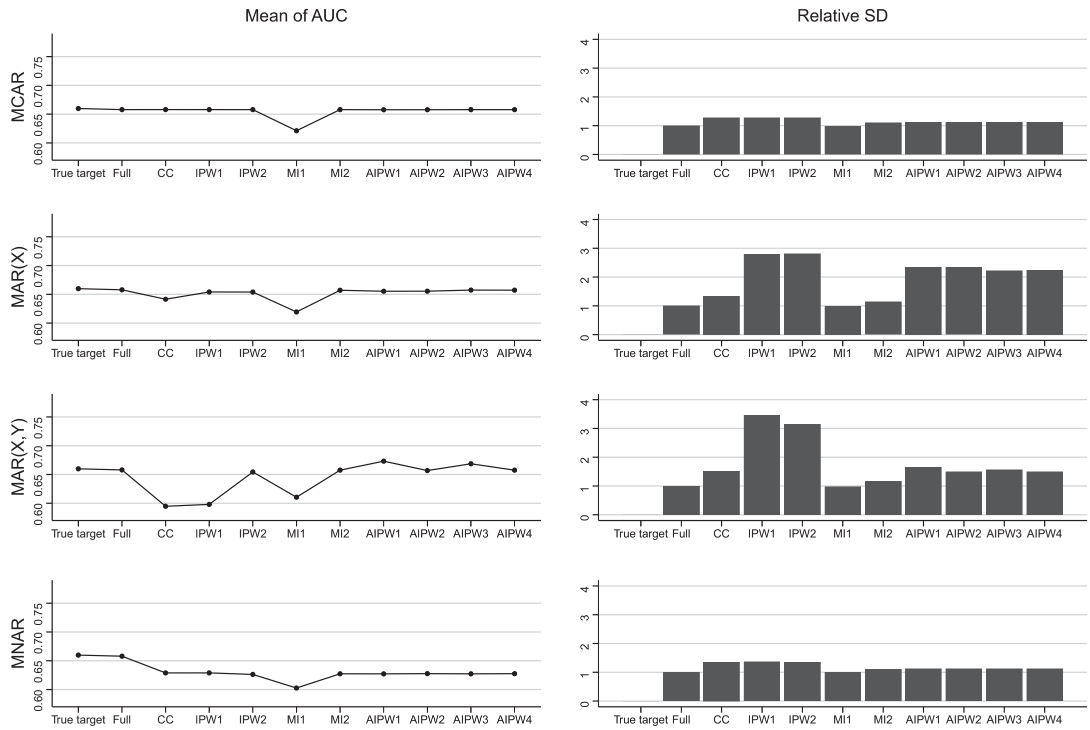

FIGURE 5 Simulation results of mean and relative SD of AUC for existing model M_3: poor model based on X_1, X_2, X_3. Column left denotes mean AUC. Column right denotes SD relative to full data analysis. The four rows are different missingness mechanisms. AUC, area under the curve

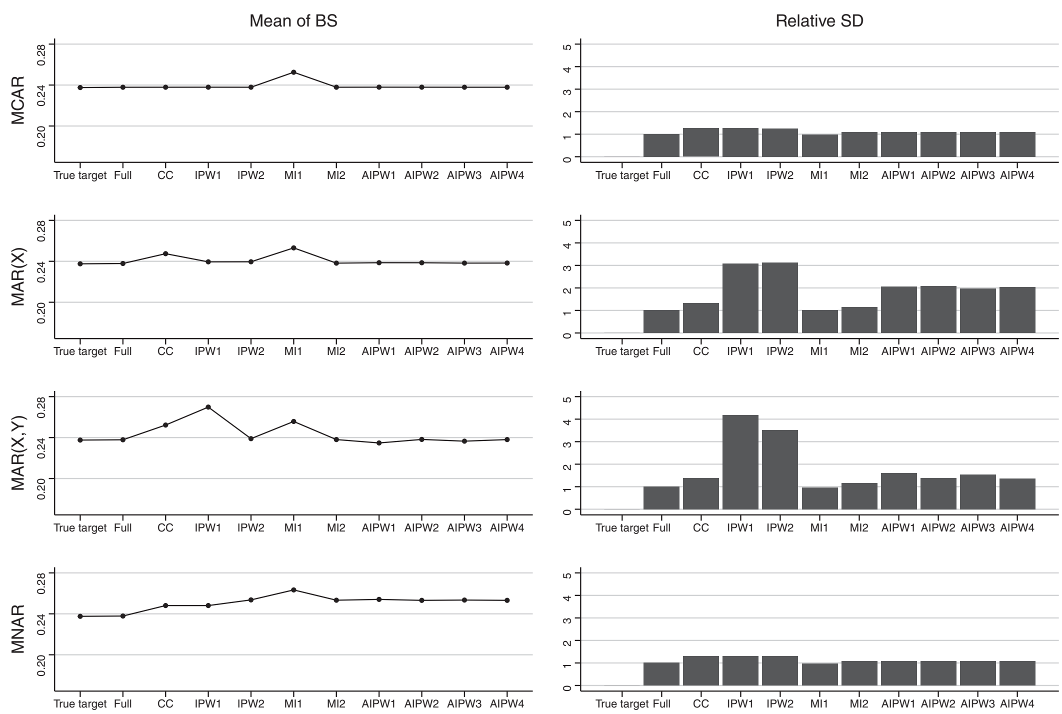

FIGURE 6 Simulation results of mean and relative SD of BS for existing model M_3: poor model based on X_1, X_2, X_3. Column left denotes mean BS. Column right denotes SD relative to full data analysis. The four rows are different missingness mechanisms. BS, Brier score

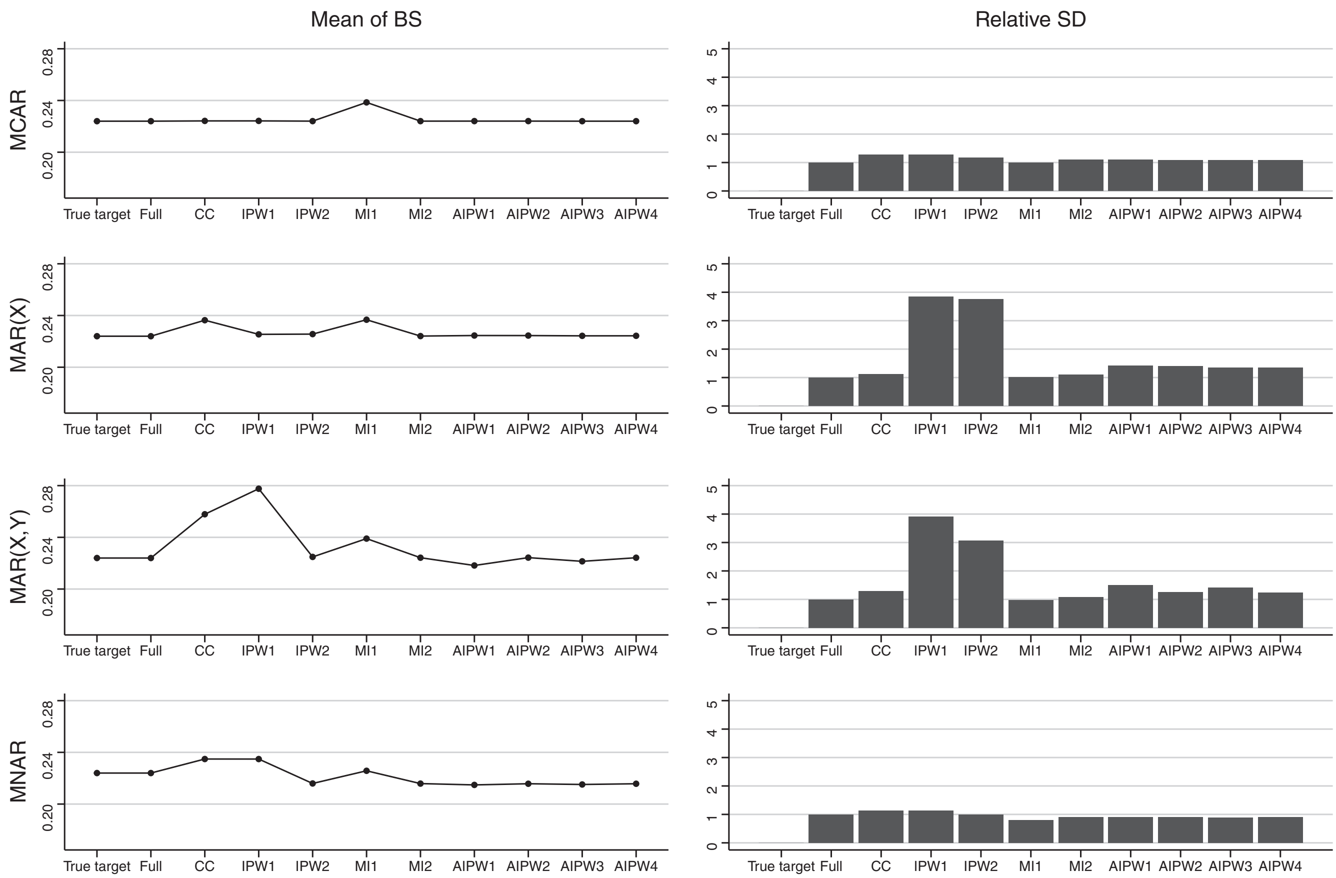

FIGURE 7 Simulation results of mean and relative SD of BS for existing model M_4: different intercept model. Column left denotes mean BS. Column right denotes SD relative to full data analysis. The four rows are different missingness mechanisms. BS, Brier score

The results are identical to those shown in Figure 1. The values of BS increased compared with situation M_1. As shown in Figure 7 the relative merits of the MI, IPW, and AIPW methods are similar to the other scenarios.

Overall, for the situations presented here, considering both bias and variability the best methods are MI2 and AIPW4. For correlated covariates, the conclusions are the same (see Appendix). With multiple missing covariates, the findings are broadly similar, but with some differences depending on the missingness pattern. The simulation results shown in the Appendix, suggest that here MI2 is the best method. The findings from additional simulations investigating the impact of sample size and percent missingness are also described in the Appendix.

We note that the model used to impute the missing X in MI2 and create X^* in AIPW3 and AIPW4 is slightly misspecified. Although it does regress X_1 on X_2, X_3 and Y, the assumed linear model is not the same as the true distribution for $X_1|X_2, X_3, Y$ based on how the data was generated from the true model. Furthermore, as noted in the consistency proof, we use an approximation to a doubly robust AIPW estimator, specifically we use $I(\beta X_i^* > \beta X_j^*)$ to approximate $E[I(\beta X_i > \beta X_j)]$. These two facts may explain the small bias in the AIPW3 method for the MAR(X_2, Y) case, because in fact neither the weight model nor the imputing model is correctly specified. However, the misspecified imputing model does not give any noticeable bias for the MI2 method. It is feasible to consider other approximations of $E[I(\beta X_i > \beta X_j)]$. Willamson et al[7] suggested a Monte Carlo approximation for general AIPW methods with missing covariates. However, in our settings we found that this lead to more bias and greater variability of the AIPW estimates than using the $I(\beta X_i^* > \beta X_j^*)$ approximation. We were surprised by this finding and do not have a satisfactory explanation of why it occurred.

4 | APPLICATION

In this section, we applied the proposed methods to evaluate the performance of an existing model for the risk of recurrence in men with Prostate Cancer. The Cancer of the Prostate Risk Assessment (CAPRA) score was published in 2005 and was based on an initial cohort consisting of >1400 men from the University of California, San Francisco.[13] A Cox proportional hazards regression model identified age, pretreatment prostate-specific antigen (PSA), Gleason score, percentage of biopsy cores positive for cancer (PPC), and clinical stage as significant factors associated with biochemical recurrence (BCR) or secondary treatment. Based on the results of the Cox analysis, points were assigned as in Table 2 to indicate relative risk. For each patient the points would be added to give an overall CAPRA score. The CAPRA score ranges from 0 to 10, and every 2-point increase in the score represents an approximate doubling of the risk. The distribution of the score and the 3-year recurrence-free survival (RFS) rate were reported in the publication, and are shown in Table 3. The AUC can be calculated from the CAPRA score itself, but the BS requires the predicted probabilities from Table 3.

We sought to estimate the performance of CAPRA using a separate dataset from the Mayo Clinic. The 1268 patients were treated with surgery between 2008 and 2012 and all patients before 2010 and half patients later were missing PPC values. So in total 90% of the patients were missing PPC. We considered 3-year RFS as a binary outcome. We included in our analysis all men who were followed more than 3 years or developed progression in 3 years. In total, 314 of the 1268 patients had a recurrence in 3 years. To validate the prediction of CAPRA score, we compared the CAPRA score with the outcome to get the AUC, and compared the RFS rate for each CAPRA score as in Table 3 with the outcome to get BS. Because 90% patients have missingness in PPC, we used PSA, Gleason score, T-stage, Age and/or the outcome to build the weight model for missingness and the imputation model of PPC in the IPW, AIPW, and MI methods. In the data analysis, `mice()` in `R` with logistic regression is used to implement MI for the missing binary PPC. `glm()` with logistic link was

TABLE 2 CAPRA score

Variable	Level	Points
PSA	2.0-6	0
	6.1-10	1
	10.1-20	2
	20.1-30	3
	>30	4
Gleason score (Primary/Secondary)	1-3/1-3	0
	1-3/4-5	1
	4-5/1-5	3
T stage	T1/T2	0
	T3a	1
Percent positive biopsy	<34%	0
	≥34%	1
Age	<50	0
	≥50	1

Abbreviations: CAPRA, Cancer of the Prostate Risk Assessment; PSA, prostate-specific antigen.

TABLE 3 CAPRA score distribution and predicted probabilities derived from the CAPRA score

CAPRA score	CAPRA score distribution	3-year RFS rate
0-1	27.9%	0.91
2	30.0%	0.89
3	20.6%	0.81
4	10.8%	0.81
5	5.8%	0.69
6	3.0%	0.54
7 or Greater	2.0%	0.24

Abbreviation: CAPRA, Cancer of the Prostate Risk Assessment.

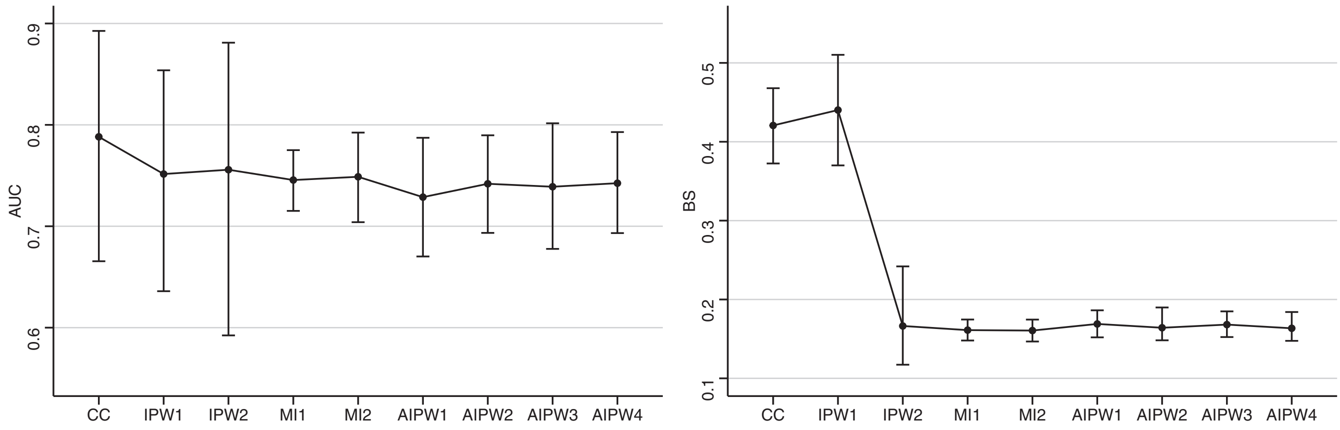

FIGURE 8 Varying estimates of mean and 95% confidence interval of AUC and Brier scores for prostate cancer example, based on how missing data are handled. AUC, area under the curve

used to build weight models and `glm()` with logistic link was used to calculate the predicted PPC in AIPW. A bootstrap was used to give 95% confidence intervals for AUC and BS.

Figure 8 shows the analysis results of different methods. The AUC ranged from 0.73 to 0.79, which is similar to other external validation studies of the CAPRA score for which the C-index for BCR ranged from 0.66 to 0.81.[14] On the other hand, the BS values were around 0.16 except for CC analysis and IPW with the weight model excluding the outcome variable, which were above 0.4. The CC analysis and IPW methods have much wider confidence intervals, while the MI and AIPW methods have comparable confidence intervals. Little's test was used and indicated the missingness is not MCAR ($P < .001$),[15] thus CC analysis is not an optimal choice. The Odds Ratio of PPC not missing and RFS observed was 24.1, indicating the missingness was strongly related to the outcome. Thus, the methods in which the weight model includes the outcome should be more reliable. The imputing model of PPC was built only on the 10% of patients with nonmissing data and was used to impute the other 90% later on, and there could be a large variation in the model, which could explain the ignorable difference between the two MI methods with or without outcome. The results for AUC and BS are different, probably because some CAPRA scores have the same RFS rate.

These results indicate the approaches to handle missing data can result in fairly large variation in model performance estimates. Based on the theoretical considerations and the simulation results, we believe the results from MI using the outcome (MI2) and AIPW using the outcome in the weight model and the imputation model (AIPW4) are the best to use, and they give very similar estimates for both BS and AUC in this example.

5 | DISCUSSION

We developed new AIPW estimators for predictive model performance metrics in the setting of missing data. This AIPW approach is shown to have good properties. We note that an AIPW estimator of the AUC has been previously proposed,[11] but for a different setting with auxiliary variables. Adapting this published approach to our setting does not lead to Equation (13), but rather an estimator with weights in the denominator as in Equation (10). When the weight model is correctly specified and with assumed independence of cases and controls, the expectation of the denominator in Equation (10) is equivalent to the denominator in Equation (13).

When there are missing observations in the internal data, MI and IPW can both be used to obtain unbiased estimates of BS and AUC if the imputation model or weight model is correctly specified. When the missingness doesn't depend on Y, IPW doesn't need to include Y in the weight model, while MI does need to include Y in the imputation model. When the missingness depends on Y, both IPW and MI need to include Y. The outcome variable should be included in the imputation model under all scenarios, because it provides information of the missing covariates. For IPW, the outcome only needs to be included in the weight model if the missingness depends on the outcome in order to get the correctly specified weight model. The findings in this article clearly support inclusion of the outcome variable Y in models that handle the missing covariates when evaluating an existing prediction model. Thus overall, even though in some situations for the IPW and AIPW methods it is not necessary, very little harm arose from including Y and there is the potential for considerable gain.

Our simulation results suggest that under small to moderate missingness AIPW can be more efficient than IPW, and also obtain approximate double robustness to misspecification of the weight model or the imputing model. Even when both models are misspecified, resulting estimates are still less biased than IPW or MI with the wrong weight model or imputing model. Further simulation shows that in terms of bias, AIPW is also less sensitive to the sample size or extreme weights comparing to IPW. Under all scenarios, MI has the best efficiency comparing to full data analysis. Under MCAR, AIPW has the same efficiency as MI, while under MAR, AIPW is less efficient than MI.

One limitation of the IPW and AIPW methods is when there are multiple covariates missing. In this situation there are different possible ways in which the weight model and the imputation model can be constructed. In the special cases of blocked missingness or monotone missingness there are natural ways to construct these models, and in the simulation study we found similar performance to that of the situation with a single missing covariate. When the missingness is scattered there are more choice of how to implement the imputation model, and our simulation results suggest that AIPW can in fact be a less desirable method than IPW. It is possible that further research may suggest alternative ways of using the weights or alternative ways of defining the AIPW estimator, that has improved performance in this and other more challenging situations. With multiple missing covariates the MI methods are still relatively easy to apply by using the chained equation approach to impute the missing values sequentially, and the simulation results suggest it is clearly more efficient.

The derivations in this article revealed that the true values of AUC and BS are population quantities that depend on both the distribution of the X covariates and the $Y|X$ distribution in the population. So one should not necessarily expect the AUC and BS to be the same from one population to the next. This is perhaps well known to others, and in fact obvious for the AUC. If one population has a much narrower range of X values, then it will be harder to discriminate subjects in that population, so the AUC will be lower, even if the model is an accurate description of the $Y|X$ distribution in both populations.

The problem we consider in this article is how to estimate the correct AUC and BS for a different population than the one that was used to develop the prediction model, when (i) we do not have access to the data that was used to develop the model and (ii) the dataset we have from the different population has some missing covariate values. There are a broad set of other problems associated with missing covariates and risk prediction models. One is how to develop a model, for which a much cited reference is Moons et al[9] Another set of problems is how to implement an existing risk prediction model for an individual subject when that subject has some missing covariates, and also will not have the outcome known. Different situations and possibilities exist here. The model developer may have set up methods to use in the case of missing data for the individual subject, such as 2^k different models, one for each pattern of missingness. It is our observation that developers of models rarely provide explicit rules for producing predicted probabilities for an individual subject with missing covariates. So implicit in the intended use of their model is that all the required input covariates will be available or attainable for the specific subject. If a particular required input covariate is known to be hard to obtain, then it would seem that the onus is on the developer of the model to provide a rule or a guidance on how their model should be used for an individual subject. In practice we think that it will frequently be the case that all the required covariates will be available because they can probably be attained at that point in time by ordering further tests or taking further measurements. Alternatively for a subject with missing values the user of the model may simply try a range of values for the missing variables, to give a range of predicted probabilities for the specific subject, analogous to sensitivity analysis. If the user of the model has access to the training data, then the question becomes how to make use of these data. Alternatively, the user of the model may have access to their own dataset, with information on both the covariates and outcomes for people in this dataset, and if the individual subject can be considered as coming from the same population as this dataset, then the question again is how to make use of these data. These different challenges have received limited attention in the statistical literature,[16,17] but have been expounded upon in a recent publication.[18]

A challenge related to the one considered in this article is how to evaluate an existing prediction model in a different population when the data from this population has missing values in some of the X variables, but also in the outcome Y for some subjects. We did not study this situation, but one option is to simply remove the people with missing values before calculating AUC and BS. Other options are to apply a MI approach or develop an extension of the IPW and AIPW methods. We hypothesize that these options would give better estimates of AUC and BS than the option of removing subjects.

Another situation worthy of study, is how to evaluate an existing prediction model, in a different population, when that different population does not have measured one of the needed input variables for the prediction model. This would seem to be an impossible task, unless extra information is available, either in the form of additional data or knowledge of the joint distribution of the missing variable with the other variables.

ACKNOWLEDGEMENT

This work was partially supported by U.S. National Institutes of Health grants CA129102 and CA059827.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Pin Li https://orcid.org/0000-0003-4508-3762
Matthew J. Schipper https://orcid.org/0000-0001-6728-4720

REFERENCES

1. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*. 2010;21(1):128.
2. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. vol. 793: John Wiley & Sons; 2019.
3. Wood AM, Royston P, White IR. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biom J*. 2015;57(4):614-632.
4. Janssen KJ, Donders ART, Harrell FE Jr, et al. Missing covariate data in medical research: to impute is better than to ignore. *J Clin Epidemiol*. 2010;63(7):721-727.
5. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res*. 2013;22(3):278-295.
6. Mao L. On causal estimation using-statistics. *Biometrika*. 2017;105(1):215-220.
7. Williamson EJ, Forbes A, Wolfe R. Doubly robust estimators of causal exposure effects with missing data in the outcome, exposure or a confounder. *Stat Med*. 2012;31(30):4382-4400.
8. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61(4):962-973.
9. Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol*. 2006;59(10):1092-1101.
10. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377-399.
11. Long Q, Zhang X, Johnson BA. Robust estimation of area under ROC curve using auxiliary variables in the presence of missing biomarker values. *Biometrics*. 2011;67(2):559-567.
12. Neuhaus JM, Jewell NP. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*. 1993;80(4):807-815.
13. Cooperberg MR, Pasta DJ, Elkin EP, et al. The University of California, san Francisco Cancer of the prostate risk assessment score: a straightforward and reliable preoperative predictor of disease recurrence after radical prostatectomy. *J Urol*. 2005;173(6):1938-1942.
14. Brajtbord JS, Leapman MS, Cooperberg MR. The CAPRA score at 10 years: contemporary perspectives and analysis of supporting studies. *Eur Urol*. 2017;71(5):705-709.
15. Little RJ. A test of missing completely at random for multivariate data with missing values. *J Am Stat Assoc*. 1988;83(404):1198-1202.
16. Marshall G, Warner B, MaWhinney S, Hammermeister K. Prospective prediction in the presence of missing data. *Stat Med*. 2002;21(4):561-570.
17. Janssen KJ, Vergouwe Y, Donders ART, et al. Dealing with missing predictor values when applying clinical prediction models. *Clin Chem*. 2009;55(5):994-1001.
18. Hoogland J, van Barreveld M, Debray TP, et al. Handling missing predictor values when validating and applying a prediction model to new patients. *Stat Med*. 2020;39(25):3591-3607.
19. Day NE, Kerridge DF. A general maximum likelihood discriminant. *Biometrics*. 1967;23:313-323.
20. Pepe MS, Thompson ML. Combining diagnostic test results to increase accuracy. *Biostatistics*. 2000;1(2):123-140.

How to cite this article: Li P, Taylor JMG, Spratt DE, Karnes RJ, Schipper MJ. Evaluation of predictive model performance of an existing model in the presence of missing data. *Statistics in Medicine*. 2021;40:3477–3498. https://doi.org/10.1002/sim.8978

APPENDIX A. DIFFERENCES BETWEEN OPTIMIZING THE LIKELIHOOD, THE AUC AND THE BS

BS measures the mean squared difference between the predicted probability and the actual outcome of an event across all subjects. The lower the BS is for a set of predictions, the better the predictions are calibrated. When we evaluate an existing model such as a logistic model on the internal dataset, the BS will be minimized when the external model is the same as internal model, that is, $F_E(Y|X) = F_I(Y|X)$.

Proof. Assume the $F_I(Y|X)$ as expit(αX) and $F_E(Y|X)$ as expit(βX).

BS

$$
\begin{aligned}
&= \sum_Y \int_x (Y - \hat{p})^2 F_I(Y|X) F_I(X) dX \\
&= \sum_Y \int_X \left(Y - \frac{1}{1+\exp(-\beta X)} \right)^2 \left(\frac{1}{1+\exp(-\alpha X)} \right)^Y \left(\frac{\exp(-\alpha X)}{1+\exp(-\alpha X)} \right)^{(1-Y)} F_I(X) dX \\
&= \int_X \left[\left(\frac{\exp(-\beta X)}{1+\exp(-\beta X)} \right)^2 \frac{1}{1+\exp(-\alpha X)} + \left(\frac{1}{1+\exp(-\beta X)} \right)^2 \frac{\exp(-\alpha X)}{1+\exp(-\alpha X)} \right] F_I(X) dX \\
&= \int_X \frac{\exp(-\alpha X) + \exp(-\beta X)^2}{(1+\exp(-\beta X))^2 (1+\exp(-\alpha X))} F_I(X) dX.
\end{aligned}
$$

If for any X, $\frac{\exp(-\alpha X)+\exp(-\beta X)^2}{(1+\exp(-\beta X))^2(1+\exp(-\alpha X))}$ is minimized, then the integral over X will be minimized.

let $A = \exp(-\alpha X)$, $B = \exp(-\beta X)$, then the function can be written as

$$
\frac{A + B^2}{(1+B)^2(1+A)}
$$

Take derivative w.r.t B, we get:

$$
\frac{2B(1+B)^2(1+A) - (A+B^2)2(1+B)(1+A)}{(1+B)^4(1+A)^2} = \frac{2(B-A)}{(1+B)^3(1+A)}
$$

When $B < A$, the function will decrease, When $B > A$, the function will increase. Thus, it will be minimized at $B = A$, that is, when $F_E(Y|X) = F_I(Y|X)$. ∎

AUC, which measures the area under the ROC Curve, indicates how well the predicted probabilities for the cases are separated from the controls. The question is under logistic models will the AUC be maximized when the external model is same as the internal model, that is, $F_E(Y|X) = F_I(Y|X)$? The answer is it depends. The coefficients in the logistic regression model are not chosen to maximize the AUC, rather the coefficients are chosen to maximize the likelihood. In practice, these two sets of coefficients will frequently, but not always, be quite similar. However, if complete discrimination is possible, the maximum likelihood logistic regression coefficients will estimate the coefficients which separate the population.[19,20]

APPENDIX B. CONSISTENCY OF IPW AND AIPW ESTIMATORS FOR BS

Considering the BS using the IPW method. Let

$$
U_i(\theta, \gamma_1) = \theta R_i W_i - (Y_i - \hat{p}_i)^2 R_i W_i,
$$

where W_i depend on weight model with parameters γ_1. Let $U_N(\theta, \gamma_1) = N^{-1} \sum_{i=1}^N U_i(\theta, \gamma_1)$, and it is straight forward that BS_{IPW} is the solution of $U_N(\theta, \gamma_1) = 0$. Let $U_E = E(U_N) = E(U_i(\theta, \gamma_1))$.

Let γ_1^* be the probability limits of γ_1 using the weight model $\Pr(R = 1|X_{\text{obs}}, Y; \gamma_1)$. When the weight model is correctly specified, $\Pr(R = 1|X_{\text{obs}}, Y; \gamma_1^*) = \Pr(R = 1|X_{\text{obs}}, Y)$, then $E(R_i W_i) = 1$, and it is clear that $U_E(\theta, \gamma_1) = 0$. Because $U_N(\theta, \gamma_1)$ converges uniformly to $U_E(\theta, \gamma_1)$, BS_{IPW} is a consistent estimator.

The proof is similar for AIPW estimator. We first demonstrate consistency for a slightly modified estimator, which we call BS_{AIPW*} with

$$BS_{AIPW*} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{p}_i)^2 R_i W_i + E[(Y_i - \hat{p}_i)^2](1 - R_i W_i)$$

Let

$$V_i(\theta, \gamma_1, \gamma_2) = \theta - \left\{(Y_i - \hat{p}_i)^2 R_i W_i + E[(Y_i - \hat{p}_i)^2](1 - R_i W_i)\right\},$$

where W_i depend on weight model with parameters γ_1 and $E[(Y_i - \hat{p}_i)^2]$ depend on the model for missing covariates with parameters γ_2. Let $V_N(\theta, \gamma_1, \gamma_2) = N^{-1}\sum_{i=1}^{N} V_i(\theta, \gamma_1, \gamma_2)$, then it is straightforward to see that BS_{AIPW*} is the solution of $V_N(\theta, \gamma_1, \gamma_2) = 0$. Let $V_E = E(V_N) = E(V_i(\theta, \gamma_1, \gamma_2))$. It is easy to see that $V_N(\theta, \gamma_1, \gamma_2)$ converges uniformly to $V_E(\theta, \gamma_1, \gamma_2)$, thus the solution to $V_N(\theta, \gamma_1, \gamma_2) = 0$ converges to the solution of $V_E(\theta, \gamma_1, \gamma_2) = 0$.

Let γ_1^* be the probability limits of γ_1 using the weight model $\Pr(R = 1|X_{obs}, Y; \gamma_1)$. When the weight model is correctly specified, $\Pr(R = 1|X_{obs}, Y; \gamma_1^*) = \Pr(R = 1|X_{obs}, Y)$, then $E(R_i W_i) = 1$. Let γ_2^* be the probability limits of γ_2 using the model for the missing covariates $F(X_{mis}|X_{obs}, Y; \gamma_2)$. When the model is correctly specified, that is, $F(X_{mis}|X_{obs}, Y; \gamma_2^*) = F(X_{mis}|X_{obs}, Y)$, then $E\{E[(Y_i - \hat{p}_i)^2] - (Y_i - \hat{p}_i)^2\} = 0$. When either working model is correctly specified, it is clear that $V_E(\theta, \gamma_1, \gamma_2) = 0$, and that the θ that solves $V_E(\theta, \gamma_1, \gamma_2) = 0$ is the true BS. Because V_N converges uniformly to V_E, BS_{AIPW*} is a consistent estimator.

For the actual estimator BS_{AIPW} described in Section 2.4 instead of calculating $E[(Y_i - \hat{p}_i)^2]$ where the expectation is over the distribution $F(X_{mis}|X_{obs}, Y; \gamma_2^*)$, we propose to use $(Y_i - \hat{p}_i{}^*)^2$ as an approximation.

APPENDIX C. ADDITIONAL SIMULATION RESULTS FOR CORRELATED COVARIATES

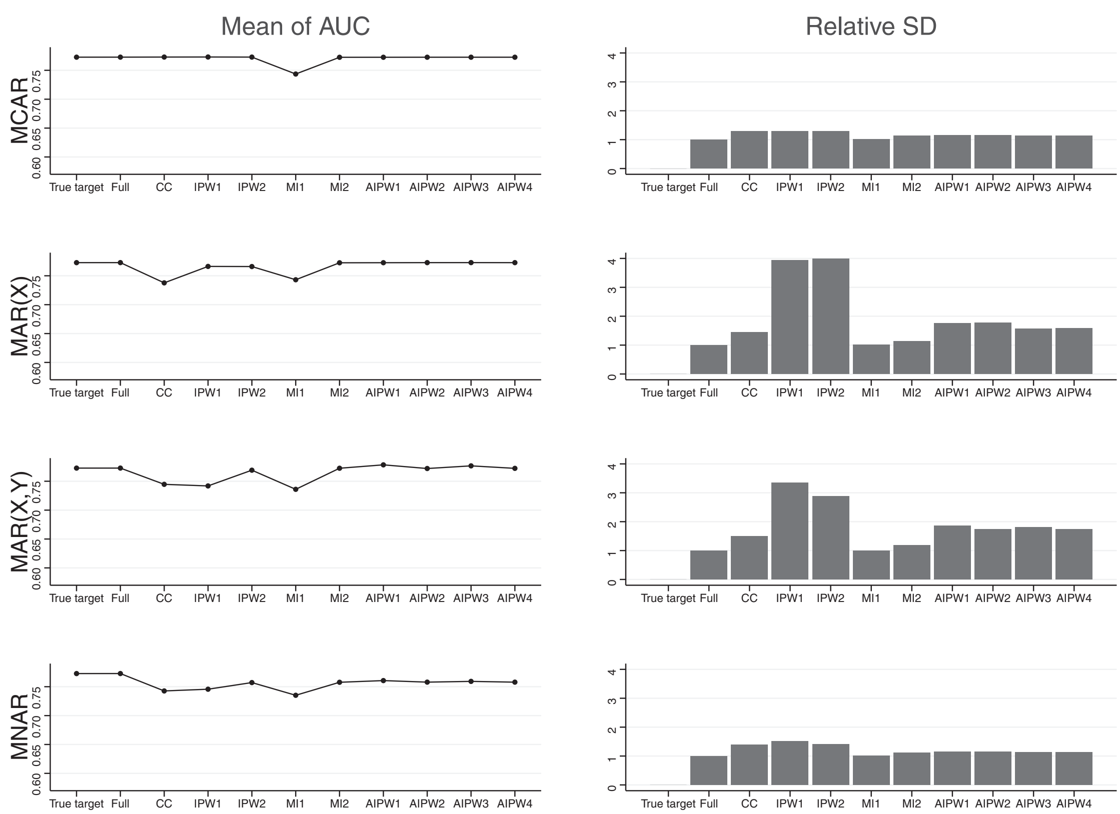

FIGURE C1 Simulation results of mean and relative SD of AUC for existing model M_1: $\mathrm{cor}(X_1, X_3) = -0.5$. Column left denotes mean AUC. Column right denotes SD relative to full data analysis. The four rows are different missingness mechanisms. AUC, area under the curve

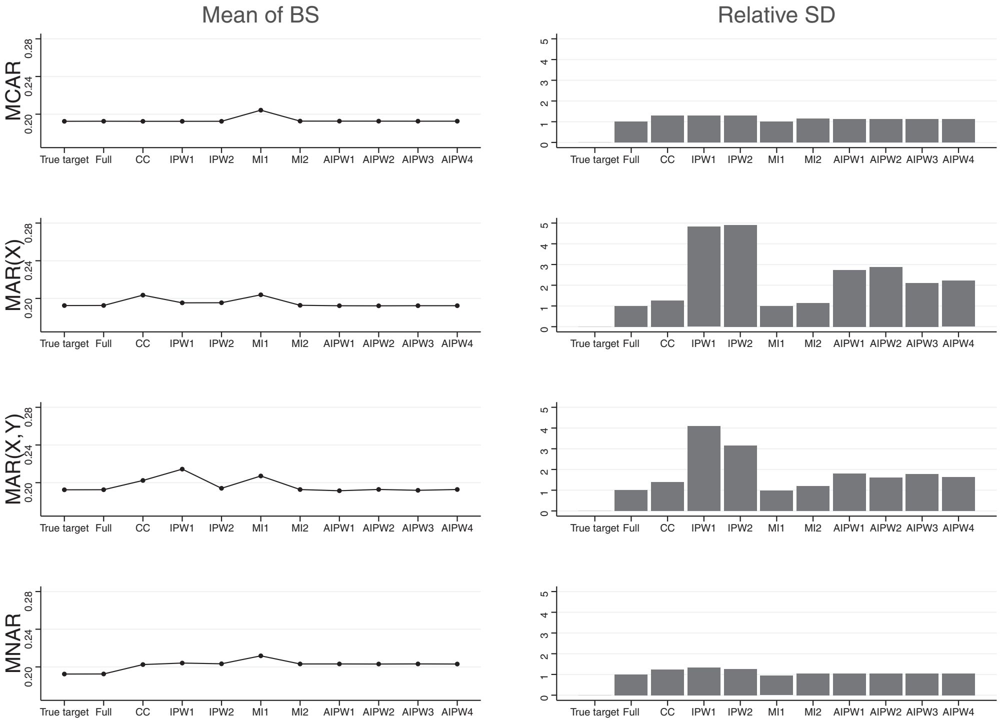

FIGURE C2 Simulation results of mean and relative SD of BS for existing model M_1: $\text{cor}(X_1, X_3) = -0.5$. Column left denotes mean BS. Column right denotes SD relative to full data analysis. The four rows are different missingness mechanisms. BS, Brier score

APPENDIX D. IMPLEMENTING AIPW AND IPW ESTIMATORS WHEN MORE THAN ONE VARIABLE HAS MISSING VALUES

We propose the IPW and AIPW estimates of AUC and BS for a single missing covariate in the main text and extend it here to more than one variable with missingness. We discuss how to build weight models and models for the missing covariates under different missing patterns.

First, we consider the block missing of covariates. Without loss of generality, consider the model with outcome Y and covariates X_1, X_2, X_3, and both X_2, X_3 are missing in some subjects. Let R_2 indicate X_2 is observed and R_3 indicate X_3 is observed, then $\Pr(R=1) = \Pr(R_2 = 1, R_3 = 1)$. The weight model can be built by $\Pr(R=1|X_1, Y)$ or $\Pr(R=1|X_1)$, using the fully observed covariates with the outcome or not. The models to impute X_2^* and X_3^* can be built separately, with $F(X_2|X_1, Y)$, $F(X_3|X_1, Y)$ or $F(X_2|X_1), F(X_3|X_1)$ from the data of subjects with $R=1$, and then obtain the predictions of X_2^* and X_3^* for all the subjects.

Next we look at a scattered pattern of missingness in the covariates. Use the same notation above with X_1 fully observed and X_2, X_3 are missing in some subjects. The weight model can be built by $\Pr(R=1|X_1, Y)$ which indicate the CCs without any missing, but may not capture the missingness for each covariate. Alternatively we can assume that the missingness of X_2 and X_3 are independent, then $\Pr(R=1) = \Pr(R_2=1)\Pr(R_3=1)$. The weight models for R_2 and R_3 can be built separately by $\Pr(R_2=1|X_1, Y)$, $\Pr(R_3=1|X_1, Y)$ or $\Pr(R_2=1|X_1)$, $\Pr(R_3=1|X_1)$, using the fully observed covariates with the outcome or not. The models to impute X_2^* and X_3^* can be built separately as in block missingness. In numerical studies we found the best results when the model to impute X_2 was built from the observations with $R_2=1$ and the model to impute X_3 was built from the observations with $R_3=1$.

For the monotone missingness, X_1 is fully observed and both X_2, X_3 are missing in some subjects. For those with X_2 observed, X_3 is missing in some subjects too, with the probability of missing X_3 can depend on the value of X_2 under the MAR scenario. Now $\Pr(R=1) = \Pr(R_3=1|R_2=1)\Pr(R_2=1)$ and we can build the model for R_2 using all the subjects and the model for R_3 using the subjects with X_2 observed. The models to impute X_2^* and X_3^* can be built separately as in block missing using the fully observed covariate X_1 with the outcome or not. Alternatively, the model to impute X_2^* can be built with $F(X_2|X_1, Y)$ or $F(X_2|X_1)$ from subjects with $R_2=1$ and get the predictions of X_2^* for all the subjects. Then the model to impute X_3^* can be built with $F(X_3|X_1, X_2, Y)$ or

$F(X_2|X_1, X_2)$ from subjects with $R_3 = 1$ and get the predictions of X_3^* using X_2^* as the predictor covariate for all the subjects.

APPENDIX E. SIMULATION RESULTS WHEN MORE THAN ONE VARIABLE HAS MISSING VALUES

We consider the same model with true coefficients as for M_1 and the covariates are independent. For block missing, similar to the single covariate missing, we consider MCAR: block missing of X_2, X_3 with probability of 0.4; MAR (X_1): block missing of X_2, X_3 depends on the value of fully observed covariate X_1; MAR (X_1, Y): block missing of X_2, X_3 depends on the value of X_1, Y; MNAR: block missing of X_2, X_3 depends on the value of X_2, X_3. The fraction of observations that are fully observed in these four situations are 60%, 60%, 50%, and 60%. Figure E1 shows the simulation results with 1000 replications for AUC, and the results are similar to Figure 1 for the single covariate missing situation.

For scattered missingness, we assume the missing of X_2 and X_3 are conditionally independent. For MCAR: missing of X_2 has probability of 0.4 and missing of X_3 has probability of 0.2; MAR (X_1): missing of X_2 depends on the value of fully observed covariate X_1 and missing of X_3 depends on X_1 too with a different probability; MAR (X_1, Y): missing of X_2 and X_3 depends on the value of X_1, Y with different probabilities; MNAR: missing of X_2 depends on the value of X_2 and missing of X_3 depends on the value of X_3. The fraction of observations that are fully observed in these four

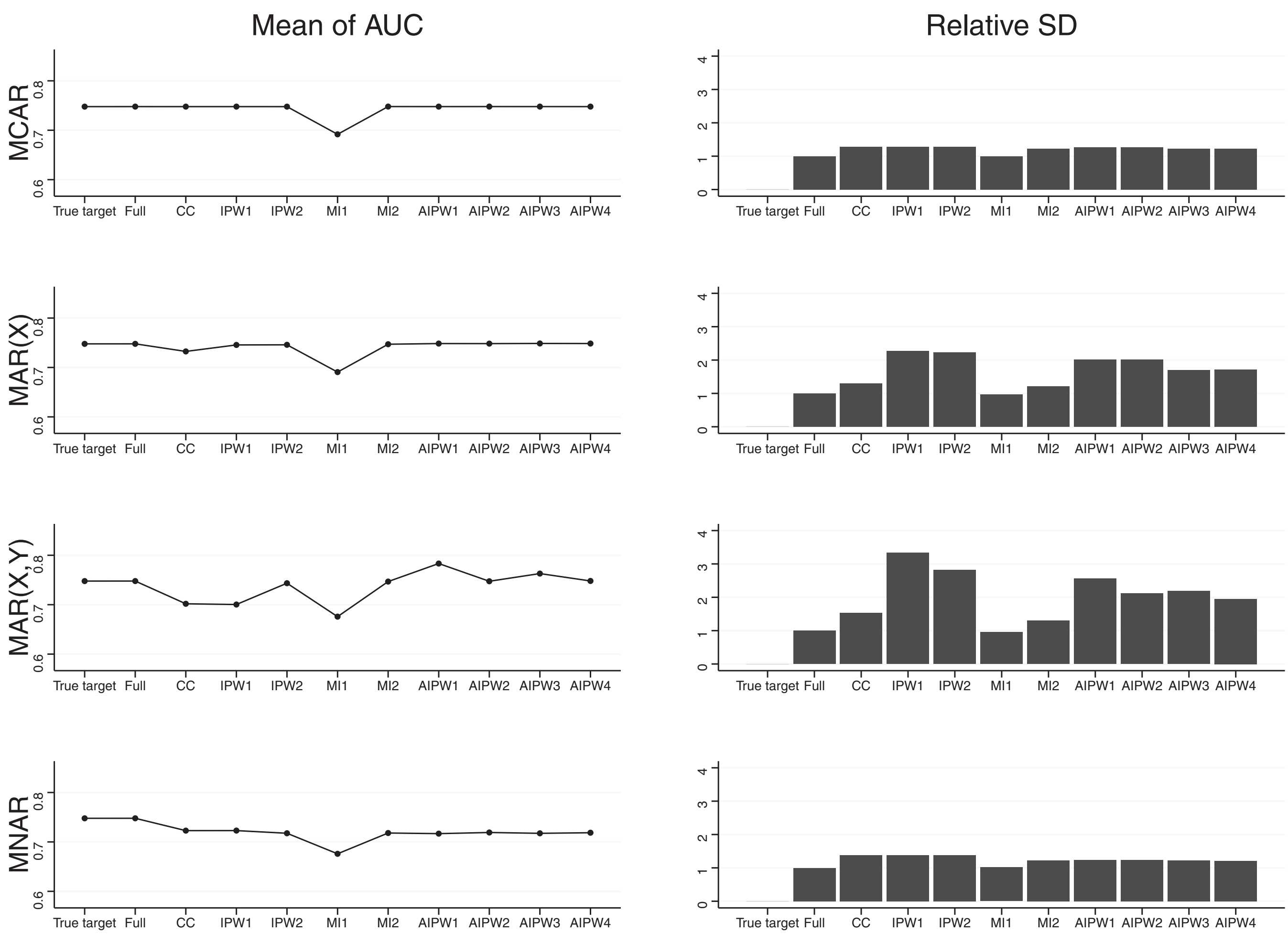

FIGURE E1 Simulation results of mean and relative SD of AUC for existing model with block missingness of two covariates. Column left denotes mean AUC. Column right denotes SD relative to full data analysis. The four rows are different missingness mechanisms. AUC, area under the curve

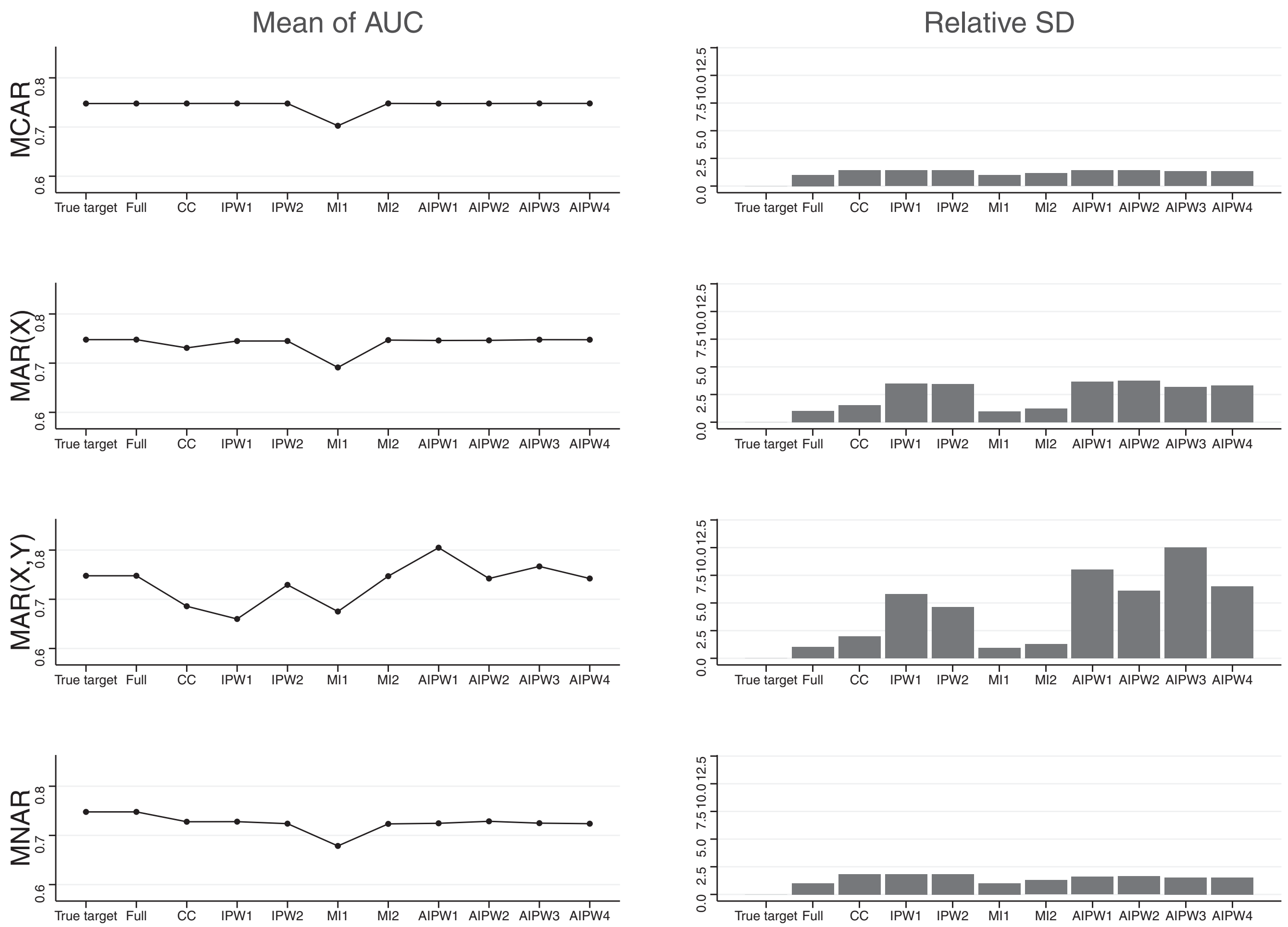

FIGURE E2 Simulation results of mean and relative SD of AUC for existing model with scattered missingness of two covariates. Column left denotes mean AUC. Column right denotes SD relative to full data analysis. The four rows are different missingness mechanisms. AUC, area under the curve

situations are 48%, 40%, 30%, and 33%. As shown in Figure E2, under MCAR, MAR(X) and MAR(X,Y), the IPW and AIPW methods can get unbiased estimates when the models for $\Pr(R_2 = 1), \Pr(R_3 = 1)$ or the model to calculate X_2^*, X_3^* are correctly specified. But the variance are much higher in comparison to MI methods, especially for AIPW under MAR(X,Y).

For monotone missing, we assume the subjects with missing in X_2 have missing in X_3 and some subjects with X_2 observed have missing in X_3 too. For MCAR: missing of X_2 has probability of 0.4 and for those with X_2 observed, missing of X_3 has probability of 0.5; MAR (X_1): missing of X_2 depends on the value of fully observed covariate X_1, and for those with X_2 observed, missing of X_3 depends on X_1 and X_2; MAR (X_1, Y): missing of X_2 depends on the value of X_1 and Y, and for those with X_2 observed, missing of X_3 depends on X_1, X_2 and Y; MNAR: missing of X_2 depends on the value of X_2, and for those with X_2 observed, missing of X_3 depends on the value of X_3. The fraction of observations that are fully observed in these four situations are 30%, 40%, 50%, and 30%. We compared different choices for the models to obtain X_3^*, either it includes X_2 or independent of X_2, and we saw no difference of the simulation results. In further simulations we saw that using X_2^* to predict X_3^* does not help when X_2, X_3 are correlated. As shown in Figure E3, under MCAR, MAR(X), and MAR(X,Y), the IPW and AIPW methods can get unbiased estimates when the weight model of $\Pr(R_2 = 1), \Pr(R_3 = 1 | R_2 = 1)$ or the model to calculate X_2^*, X_3^* are correctly specified. The AIPW methods are more efficient than IPW methods.

In conclusion, the extension of the IPW and AIPW methods to multiple covariates missing is feasible and have good performance under block missing and monotone missing.

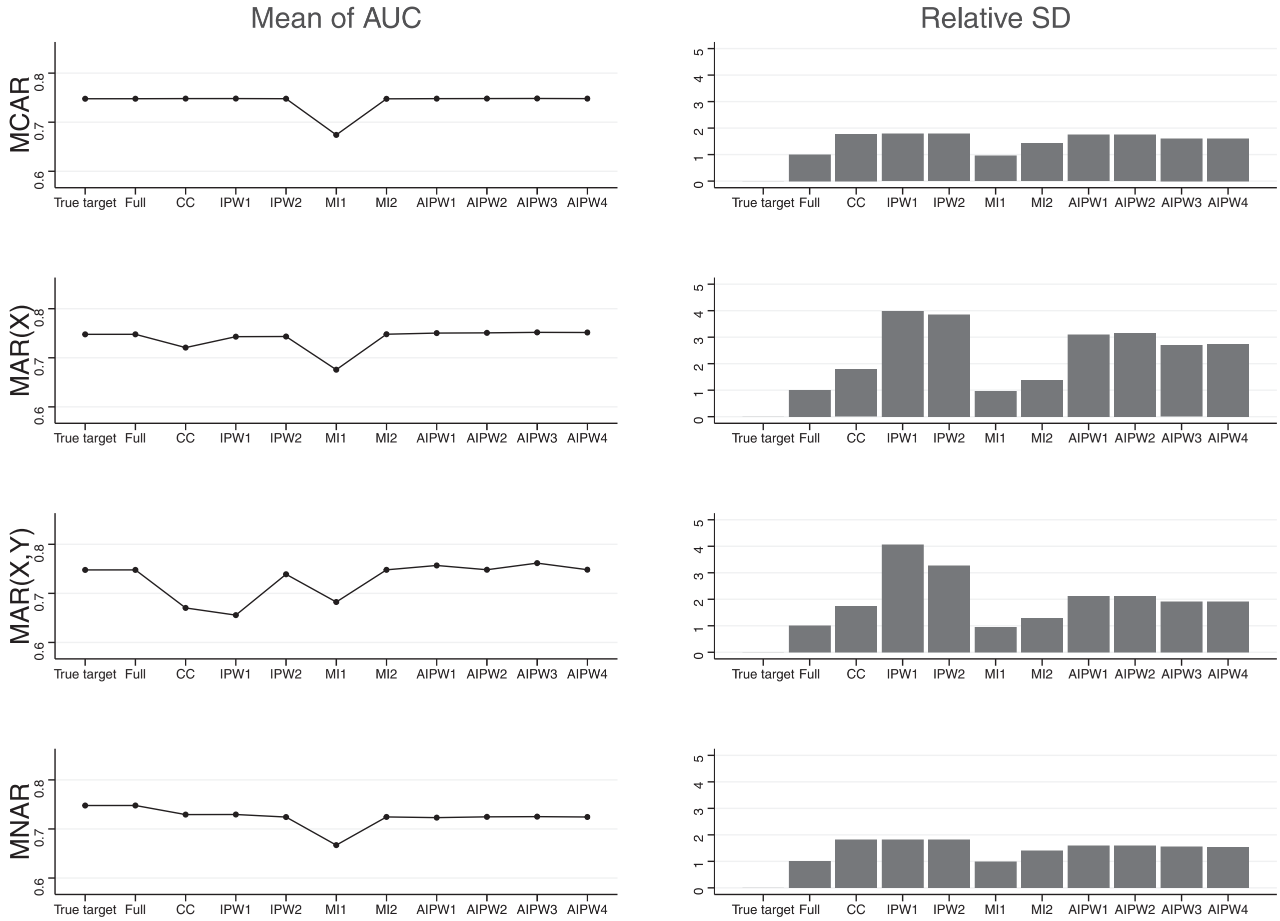

FIGURE E3 Simulation results of mean and relative SD of AUC for existing model with monotone missingness of two covariates. Column left denotes mean AUC. Column right denotes SD relative to full data analysis. The four rows are different missingness mechanisms. AUC, area under the curve

APPENDIX F. FINDINGS FROM SIMULATIONS WHERE SAMPLE SIZE AND AMOUNT OF MISSINGNESS IS VARIED

In further work, we investigated the impact of sample size and percent missingness on the performance of the methods, and also considered an alternative AIPW estimator. With smaller sample size, we observed that IPW2 is slightly biased under MAR and that the SD of IPW methods are smaller and similar to AIPW methods. On the other hand, with bigger sample size, the SD of IPW is a lot larger than that of AIPW. We found that small sample size has most impact on the IPW and AIPW performance in situations where there are some extreme weights. Truncating the very high weights does reduce the variability of the IPW and AIPW methods, but also increase their bias.

In the simulations presented in Figures 1 to 7, the missingness rate of X_1 is about 40% to 50%. With less missingness of X_1, the differences between the methods are smaller under all missing mechanisms. With 80% missing of X_1 the performance of the IPW and AIPW methods do deteriorate. For the M1 setting IPW2 is biased for both AUC and BS under MAR. For AUC, AIPW3 is more biased than AIPW1 under MAR(X_2, Y), and SD of AIPW1 and AIPW2 is larger than IPW. The worse performance is strongly affected by the distribution of the weights, and deteriorates substantially when there are extreme weights.

For the results presented in the article we found very little difference between the alternative ways of calculating the AUC, that is either using the C-index or by calculating the area under the estimated ROC curve via Equation (14). With 80% missingness rate for X_1 we did find differences between the methods. The ROC version AUC_{AIPW} results in more biased AUC than the C-index version AUC_{AIPW} under MAR, and furthermore AIPW2 and AIPW4 showed some bias.