Handling missing values in risk prediction modeling: a comparative simulation study on parametric and machine learning multiple imputations

Yuxin Wu

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2023

Reading Committee:

Yu-Ru Su, Chair

Rebecca Yates Coley

Program Authorized to Offer Degree:

Biostatistics

University of Washington

Abstract

Handling missing values in risk prediction modeling: a comparative simulation study on parametric and machine learning multiple imputations

Yuxin Wu

Chair of the Supervisory Committee:
Yu-Ru Su
Department of Biostatistics - Public Health

Risk prediction is a critical tool in preventive medicine, enabling precision prevention for diseases. Electronic health record (EHR) data offers a rich source for constructing risk models, capturing detailed clinical information from patient cohorts. However, missing data poses a prevalent challenge in EHR analysis, and multiple imputation (MI) is a popular strategy for handling missing data. In this thesis, we employed simulations to compare different MI methods (parametric MI, MI using Random Forest, MI using Gradient Boosting Machines and MI using Principal Component Analysis) within the context of risk prediction modeling. Our investigation focused on evaluating predictive performance, encompassing measures of predictive accuracy and precision, for risk prediction models developed and assessed in datasets processed with various MI strategies. Furthermore, we explored two facets: (1) the impacts of including or omitting the outcome variable during MI, and (2) the impacts of model misspecification of higher-order effects during MI. We also used breast surveillance mammogram examination data from breast cancer survivors in the Breast Cancer Surveillance Consortium (BCSC) as the input for part of the bootstrapping and data illustration complementary to the simulation study. Our results revealed that the adoption of machine learning-based imputation methods did not lead to superior model performance compared to traditional parametric imputation. We recommend against including the outcome variable in

the imputation model for the test set since it may raise concerns of over-optimistic predictive performance. Although it is not the focus of this thesis, we also recommend being cautious of using Random Forest as the risk prediction model for similar prediction modeling settings.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I am profoundly grateful for the guidance and support I have received throughout the journey of completing my Master's thesis. I extend my heartfelt appreciation to my advisor, Dr. Yu-Ru Su, whose unwavering dedication, insightful feedback, and invaluable mentorship have been instrumental in shaping the course of my research. I also extend my sincere gratitude to my second reader, Dr. Yates Coley, for their valuable insights and constructive suggestions that enriched the quality of my work.

I want to thank the Breast Cancer Surveillance Consortium (BCSC) for sharing the data with me. The insights gained from the data have greatly enriched the depth and quality of this study. Data collection in the BCSC for the information used in this study was additionally supported, in part, by funding from the NCI (U54CA163303), the Patient-Centered Outcomes Research Institute (PCS-1504–30370), and the Agency for Health Research and Quality (R01 HS018366–01A1). The collection of cancer and vital status data used in this study was supported in part by several state public health departments and cancer registries throughout the United States. For a full description of these sources, please see: https://www.bcsc-research.org/about/work-acknowledgement. All statements in this thesis, including its findings and conclusions, are solely those of mine and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee, nor those of the NCI, the NIH, or the Agency for Health Research and Quality. I also want to thank the participating women, mammography facilities, and radiologists for the data they have provided for this study.

I am deeply grateful to the Department of Biostatistics for providing me with an exceptional academic environment in the past two years. Specifically, I would like to thank Prof.

Ken Rice and Prof. Amy Willis, whose instruction in BIOST 514 and 515 ignited my passion for biostatistics. Their guidance and expertise not only imparted valuable knowledge but also fostered a profound interest in the field that has been a driving force behind my research pursuits. I am also immensely thankful to Prof. Katie Kerr for her exceptional mentorship in the consulting course. She gave detailed feedback on each appointment summary I wrote and helped me develop my communication skills. Additional thanks to Prof. Yen-Chi Chen and Prof. Fang Han, I enjoyed the fantastic lectures on statistical inference, which made the statistical theories less intimidating.

Last but not least, I would like to express my deepest gratitude to my parents for their unconditional love, encouragement, and belief in my abilities. Your unwavering support has been my driving force, and I am truly fortunate to have you by my side. Additionally, I want to extend my thanks to my husband, whose unwavering patience, understanding, and encouragement have been my constant source of strength.

Chapter 1

INTRODUCTION

Risk prediction plays a crucial role in preventive medicine as it is an essential tool for risk-based precision prevention for diseases. Electronic health record (EHR) data emerges as a powerful and valuable source for building risk models since EHRs contain detailed clinical information collected from patient cohorts, reflecting the characteristics of the general population. However, missing data is a ubiquitous issue that researchers often encounter while analyzing data in particular EHRs. Missing data can occur in EHRs due to various reasons such as (1) non-response from a survey (e.g. family history of diseases), and (2) various levels of data ascertainment from different data sources. Based on different causes of missing data, Rubin [35] described three missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Missing data are MCAR if whether a value is missing is unrelated to both the observed and unobserved values. This means that the missingness can be safely ignored in the analysis, and the complete-case analysis would be unbiased. Nevertheless, MCAR is a strong assumption that is often unrealistic in practice, and it can be difficult to verify its validity. Missing data are MAR if whether a value is missing depends on the set of observed responses but is unrelated to the missing values. This means that the missingness is not random, but it can be accounted for in the analysis by including the observed data in the modeling. For example, if the missingness of a variable is related to another observed variable, including that variable in the analysis can account for the missingness and reduce bias. Missing data are MNAR in scenarios where the probability of missingness depends on both the observed and unobserved data and cannot be predicted by the observed data alone. This means that the missingness is not random and cannot be accounted for in the analysis by including the observed data.

MNAR is the most challenging mechanism to deal with, as it is often impossible to know the reasons for the missingness or to predict the missing values. Therefore, understanding the mechanisms of missingness is essential for handling missing data in the analysis. Ignoring or misinterpreting the mechanism of missingness can lead to biased and invalid results [14].

There are various strategies for handling missing data depending on the missing mechanism. One of the most common approaches is to exclude observations with missing data (listwise deletion), which can be a straightforward solution but may reduce the sample size and potentially introduce selection bias [11]. Another approach is to use imputation techniques to estimate the missing values based on the observed data. Imputation methods can be classified as single imputation (SI) and multiple imputation (MI). SI involves replacing each missing value with a single estimated value, such as the mean/median/mode of the observed data, or a predicted value based on a regression model [15]. However, SI does not reflect the uncertainty about the imputed values and can lead into biased estimates even if data are MCAR [25]. In contrast, MI [30] involves creating multiple complete datasets, each with different imputed values based on plausible statistical models. MI provides more accurate and reliable estimates of the missing values and allows for valid statistical inference. MI had been first worked on by Rubin [36], and most of the applications have been on solving missing data problems using parametric imputation methods such as regression imputation. MI estimates the missing values `m` times conditional on the observed values, with each time incorporating a random variation to account for uncertainty about the missing values. Ultimately, we could obtain `m` completed datasets with the same observed data but different imputed missing data. Multiple imputation by chained equations (MICE), also called "Fully Conditional Specification", is a commonly used technique in MI [47]. It imputes missing cells in variables in a dataset through repeated steps of an iterative series of predictive models. In each iteration cycle, each specific variable in the data set is imputed based on a distribution conditional on other variables. MICE provides a lot of flexibility in the sense that the model to impute each variable doesn't have to be parametric. Semi-parametric, non-parametric and machine-learning methods can also be used to estimate the missing values [54, 42, 13, 28].

Compared to parametric methods, non-parametric and machine-learning approaches could favorably provide more flexibility in their formulation [6] and can be powerful when adequate training samples are provided. If there are complex relations among variables, using the popular R package `mice` [46] with default settings would produce unsatisfactory results unless users had manually specified any potential non-linear or interaction effects in the imputation model for each incomplete variable. Stekhoven and Bühlmann [42] proposed and implemented the R package `missForest` for predicting missing values using the Random Forest algorithm. `missForest` has shown high imputation accuracy with various missing rates[51]. Solaro et al. also [41] conducted an extensive simulated investigation under SI setting to compare the imputation accuracy of three different non-parametric methods in the presence of multiple missing data patterns. Moreover, some works have been done [37, 23] in comparing the bias and precision of model parameters obtained using Random Forest-based MI and parametric MI by simulation studies. It shows that the Random Forest method led to less biased parameter estimates and better confidence interval coverage when there are nonlinear effects or interactions among the variables. However, to our knowledge, a systematic investigation comparing the predictive performance of risk prediction models built using parametric MI and MI approaches leveraging non-parametric and machine learning techniques is still lacking.

In the context of building regression or prediction models, another factor that may impact the imputation approach is the inclusion of the outcome variable in the imputation process. Moons et al. [32] conducted simulations and showed that regression coefficients based on MI including the outcome variable were close to the ground truth, while MI without the outcome variable yielded biased coefficients. In the context of risk prediction, Sisk et al. [38] found that SI consistently led to models with better predictive accuracy when the outcome is omitted from the imputation. However, this is because the inclusion of the outcome in the imputation model artificially strengthens the relationship between the predictors and the outcome [1]. MI can overcome this issue by introducing randomness during imputation.

In this work, we conducted extensive simulations to compare different MI methods that

use non-parametric and machine learning techniques including Random Forest, Gradient Boosting Machines as well as Principal Component Analysis to the parametric imputation method (regression imputation) in the context of risk prediction modeling. Specifically, we investigated the predictive performance, including predictive accuracy and precision, of risk prediction models developed and evaluated in datasets processed with various MI approaches. In addition, we investigated (1) the impacts of including and removing the outcome variable during MI, and (2) the impacts of model misspecification of higher-order effects during MI. To reflect the complexity of real-world data, we used breast surveillance mammogram examination data from breast cancer survivors in the Breast Cancer Surveillance Consortium (BCSC) as the input for part of the bootstrapping and data illustration complementary to the simulation study.

Chapter 2

METHODS

2.1 Multiple imputation approaches

MI is one of the imputation approaches to handle missing data that involves creating multiple imputed data sets and estimating the parameters of interest for each imputed data set separately, before pooling the results using Rubin's rules to produce a final estimate that incorporates the uncertainty in the imputation process. The multiple imputation procedure can be generally summarized as following steps (Figure 2.1):

1. Identify the variables with missing data and the mechanism of missingness.

2. Create `m` (typically 1 imputation for every 1% missingness) imputed datasets by replacing the missing values with plausible values, based on a statistical model that reflects the relationship between the missing data and other observed data.

3. Fit the prediction model of interest to each imputed dataset separately. The model performance metrics and their standard errors are estimated separately for each imputed dataset.

4. Combine the results from the `m` imputed datasets using Rubin's rules.

Below we introduced various statistical approaches that can be considered in Step 2 above.

2.1.1 Multiple imputation using regression

The regression model used for MI can take various forms, such as linear regression, logistic regression, or generalized linear models, depending on the type of variable being imputed

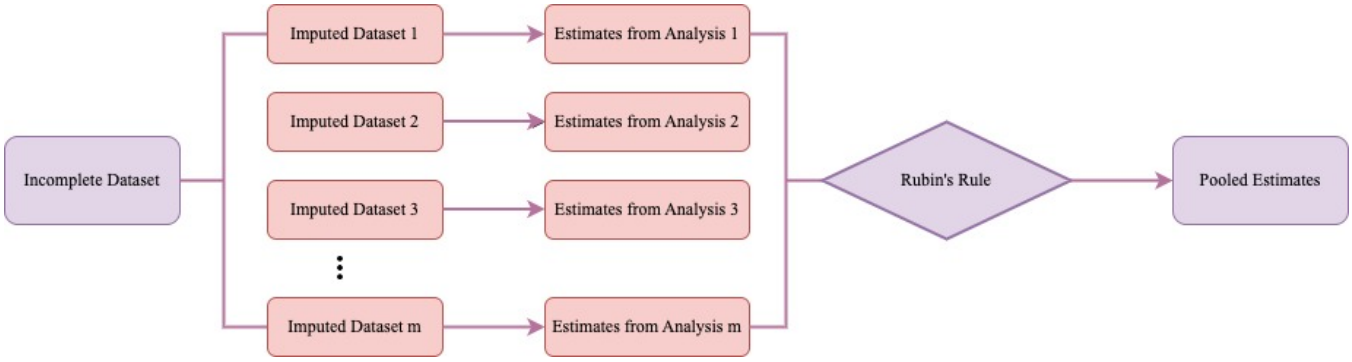


Figure 2.1: Procedure for Multiple Imputation

and the distributional assumptions of the data. The model may also include interactions, polynomial terms, or other nonlinear transformations to capture more complex relationships between the variables. The process involves fitting a regression model on each variable with missing data, using the other variables as predictors, and then generating multiple imputed data sets by randomly sampling plausible values from the fitted conditional distribution.

Multiple imputation using regression is typically more straightforward to implement and may be more appropriate when the relationships among variables are well understood and can be accurately specified. However, one main challenge in real-world practice is to determine which variables to include and which interactions to consider in the imputation model. Model misspecification may lead to biased coefficient estimates and impact downstream predictive performance of a risk prediction model.

There are several R packages available to implement multiple imputation using regression. For example, the `mice` package provides a comprehensive set of tools for implementing multiple imputation using various imputation models, including regression-based imputation.

2.1.2 Multiple imputation using random forest

Bagging, also known as bootstrap aggregation, is an ensemble technique in which we fit a decision tree on different bootstrap samples of the training dataset. One example of

bagging ensemble is the random forest model. Random forest (RF), as its name implies, consists of a large number of independent decision trees that operate as an ensemble. The algorithm involves randomly selecting a subset of the predictor variables at each split and fitting multiple decision trees to the data, and then combining the results from individual trees to obtain a final prediction. The RF algorithm is particularly useful for handling high-dimensional data and nonlinear relationships between variables. In the context of MI, the algorithm is used to predict the missing values of a specified variable based on the observed and imputed values of other variables as the predictors.

There are several existing R packages that could perform multiple imputation using random forests, including `missForest`, `mice`, and `CALIBERrfimpute`, etc.

2.1.3 Multiple imputation using gradient boosting

Boosting is an ensemble technique in which the decision trees are not built independently, but sequentially. This technique employs the logic in which the subsequent models learn from the mistakes of the previous trees. The predictors can be chosen from a range of models like decision trees, regressors, classifiers, etc. Because new models are learning from mistakes committed by previous trees, it takes less time/iterations to reach close to actual predictions than bagging. However, one has to carefully choose the stopping criteria to avoid overfitting on training data. Belonging to boosting algorithms, gradient boosting [18] is one of the machine learning algorithms for regression as well as classification, which provides robust, reliable, and sufficiently accurate prediction in engineering applications[33]. Gradient boosting machines combine the predictions from multiple sequential models to create a strong model. The algorithm iteratively trains weak models on the error residuals of the previous models, and then combines their predictions by weighting them according to their individual performance. The procedures of Friedman's algorithm (for regression) are as follows [21]:

1. Initialize a prediction function: $f_0(x) = \text{argmin}_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$

2. For m = 1 to M:

(a) For $i = 1, 2, \ldots, N$, compute the negative gradient:

$$r_{im} = -[\frac{\partial L(y_i, f(x_i)))}{f(x_i)}]_{f=f_{m-1}}$$

(b) Train a weak model on the residuals: fit a regression tree to the targets r_{im} giving termial regions $R_{jm}, j = 1, 2, \ldots, J_m$

(c) For $j = 1, 2, \ldots, J_m$, compute

$$\gamma_{jm} = \text{argmin}_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

(d) Then update the prediction function by adding the weighted predictions of the weak model to the current prediction function: $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$

3. The final prediction function is the sum of all the weak models' weighted predictions: $\hat{f}(x) = f_M(x)$

XGBoost [9], which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree machine learning library, extending the basic gradient boosting algorithm proposed by Friedman. It improves the performance of the original algorithm by incorporating several enhancements such as regularization, parallel processing, and efficient memory usage. XGBoost is also able to maintain complex relationships observed in the data, such as interactions and non-linear relationships. Deng and Lumley [13] proposed a scalable multiple imputation framework `mixgb`, which offers a scalable solution for imputing large datasets using XGBoost, bootstrapping, and predictive mean matching. `mixgb` is built under Fully Conditional Specification, where XGBoost imputation models are built for each incomplete variable. XGBoost is limited to the given data, so it often underestimates the variance of imputed data. In this case, Predictive Mean Matching (PMM) could be applied [31], which randomly selects one of the nearest observed data points to each imputed value and replaces an imputed value with an observed data point so that to ensure

that the introduced variance will have the same structure as the variance in the population. `mixgb` can automatically handle different types of variables. Categorical variables do not need to be encoded by users themselves. Users can also choose different settings regarding bootstrapping and the types of PMM to enhance imputation performance. In this study, we specified `bootstrap = FALSE` to avoid bootstrapping for MI and `pmm.type = "auto"` to allow imputations with PMM for numeric variables and without PMM for categorical variables.

2.1.4 Multiple imputation using principal component analysis

Given n observations of p random variables denoted as X $(n \times p)$, principal component analysis (PCA) consists of finding a matrix $\hat{X}$ with rank S which minimizes the least squares criterion $||\hat{X} - X||^2$ with $||\cdot||$ the Frobenius norm. The solution could be obtained by using the singular value decomposition (SVD) of the matrix X as UDV^T [21] where U is an $n \times S$ orthogonal matrix $(U^TU = I_p)$ whose columns u_j are called the left singular vectors; V is a $p \times S$ orthogonal matrix $(V^TV = I_p)$ with columns v_j called the right singular vectors; D is a $S \times S$ diagonal matrix, with diagonal elements $d_1 \geq d_2 \geq \ldots \geq d_s \geq 0$ known as singular values. The columns UD are called the principal components of X. The expression of the general term of X is given by:

$$x_{ij} = \tilde{x}_{ij} + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2) \tag{2.1}$$

$$= \sum_{s=1}^{S} \tilde{x}_{ij}^{(s)} + \varepsilon_{ij} \tag{2.2}$$

$$= \sum_{s=1}^{S} \sqrt{d_s} u_{is} v_{js} + \varepsilon_{ij} \tag{2.3}$$

where the estimate of σ^2 corresponds to dividing the sum of the squared residuals by the number of entries minus the number of independent model parameters [8]. Point estimate would be $\hat{X}_{ij}^{PCA} = \sum_{s=1}^{S} \sqrt{d_s} u_{is} v_{js}$. Instead of the classical PCA estimator, us-

ing a regularized PCA is more suggested [34]. The regularized PCA solution would be $\hat{X}_{ij}^{rPCA} = \sum_{s=1}^{S} \hat{\Phi}_s \sqrt{d_s} u_{is} v_{js}$, where $\hat{\Phi}_s = \frac{d_s - \frac{np}{\min(n-1,p)}\hat{\sigma}^2}{d_s}$, for all s from 1 to S.

Josse and Husson [27] proposed a method of single imputation based on a PCA model. Audigier et al. [3] extended it to multiple imputation and incorporated the uncertainty in estimating parameters of the PCA imputation model using a Bayesian approach. In their method, they included a data augmentation algorithm which can be straightforwardly used to get multiple imputed data sets as described below. Given $\hat{X}$ and the estimate of σ^2, one can impute the missing values x_{ij} by a draw from the predictive distribution $N(\hat{X}_{ij}^{rPCA}, \frac{\hat{\sigma}^2 \sum_{s=1}^{S} \Phi_s}{\min(n-1,p)})$.

There are several functions in the `missMDA` package that can perform MI. Specifically, `MIPCA` can be used in continuous variables, `MIMCA` is used to deal with categorical variables while `MIFAMD` is useful for both continuous and categorical variables. If there are only continuous variables, `MIPCA` is used to generate `nboot` imputed datasets from a PCA model. The observed values remain the same from one imputed data set to another while the imputed values change. By default, `MIPCA` uses the parametric bootstrap (`method.mi="Boot"`). To perform MI using a Bayesian treatment as described above, we may specify the argument `method.mi="Bayes"`.

2.2 Multiple imputation of interaction and non-linear terms

Practically, additional knowledge about derived data (e.g. interaction or non-linear terms) is often present but not explicitly modeled. If such relationships are not specified, the imputation model may produce inconsistent imputations. There are several approaches to imputing interaction and non-linear terms in the imputation model.

2.2.1 Just Another Variable

The Just Another Variable (JAV) approach [52] or called "Transform, then impute" is a method used in statistical analysis to impute non-linear terms of a predictor or interaction terms between two or more variables. In this approach, the higher-order term is created

by adding a new variable to the model that represents the non-linear transformation of a variable or the product of the original predictor variables. This new variable is then treated as any other predictor variable in the model. Simulation studies have shown that under the assumption of multivariate normality, the unconstrained JAV approach has good coverage properties for the case when the interaction is the product of two continuous variables [49].

2.2.2 Passive imputation

Passive imputation is another approach in handling higher-order terms during imputation. The method involves first imputing the original variable(s) and then deriving the higher-order term directly from the corresponding variable(s). This approach preserves the relationship between the original variable(s) and the higher-order terms.

2.2.3 Substantive model compatible fully conditional specification

The Substantive Model Compatible Fully Conditional Specification (smcfcs) method [5] is a multiple imputation method that ensures that each partially observed variable is imputed from an imputation model which is compatible with a user-specified model for the outcome (which is typically the substantive model of interest). The smcfcs involves specifying a substantive model that includes the interaction terms of interest, along with any other relevant variables and covariates.

2.3 Performance evaluation

Risk models built and evaluated based on the four MI approaches were first compared based on the measure of calibration. Calibration refers to the agreement between observed outcomes and predictions [22] and evaluation of calibration is important if model predictions are used to inform patients or physicians to make decisions [44]. We assessed calibration using:

1. Expected-to-observed event ratio (E/O ratio): it is a metric that compares the number of expected events with the number of observed events. A perfect model would have an

E/O ratio of 1, indicating that the model's predictions match the observed outcomes exactly. A well-calibrated model has the 95% confidence interval for the E/O ratio overlapping 1.

2. Cox calibration intercept: it measures the extent to which the model over- or underestimates the probability of an event occurring. A perfect model would have an intercept of 0, indicating that the model's predictions are unbiased. A well-calibrated model has the 95% confidence interval for the calibration intercept overlapping 0.

3. Cox calibration slope: it measures the relationship between the predicted probabilities and the actual probabilities of events occurring. A slope of 1 indicates a perfect calibration, while values less than or greater than 1 indicate over- or underfitting, respectively. A well-calibrated model has the 95% confidence interval for the calibration slope overlapping 1.

Second, the discriminatory accuracy of the risk models built and evaluated based on the four MI approaches was compared via the area under the receiver operating characteristic curve (AUC). The standard error for each performance metric was also reported.

Chapter 3

SIMULATION STUDIES

We conducted data simulations consisting of four different steps. Firstly, complete data sets were generated based on different scenarios (low/high dimensional, rates of the outcome, numbers of observations, etc.). Then, we applied several amputation mechanisms to induce missing values in the complete data sets. Thirdly, different imputation methods described in the Method section were utilized to fill in the missing values of the simulated incomplete data sets. Finally, we built prediction models using the imputed data sets and compared the predictive performance of these risk prediction models. We repeated the above process 100 times.

3.1 Data generation

We created simulated data sets with different rates of outcome to compare the performance of methods. We generated 100 simulated data sets of N observations for each scenario where the dependent variable Y was binary.

3.1.1 Low dimensional data sets

The low dimensional datasets consist of 6 mix-type predictors. Independent variables were generated from each of the six distributions (Figure 3.1):

1. $X_1 \sim$ Normal(2.5, 1.2)

2. $X_2 \sim$ Normal(5.7, 2.5)

3. $X_3 \sim$ Gamma(2.8, 3.5)

4. $X_4 \sim$ Normal(2.2, 1.5)

5. $X_5 \sim$ Categorical (3, 0.8, 0.1, 0.1)

6. $X_6 \sim$ Binom(1, 0.9)

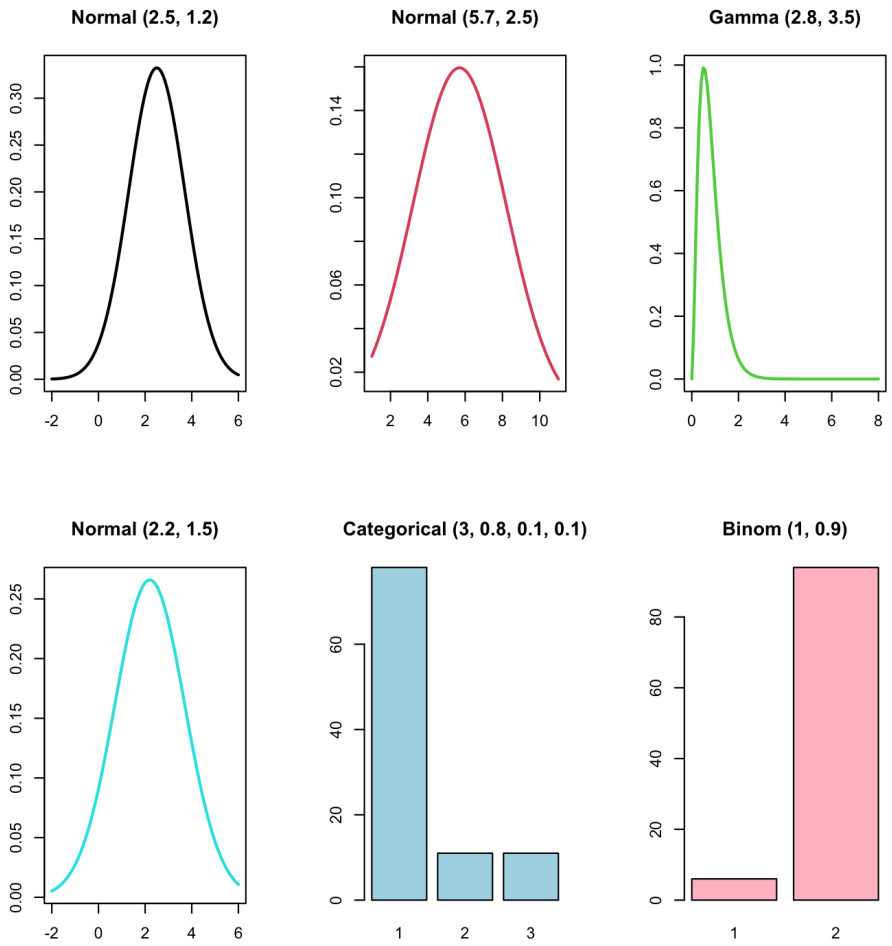


Figure 3.1: Distributions for simulated independent variables

We then generate the binary outcome variable using a logistic regression model as below.

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{5.2}\mathbb{1}(X_5 = 2) + \beta_{5.3}\mathbb{1}(X_5 = 3) + \beta_{6.2}\mathbb{1}(X_6 = 2) \tag{3.1}$$

We tried different values of β_0 and the number of observations (N) for different rates of the outcome. The clinical setting is often characterized by rare events, which can make it difficult but valuable to predict and simulate such events. We simulated data sets with rates of the outcome as 10% as well as the more rare case, 6%, to mimic the real clinical settings. We set $\beta_0 = -5.10$ and $N = 9000$ so that the probability of $Y = 1$ is $p = 0.06$; we also set $\beta_0 = -4.40$ and $N = 7500$ so that the probability of $Y = 1$ is $p = 0.10$. We always set the other coefficients as $\beta_1 = 0.23, \beta_2 = 0.09, \beta_3 = 0.15, \beta_4 = 0.45, \beta_{5.2} = 0.20, \beta_{5.3} = 0.77, \beta_{6.2} = 0.07$.

3.1.2 Low dimensional data sets with interactions

To assess the impact of model misspecification in MI, we also created 100 simulated data sets to compare the performance of methods when there were interaction terms between predictors in their associations to the outcome. Using the same distributions of independent variables above, we built a logistic regression model:

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{5.2}\mathbb{1}(X_5 = 2) + \beta_{5.3}\mathbb{1}(X_5 = 3) + \beta_{6.2}\mathbb{1}(X_6 = 2) \tag{3.2}$$

$$+ \beta_7 X_7, \text{ where } X_7 = X_2 X_3 \tag{3.3}$$

We set $\beta_0 = -6.40$ and $N = 5000$ so that the probability of $Y = 1$ is $p = 0.06$; we also set $\beta_0 = -5.70$ and $N = 7500$ so that the probability of $Y = 1$ is $p = 0.10$.

3.1.3 High dimensional data set

In general, clinical prediction models may have anywhere from a few predictors to dozens or even hundreds of predictors. Some models may only include a few important variables

that have been shown to be strongly associated with the outcome of interest, while other models may incorporate a larger number of variables to capture more nuanced relationships and potential confounding factors. In the previous scenarios, we only included six predictors in simulated dataset, which may not reflect real-world data in a clinical setting. To present a more challenging scenario, we also simulated a relatively high-dimensional data set with 64 predictors. We used the R package `pensim` [19, 50] to simulate predictors with a specified correlation structure.

We simulated 64 variables forming four different groups. Variables from different groups are independent. The first 20 variables (group "a") are uncorrelated, and the binary outcome is associated with only the first variable in the group "a" with a coefficient of 0.5. The next 16 variables (group "b") have a correlation of 0.7 to each other variable in that group, and the binary outcome is associated with only the first variable in the group "b" with a coefficient of 0.6. The next 8 variables (group "c") have a correlation of 0.7 to each other variable in that group, and the binary outcome is associated with the first variable in the group "c" with a coefficient of 0.3; The final 20 variables (group "d") are uncorrelated with each other but are all associated with the binary outcome with a coefficient of 0.2. Binary outcomes for $N = 3000$ samples are simulated as a Bernoulli distribution. We set the intercept as -2.40 such that the probability of $Y = 1$ (p) is roughly 0.10.

3.2 Data amputation

MI can handle both MCAR and MAR. In our simulation studies, we induced missing values in our data sets based on the MAR assumption.

Specifically, for the low dimensional data sets, we set the probability of X_2, X_3 being missing to 30%, while we also introduced 18% missing observations in X_4, X_5, X_6 (Figure 3.2). For the high dimensional data set, we induced 10% missing observations in each of the first 10 variables $(X_1, X_2, \ldots, X_{10})$, 20% missing observations in the combination of $X_{30}, X_{31}, \ldots, X_{45}$. For the low dimensional data sets with interactions, we did not generate missing values for the interaction term X_7 on purpose, but it would also have missing obser-

Table 3.1: The specific relationships among the variables in each scenario

Scenarios	Models	Rates of outcome (p)	Numbers of observations
Low dimensional	$\text{logit}(p) = -5.10 + 0.23X_1 + 0.09X_2 - 0.15X_3 + 0.45X_4 + 0.20\mathbb{1}(X_5 = 2) + 0.77\mathbb{1}(X_5 = 3) + 0.07\mathbb{1}(X_6 = 2)$	0.06	9000
Low dimensional	$\text{logit}(p) = -4.40 + 0.23X_1 + 0.09X_2 - 0.15X_3 + 0.45X_4 + 0.20\mathbb{1}(X_5 = 2) + 0.77\mathbb{1}(X_5 = 3) + 0.07\mathbb{1}(X_6 = 2)$	0.10	7500
Low dimensional with interactions	$\text{logit}(p) = -6.40 + 0.23X_1 + 0.09X_2 - 0.15X_3 + 0.45X_4 + 0.20\mathbb{1}(X_5 = 2) + 0.77\mathbb{1}(X_5 = 3) + 0.07\mathbb{1}(X_6 = 2) + 0.22X_7$, where $X_7 = X_2X_3$	0.06	5000
Low dimensional with interactions	$\text{logit}(p) = -5.70 + 0.23X_1 + 0.09X_2 - 0.15X_3 + 0.45X_4 + 0.20\mathbb{1}(X_5 = 2) + 0.77\mathbb{1}(X_5 = 3) + 0.07\mathbb{1}(X_6 = 2) + 0.22X_7$, where $X_7 = X_2X_3$	0.10	7500
High dimensional*	$\text{logit}(p) = -2.40 + a_1X_1 + a_2X_2 + \ldots + a_{20}X_{20} + b_1X_{21} + b_2X_{22} + \ldots + b_{16}X_{36} + c_1X_{37} + c_2X_{38} + \ldots + c_8X_{44} + d_1X_{45} + d_2X_{46} + \ldots + d_{20}X_{64}$	0.10	3000

* High dimensional simulated data set includes four groups of independent variables (group "a" - "d"). We specified correlation between variables within each group as well as associations with the outcome variable.

vations because of missing in X_2 and X_3. For each amputated data set, we randomly split the data into 80% and 20% for training and testing sets.

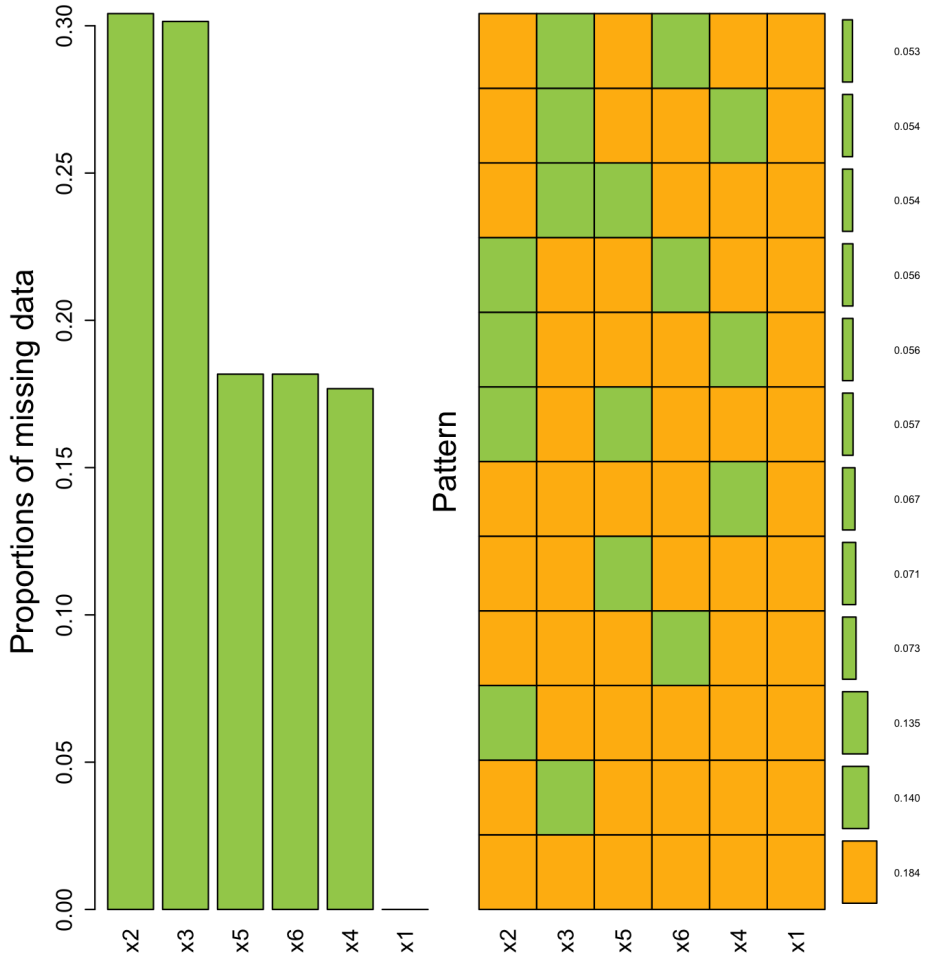


Figure 3.2: Proportions and patterns of missing data in low dimensional data sets with the rate of outcome = 0.1

3.3 Data imputation

For each amputated data set, the missing values were imputed 5 times in the training and test sets separately using four different methods: Parametric MI ($\mathbf{MI_{para}}$), MI using random for-

est ($\mathbf{MI_{RF}}$), MI using gradient boosting ($\mathbf{MI_{GB}}$) and MI using principal component analysis ($\mathbf{MI_{PCA}}$). We used R package `mice` to conduct $\mathbf{MI_{para}}$ and $\mathbf{MI_{RF}}$ (For $\mathbf{MI_{RF}}$, we specified `meth = "rf"`). We used R package `mixgb` to conduct $\mathbf{MI_{GB}}$. We firstly applied function `mixgb_cv` to tune the number of boosting rounds `nrounds`, and then applied the function `mixgb` with the optimal `nrounds`. We used R package `missMDA` to conduct $\mathbf{MI_{PCA}}$. We used `MIPCA` when there were only continuous variables in the imputation model and `MIFAMD` when there were both continuous and categorical variables. We applied `estim_ncpPCA` or `estim_ncpFAMD` to tune the number of components `ncp` by cross-validation, and then used the optimal `ncp` to perform MI. For each method, to explore the impacts of incorporating the outcome variable during MI, we imputed the missing values in three different scenarios: (1) included outcome variable in both training and test sets during MI (**with_y**); (2) included outcome variable in training set only during MI (**hybrid**); (3) omitted outcome variable in both training and test sets during MI (**no_y**). Moreover, for $\mathbf{MI_{para}}$, we also compared the three different approaches (JAV, passive imputation and smcfcs) in imputing the interaction terms. In addition, we aimed to investigate the impact of model misspecification for MI using regression: (1) if the original complete data set did not have any interaction between the covariables, whether imputing unnecessary interaction terms would affect the performance of the prediction model; **AND** (2) if the original complete data set had an interaction between the predictors, whether not imputing the known interaction terms would affect the performance of the prediction model. Specifically, smcfcs was only conducted for **with_y** since the outcome variable is required to be specified during MI.

Overall, we conducted a two-layer comparison for the MI using regression and a single-layer comparison for each of the machine learning-based methods (See Figure 3.3 and 3.4).

3.4 Model fitting

We conducted predictive modeling on each of the imputed data sets and stored models. For the low dimensional data sets, Logistic Regression was applied to each imputed data set by regressing the binary outcome variable on the 6 predictors. For the high dimensional

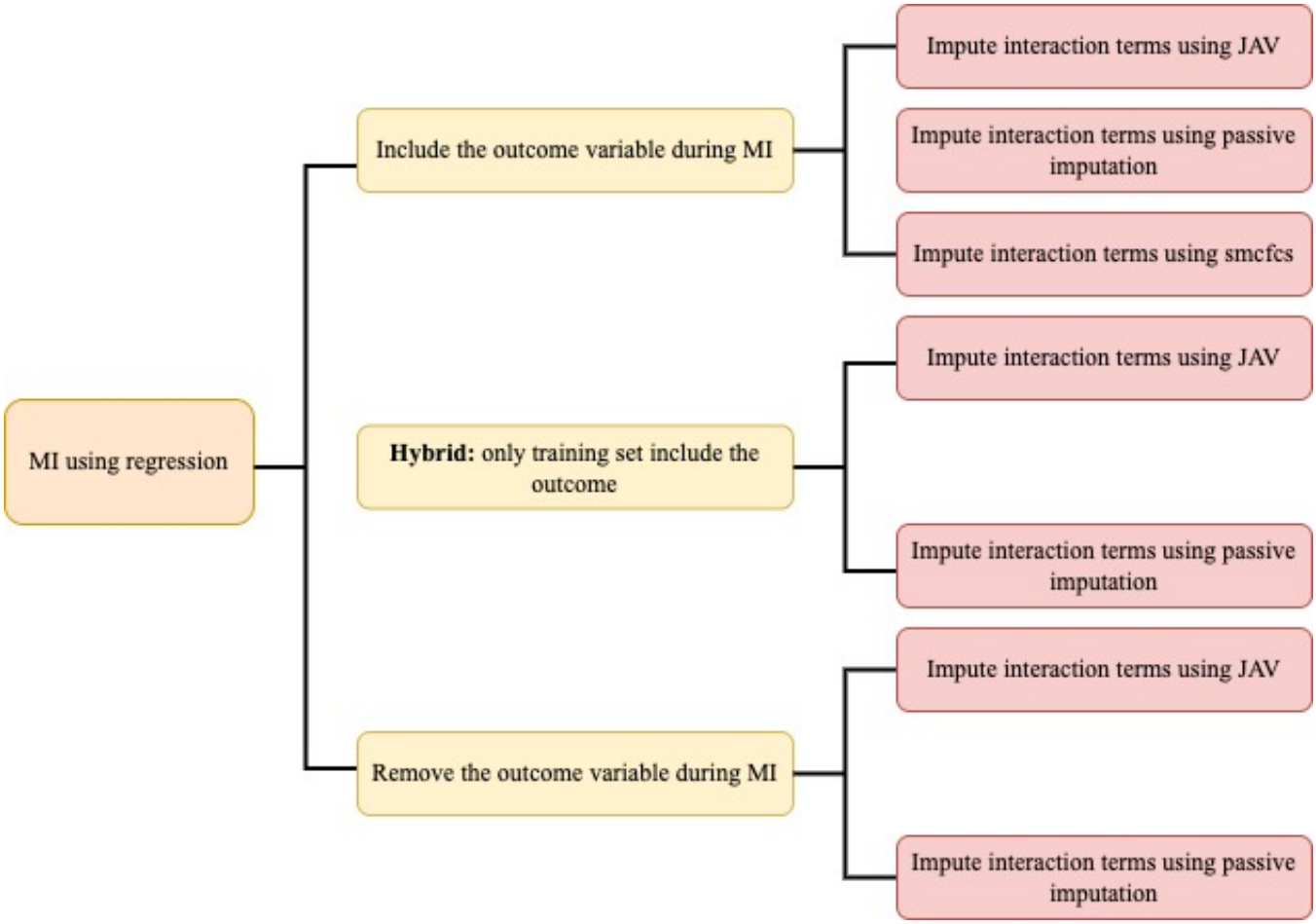


Figure 3.3: Two-layer comparison within the method of MI using regression

data sets, Ridge Regression and Random Forest were performed and compared instead. Specifically, we fitted the models on the training sets and evaluated their performance (first calibration and then discrimination) on the test sets. To derive the final values of evaluation metrics, we generated these metrics across all 5 imputations and then combined them using Rubin's rule. We constructed 95% confidence intervals (CIs) in the testing sets via 100 simulations.

For Ridge Regression, we used the function `cv.glmnet` in the R package `glmnet` to perform 5-fold cross-validation for tuning parameter selection. For Random Forest, we utilized

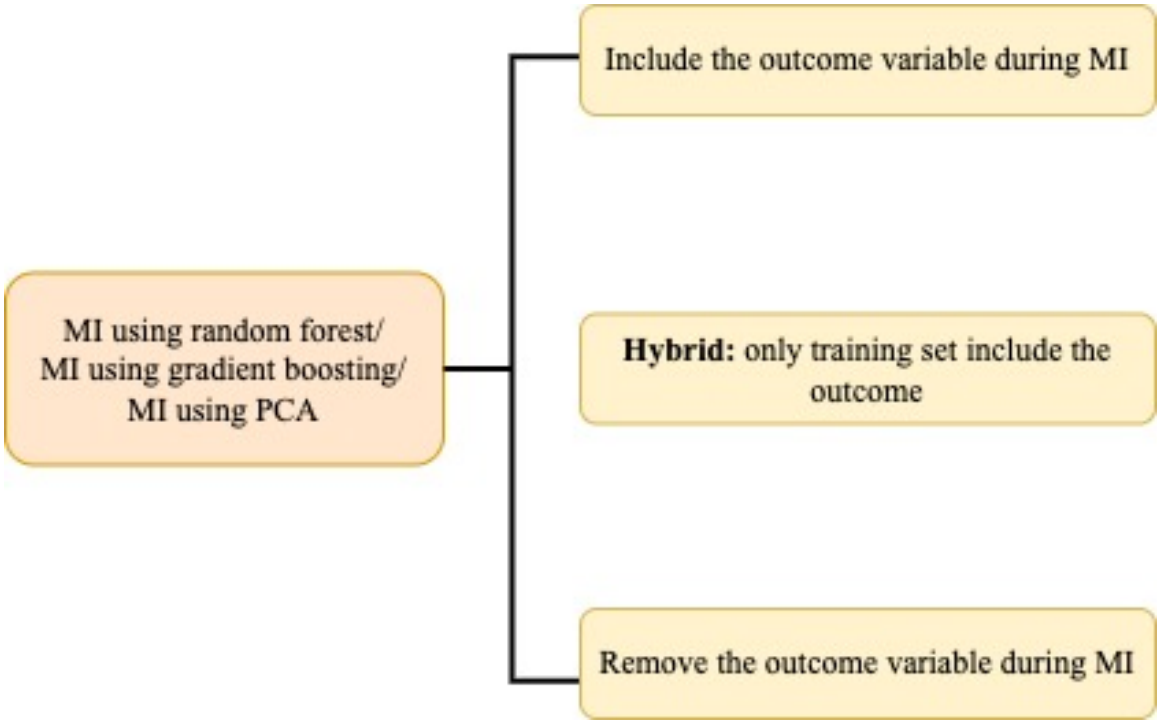


Figure 3.4: Single-layer comparison within each of the machine-learning methods

the `caret` package to help us select two tuning parameters: `mtry` (the number of variables to randomly sampled at each split) and `ntree` (the number of trees to build in the Random Forest model). For `mtry`, we tried $0.5\sqrt{p}$, $\sqrt{p}$, $2\sqrt{p}$ where p refers to the number of predictors in the model; for `ntree`, we considered 500, 1000 and 1500. A 5-fold CV was used for tuning parameter selection.

3.5 Results

3.5.1 Calibration

For low dimensional setting where the rate of outcome was 6%, all MI approaches showed that the 95% CI for E/O ratios and calibration intercepts overlapped 1 and 0, respectively (Figure 3.5a, 3.6a), indicating no systematic biases. Moreover, all MI approaches under the **hybrid** scenario had calibration slopes (Figure 3.7a) significantly less than 1 ($\mathbf{MI_{para}}$: 0.79

(95% CI: [0.77, 0.81]); $\mathbf{MI_{RF}}$: 0.84 (95% CI: [0.82, 0.86]); $\mathbf{MI_{GB}}$: 0.81 (95% CI: [0.79, 0.83]); $\mathbf{MI_{PCA}}$: 0.78 (95% CI: [0.76, 0.80])), suggesting overfitting. All MI approaches (either under **with_y** scenario or **no_y** scenario) showed calibration slopes close to 1.

We got similar results for low dimensional setting where the rate of outcome was 10%. All MI approaches showed that the 95% CI for E/O ratios and calibration intercepts overlapped 1 and 0 respectively (Figure 3.5b, 3.6b), indicating no systematic biases. All MI approaches under the **hybrid** scenario had calibration slopes (Figure 3.7b) significantly less than 1 ($\mathbf{MI_{para}}$: 0.78 (95% CI: [0.76, 0.80]); $\mathbf{MI_{RF}}$: 0.83 (95% CI: [0.81, 0.85]); $\mathbf{MI_{GB}}$: 0.80 (95% CI: [0.78, 0.82]); $\mathbf{MI_{PCA}}$: 0.78 (95% CI: [0.76, 0.80])), suggesting overfitting. All MI approaches (either under **with_y** scenario or **no_y** scenario) showed calibration slopes close to 1.

For low dimensional setting with interactions where the rate of outcome was 6%, all MI approaches showed that the 95% CI for E/O ratios and calibration intercepts overlapped 1 and 0 respectively (Figure 3.5c, 3.6c), indicating no systematic biases. All MI approaches under the **hybrid** scenario had calibration slopes (Figure 3.7c) significantly less than 1 ($\mathbf{MI_{para}}$: 0.66 (95% CI: [0.64, 0.68]); $\mathbf{MI_{RF}}$: 0.72 (95% CI: [0.70, 0.74]); $\mathbf{MI_{GB}}$: 0.71 (95% CI: [0.69, 0.73]); $\mathbf{MI_{PCA}}$: 0.61 (95% CI: [0.59, 0.63])), suggesting overfitting. All MI approaches (either under **with_y** scenario or **no_y** scenario) showed calibration slopes close to 1.

For low dimensional setting with interactions where the rate of outcome was 10%, all MI approaches under the **hybrid** scenario showed that the 95% CI for E/O ratios and calibration intercepts overlapped 1 and 0 respectively (Figure 3.5d, 3.6d), indicating no systematic biases. All MI approaches under the **hybrid** scenario had calibration slopes (Figure 3.7d) significantly less than 1 ($\mathbf{MI_{para}}$: 0.67 (95% CI: [0.65, 0.69]); $\mathbf{MI_{RF}}$: 0.73 (95% CI: [0.71, 0.75]); $\mathbf{MI_{GB}}$: 0.70 (95% CI: [0.68, 0.72]); $\mathbf{MI_{PCA}}$: 0.66 (95% CI: [0.64, 0.68])), suggesting overfitting. All MI approaches (either under **with_y** scenario or **no_y** scenario) showed calibration slopes close to 1.

For high dimensional setting when the fitted model was ridge regression, all MI approaches

demonstrated that the 95% CI for E/O ratios and calibration intercepts overlapped 1 and 0 respectively (Figure 3.5e, 3.6e), indicating no systematic biases. All approaches had calibration slopes (Figure 3.7e) overlapping 1, suggesting no under/overfitting. When using Random Forest to fit the risk prediction model (Figure 3.5f, 3.6f, 3.7f), all MI approaches had E/O ratios significantly greater than 1 and calibration intercepts significantly less than 1, indicating overestimation; all MI approaches showed overfitting as their calibration slopes were significantly less than 1 (range: 0.86 - 0.90).

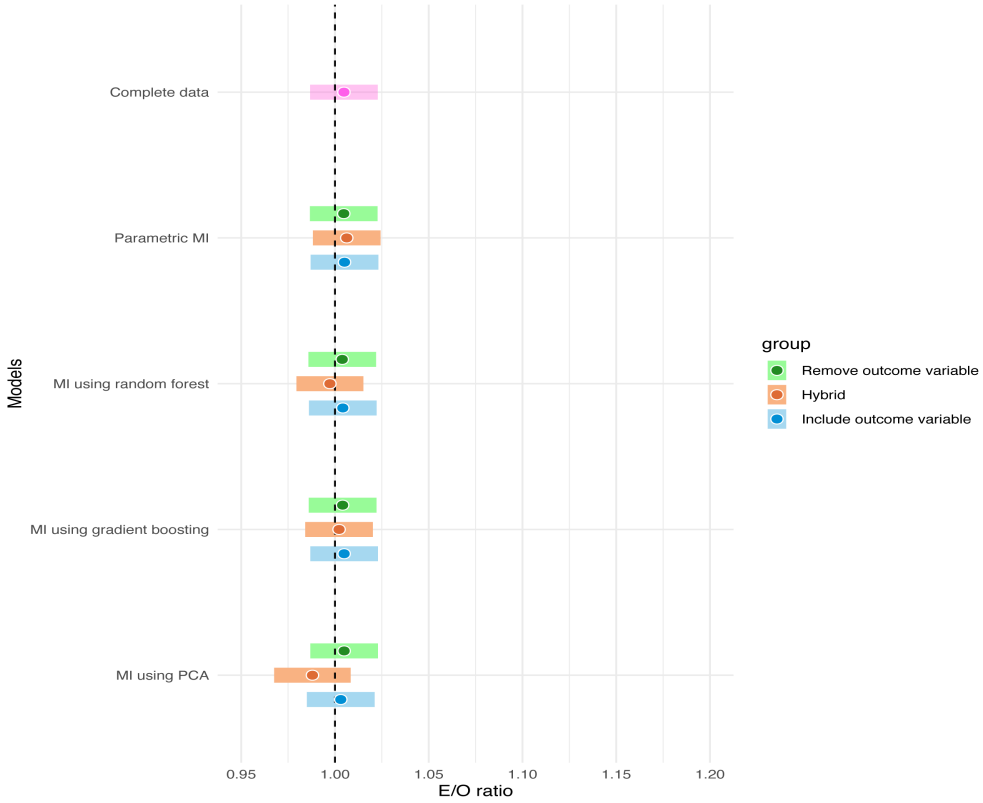
3.5.2 Discrimination

For low dimensional settings, all MI approaches had similar AUCs under **no_y** scenarios and **hybrid** scenarios, except for $\mathbf{MI_{PCA}}$. $\mathbf{MI_{PCA}}$ under the **hybrid** scenario had slightly lower AUCs than the **no_y** scenario. However, approaches under **with_y** scenarios led to higher AUCs. For the $\mathbf{MI_{PCA}}$, the AUCs under the **with_y** scenario were even substantially higher than that using complete dataset. For example, for low dimensional setting where the rate of outcome was 6%, the AUC (Figure 3.8a) of model fitted using complete data was 0.71 (95% CI: [0.70 - 0.72]), whereas the AUC using $\mathbf{MI_{PCA}}$ approach under the **with_y** scenario was 0.73 (95% CI: [0.72 - 0.74]).

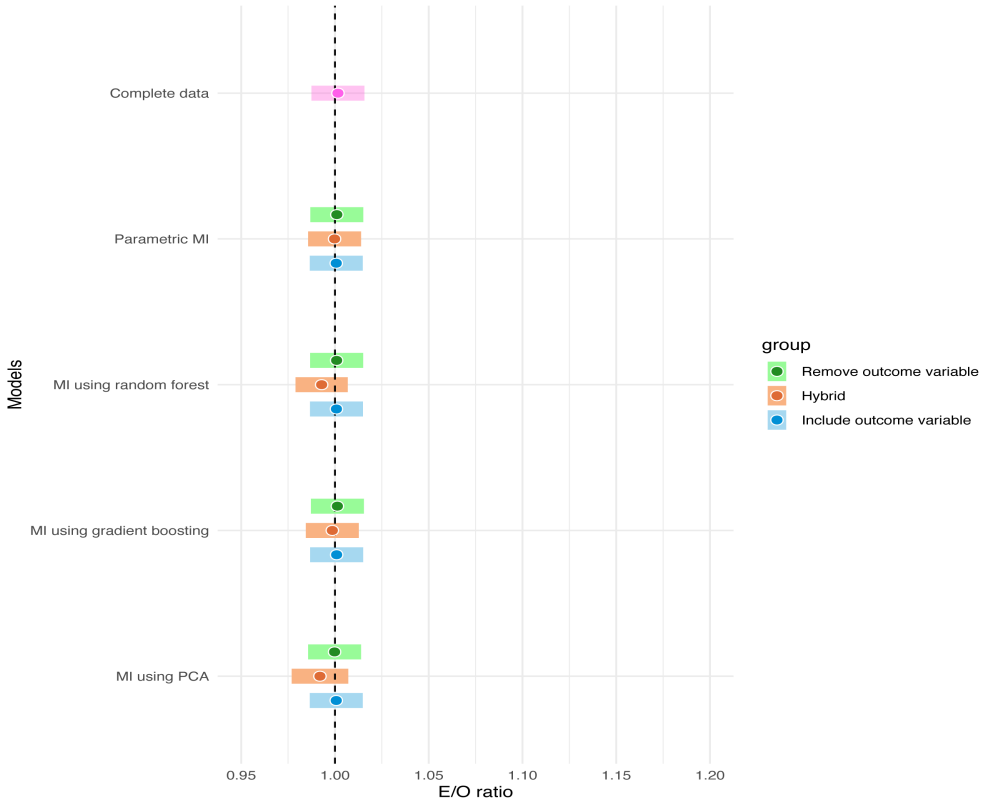
For high dimensional settings (Figure 3.8e, 3.8f), the model fitted using the Ridge Regression had overall higher AUCs than the model fitted using the Random Forest. $\mathbf{MI_{para}}$, $\mathbf{MI_{RF}}$ and $\mathbf{MI_{GB}}$ had higher AUCs under **with_y** scenarios than **no_y** scenarios; while the AUCs for the **hybrid** scenario were between them. For the $\mathbf{MI_{PCA}}$ approach, we could not observe substantial differences between those three scenarios (**no_y**, **hybrid** or **with_y**).

3.5.3 Impact of imputing interaction terms

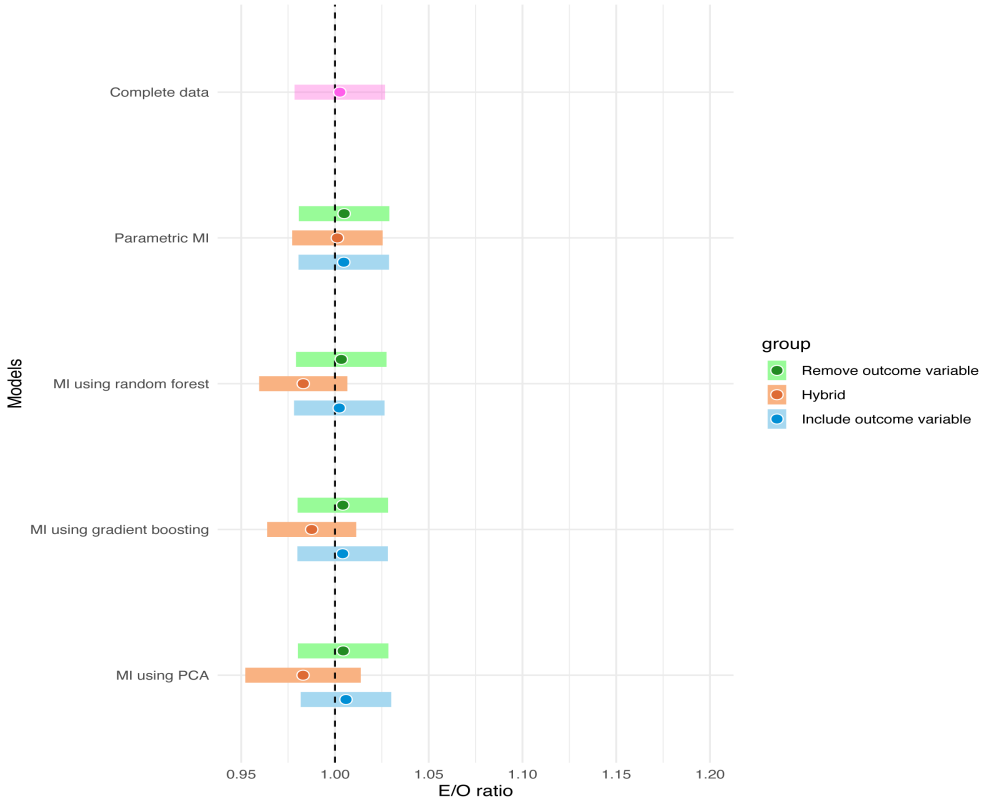
For low dimensional settings without interaction terms in the underlying model, we didn't observe meaningful differences in model performance between imputing interaction terms during MI and not imputing interaction terms during MI for all scenarios (**with_y**, **hybrid**



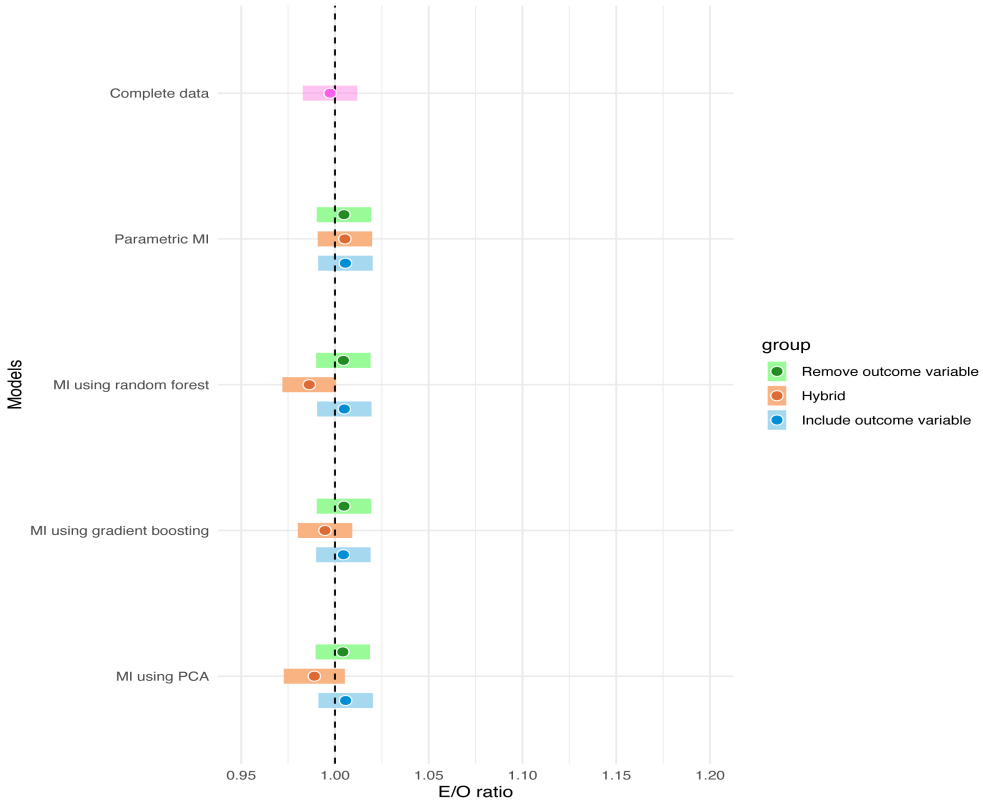
(a) Low dimensional (Rate of outcome: 6%)



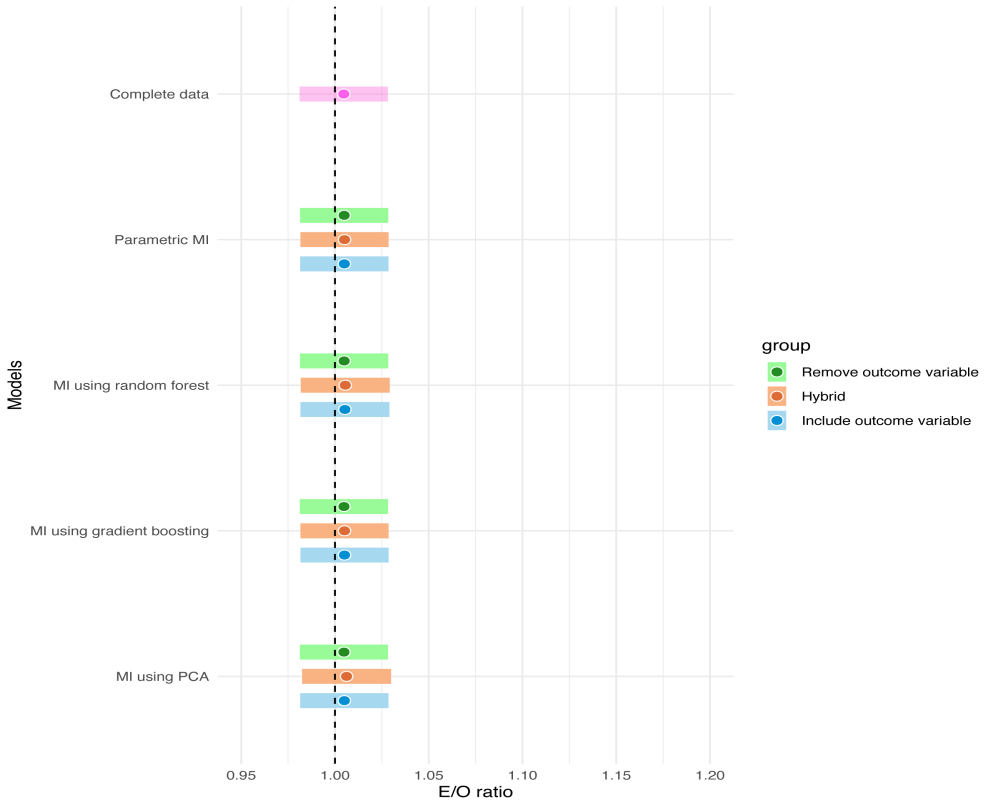
(b) Low dimensional (Rate of outcome: 10%)



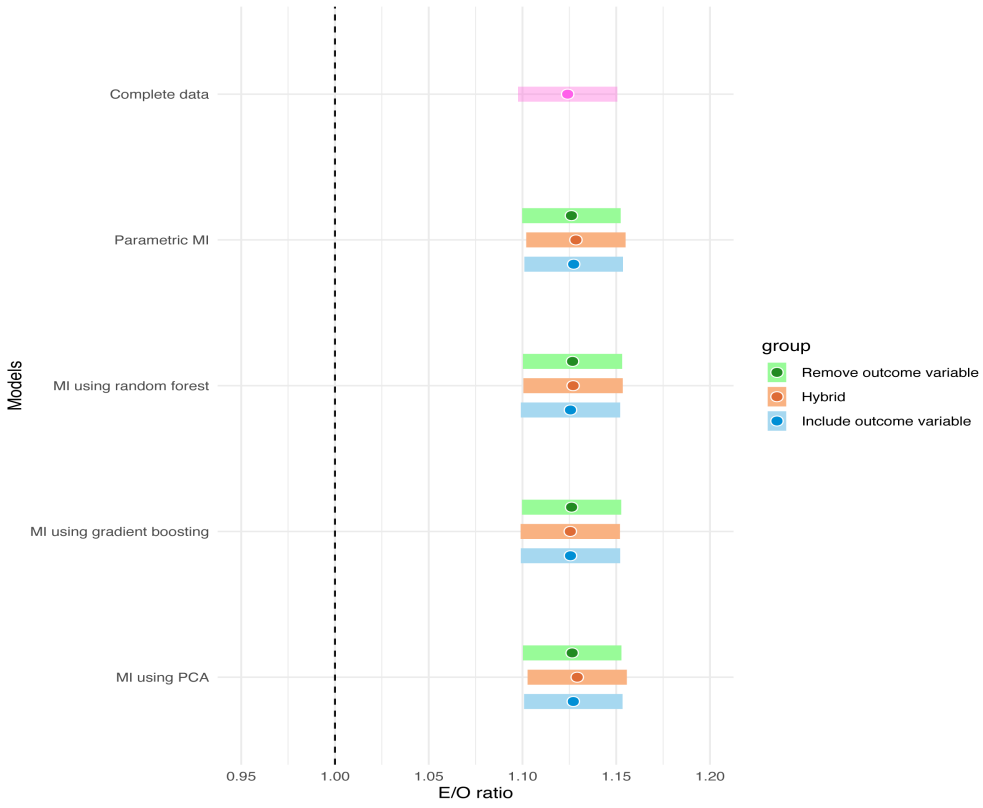
(c) Low dimensional with interactions (Rate of outcome: 6%)



(d) Low dimensional with interactions (Rate of outcome: 10%)

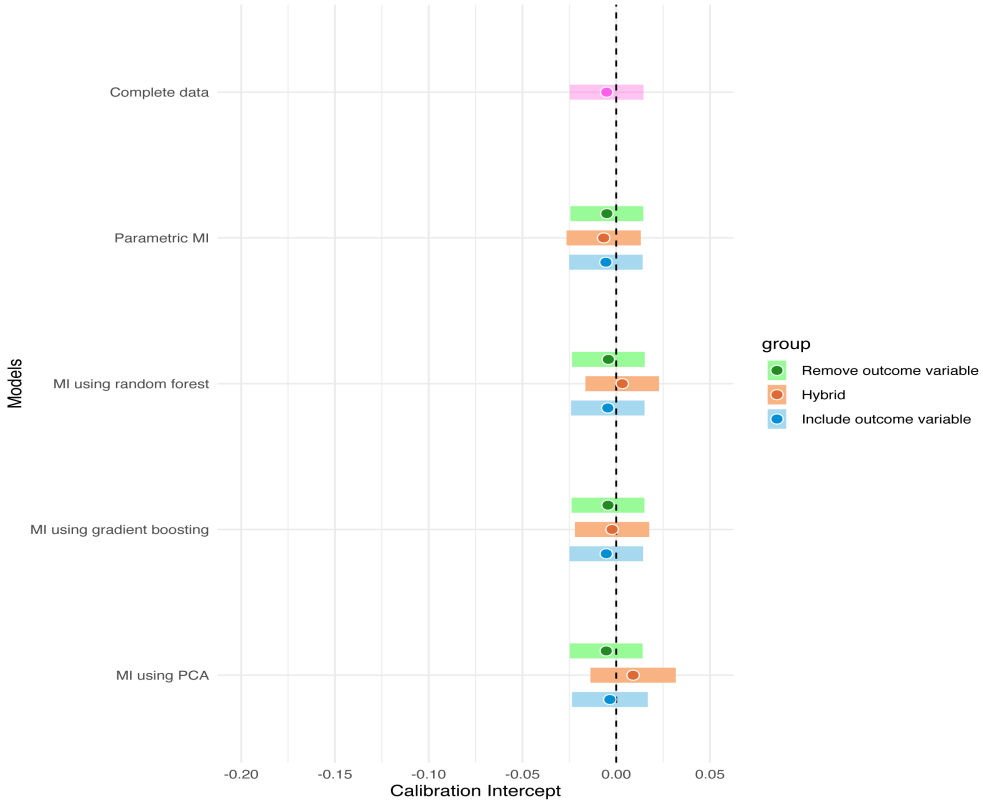


(e) High dimensional fitted with ridge regression (Rate of outcome: 10%)

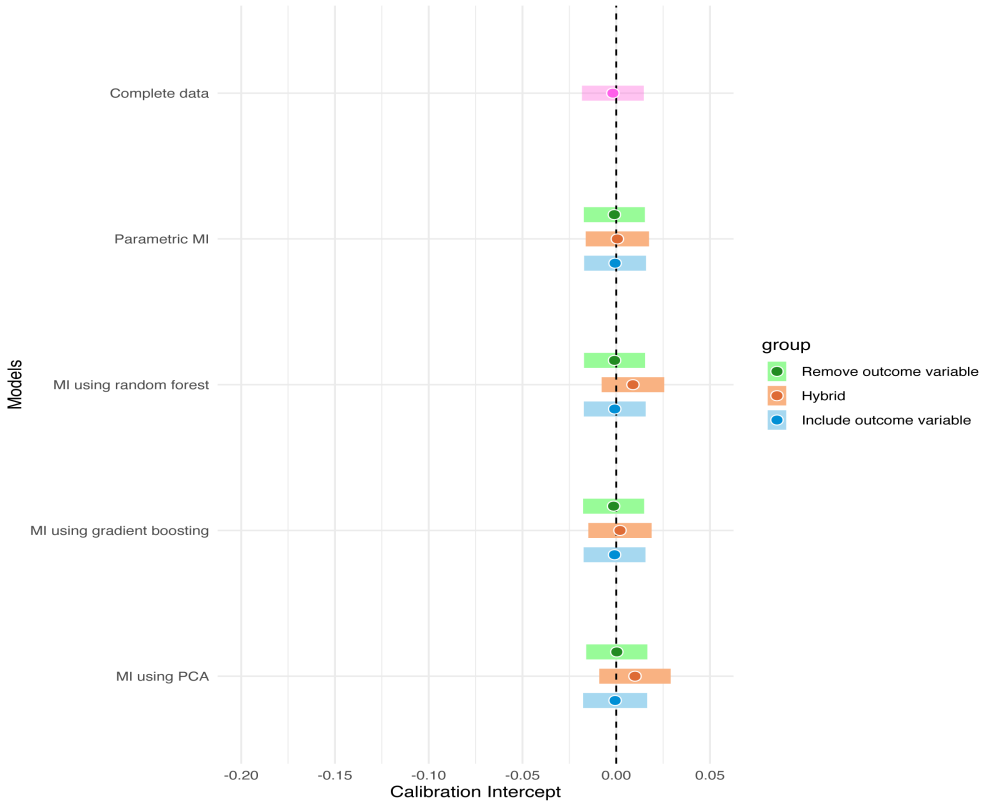


(f) High dimensional fitted with Random Forest (Rate of outcome: 10%)

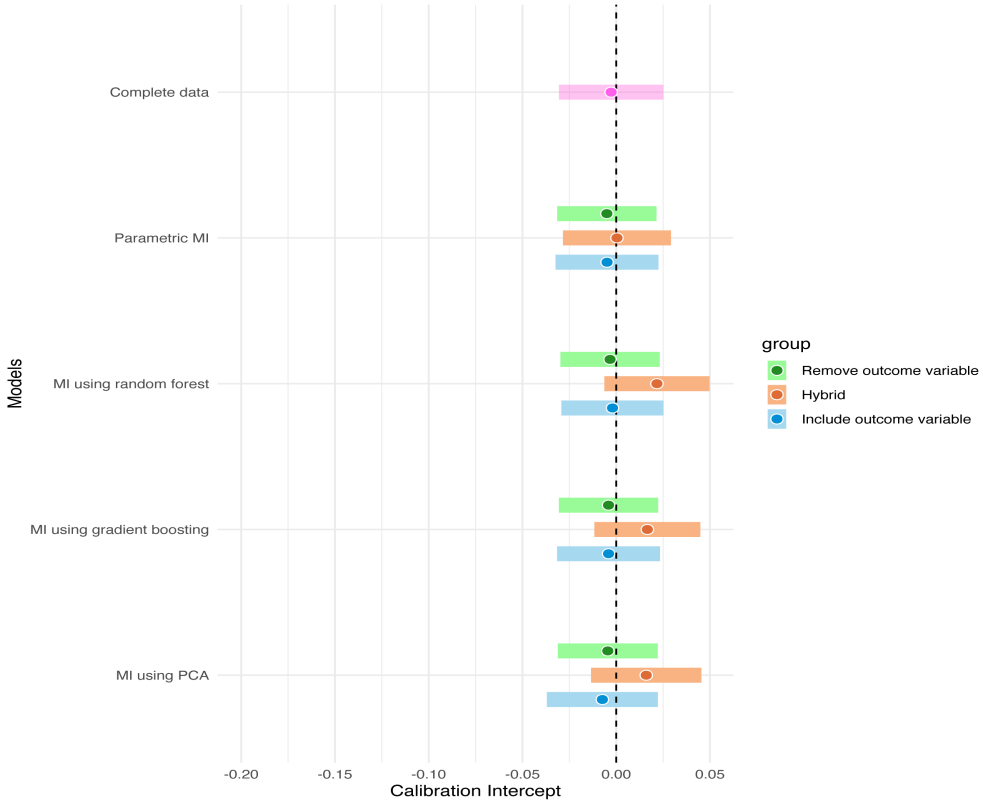
Figure 3.5: Expected to observed events ratio (E/O ratio) as well as the corresponding 95% CI for fitted models under different simulation scenarios. The dashed vertical lines showed the ideal value for E/O ratio (1).



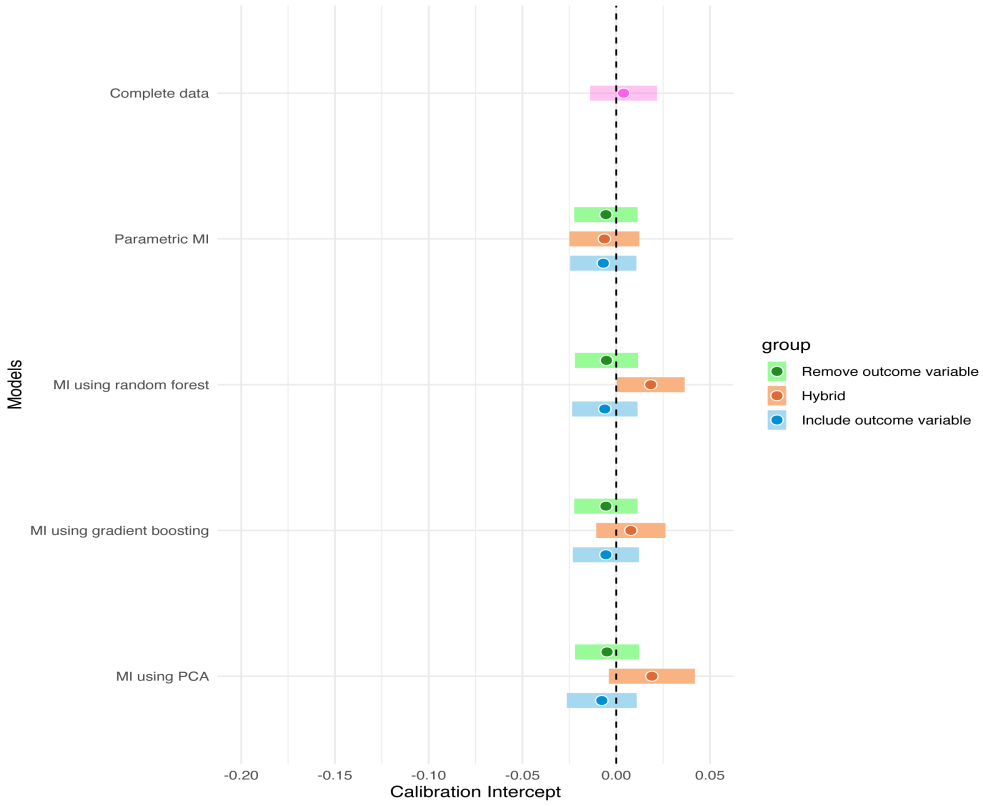
(a) Low dimensional (Rate of outcome: 6%)



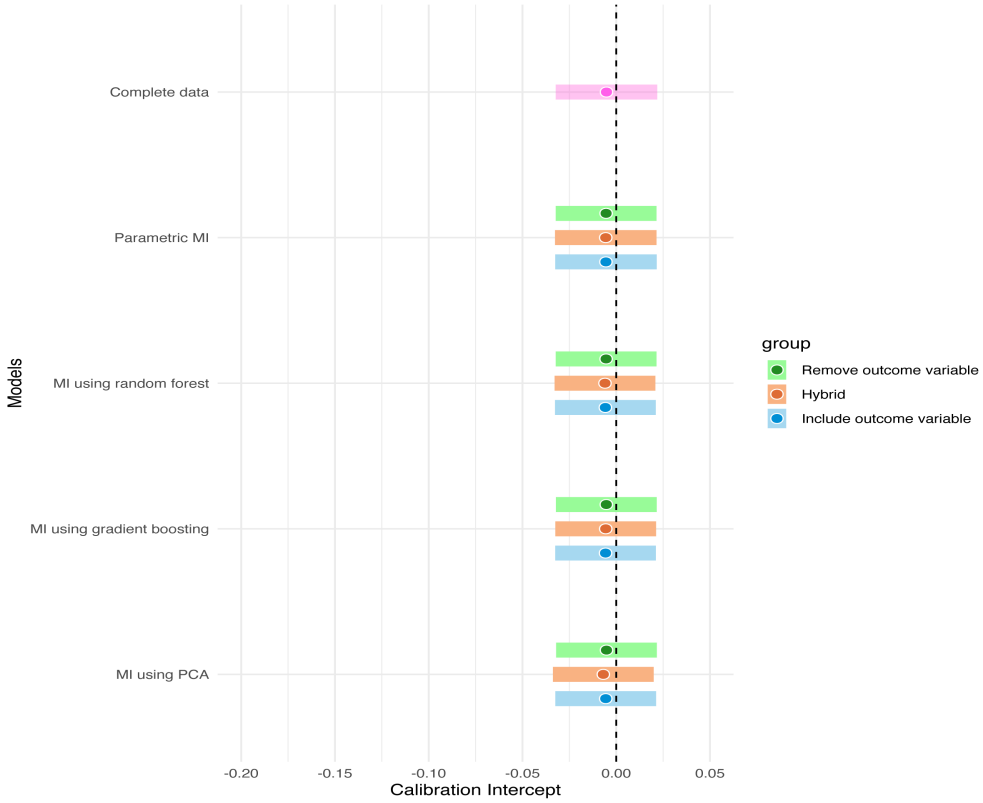
(b) Low dimensional (Rate of outcome: 10%)



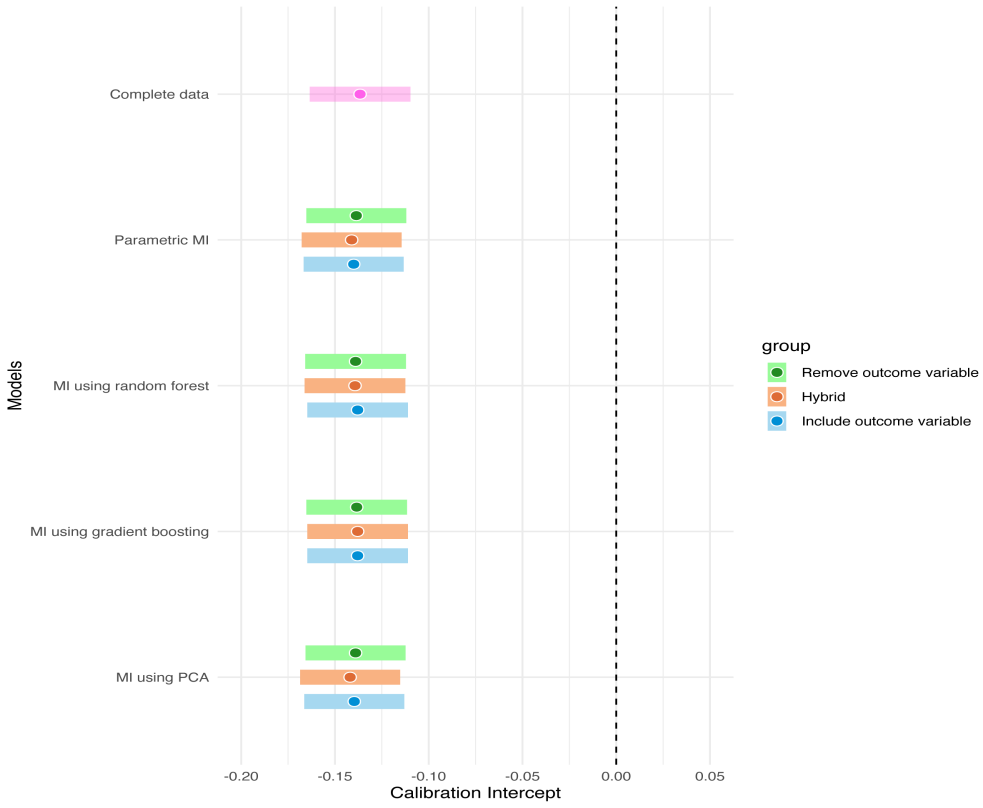
(c) Low dimensional with interactions (Rate of outcome: 6%)



(d) Low dimensional with interactions (Rate of outcome: 10%)

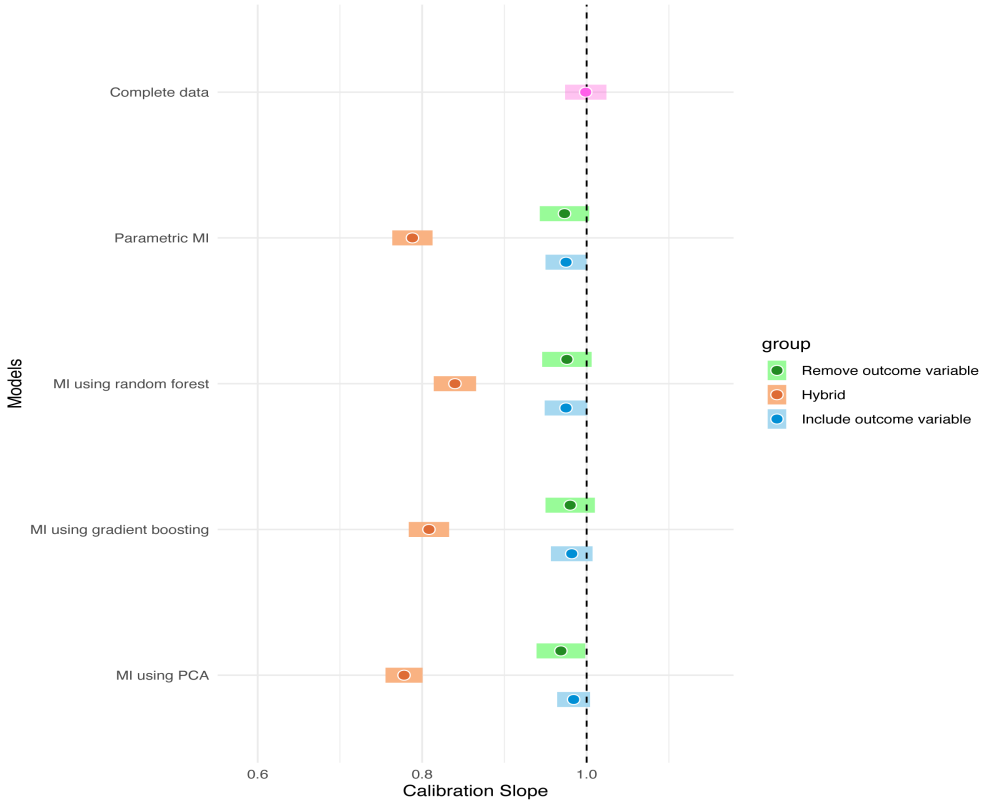


(e) High dimensional fitted with ridge regression (Rate of outcome: 10%)

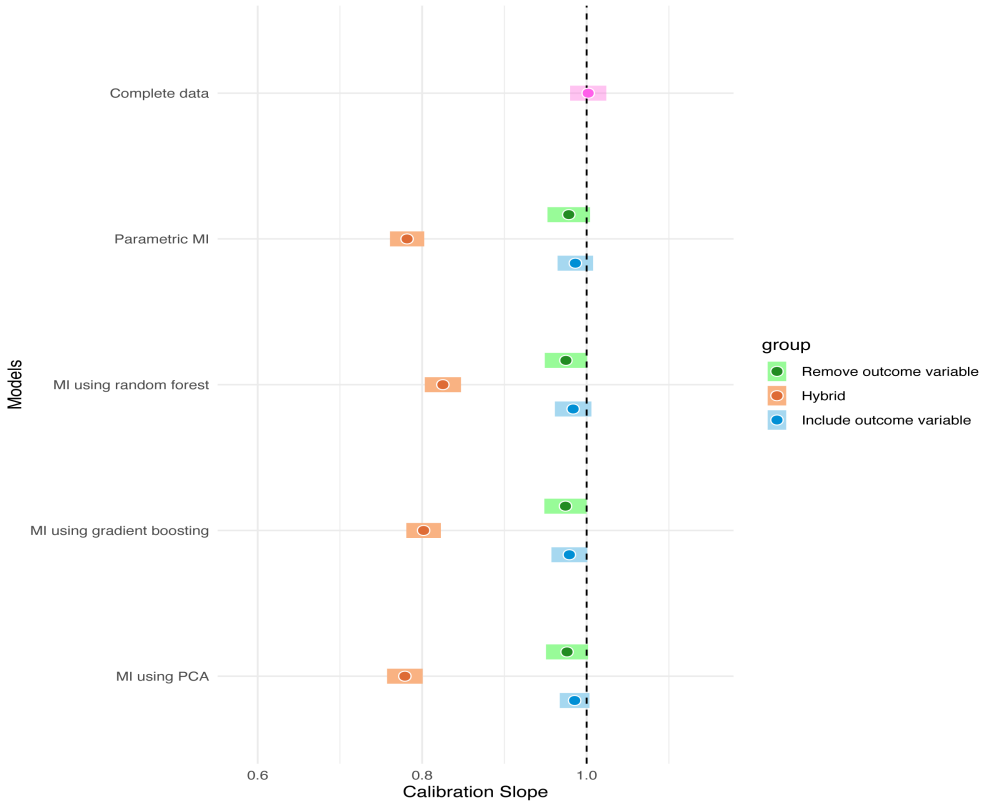


(f) High dimensional fitted with Random Forest (Rate of outcome: 10%)

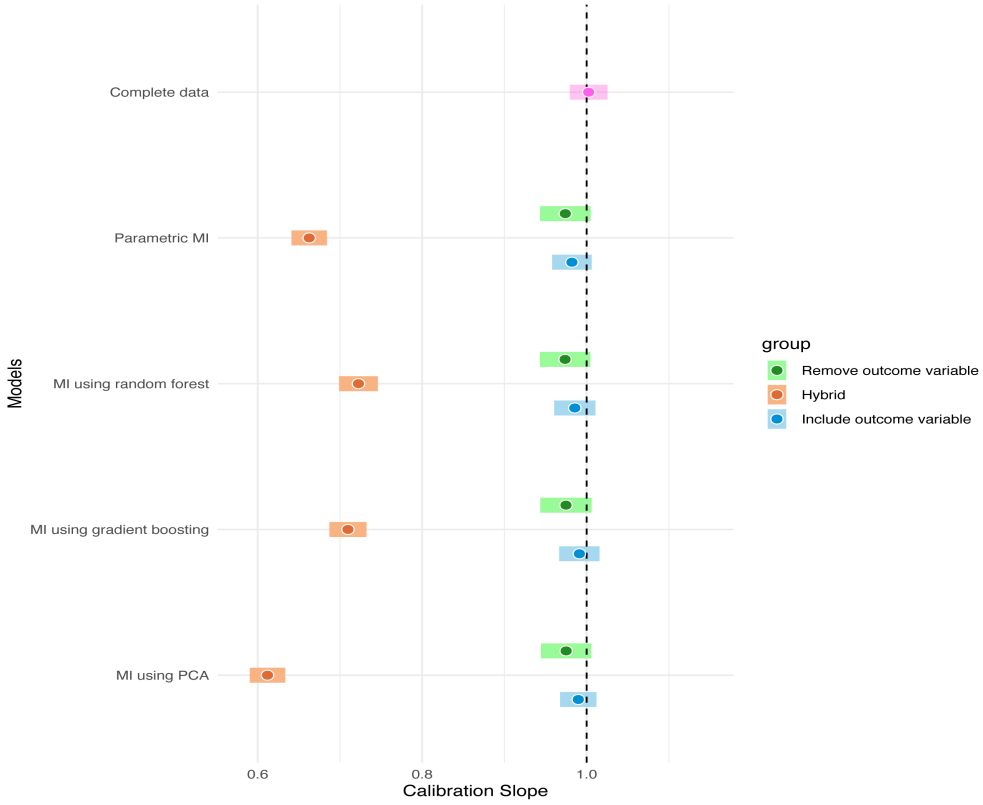
Figure 3.6: Calibration intercept as well as the corresponding 95% CI for fitted models under different simulation scenarios. The dashed vertical lines showed the ideal value for calibration intercept (0).



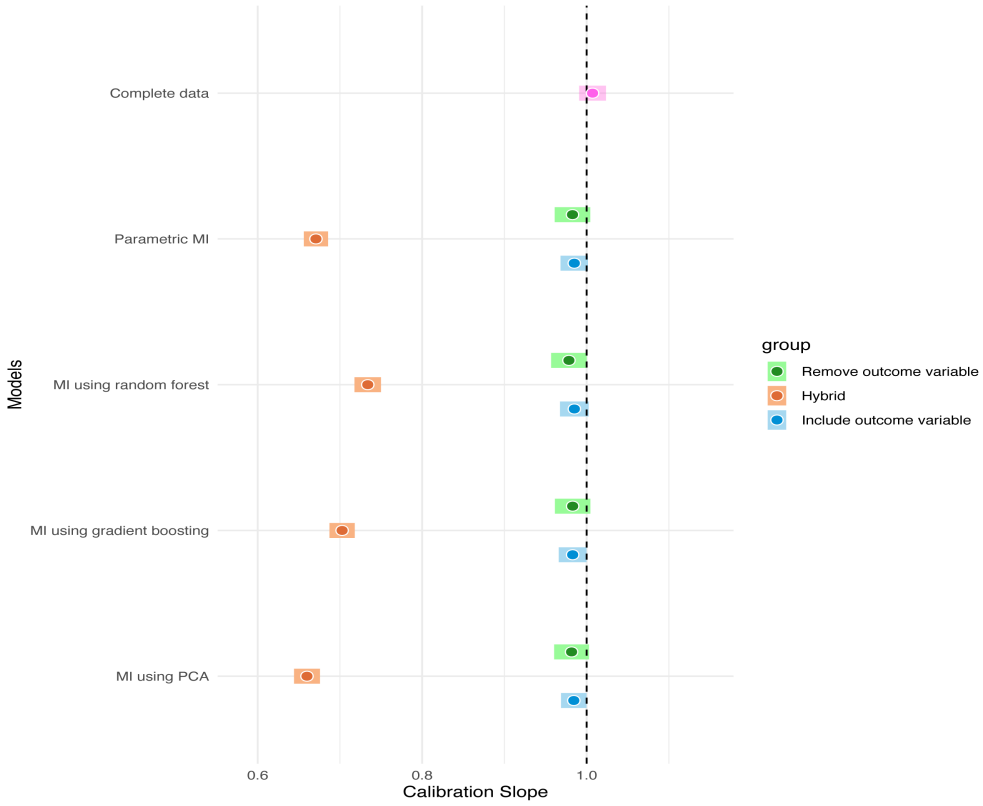
(a) Low dimensional (Rate of outcome: 6%)



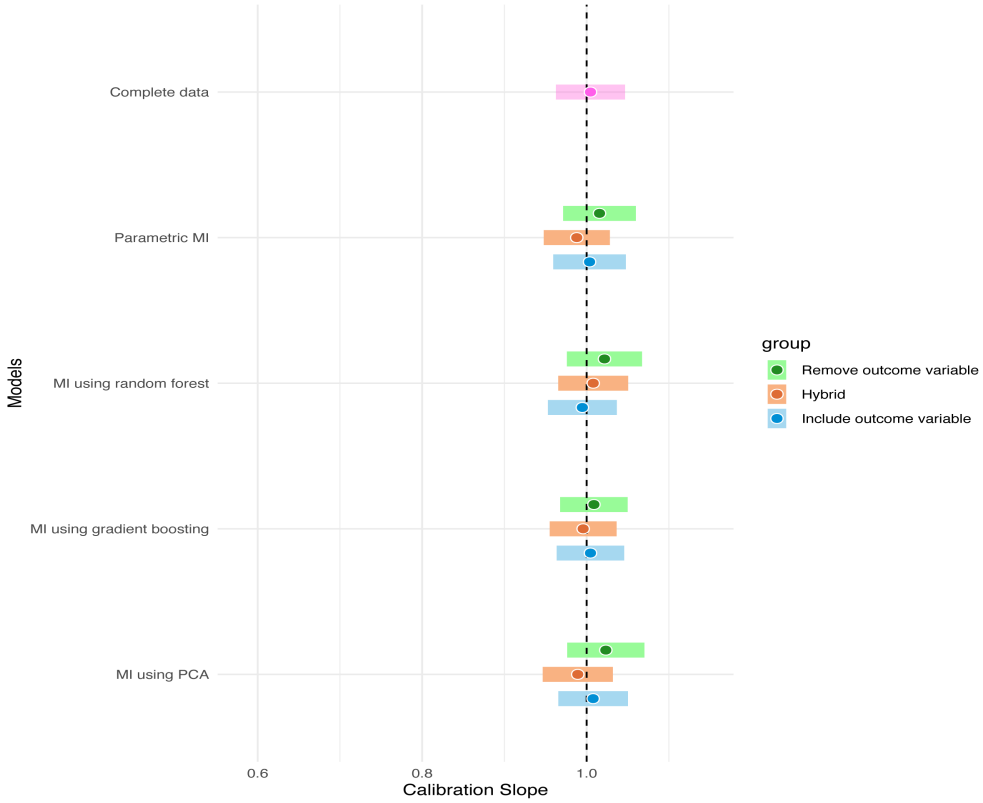
(b) Low dimensional (Rate of outcome: 10%)



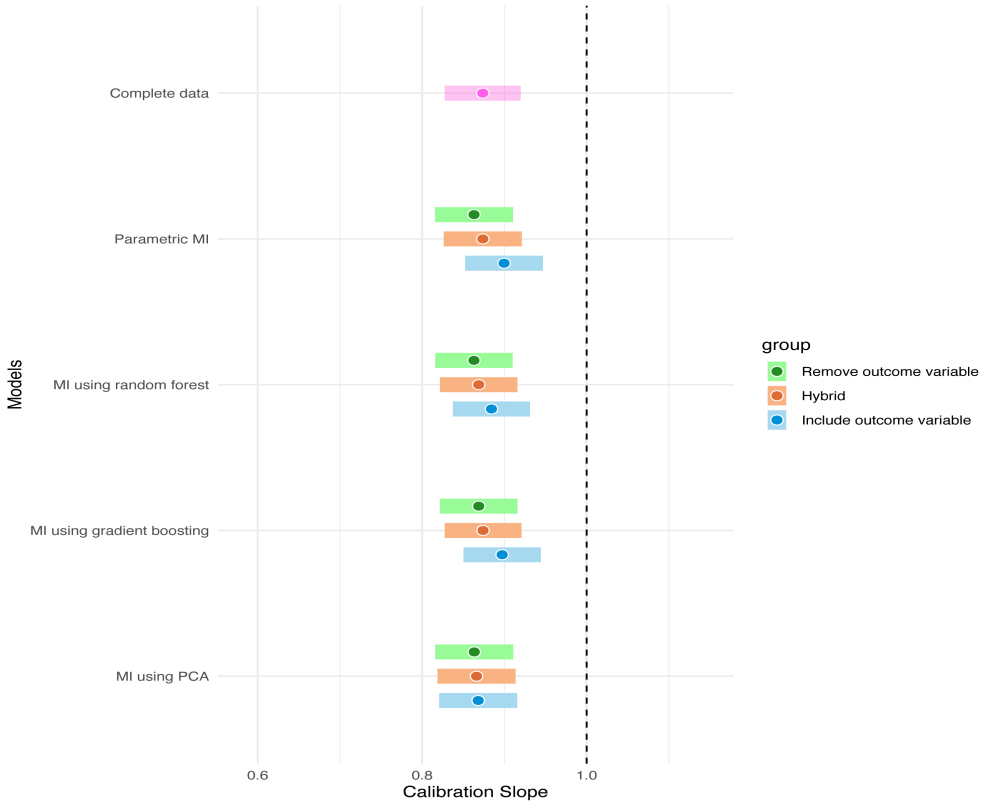
(c) Low dimensional with interactions (Rate of outcome: 6%)



(d) Low dimensional with interactions (Rate of outcome: 10%)

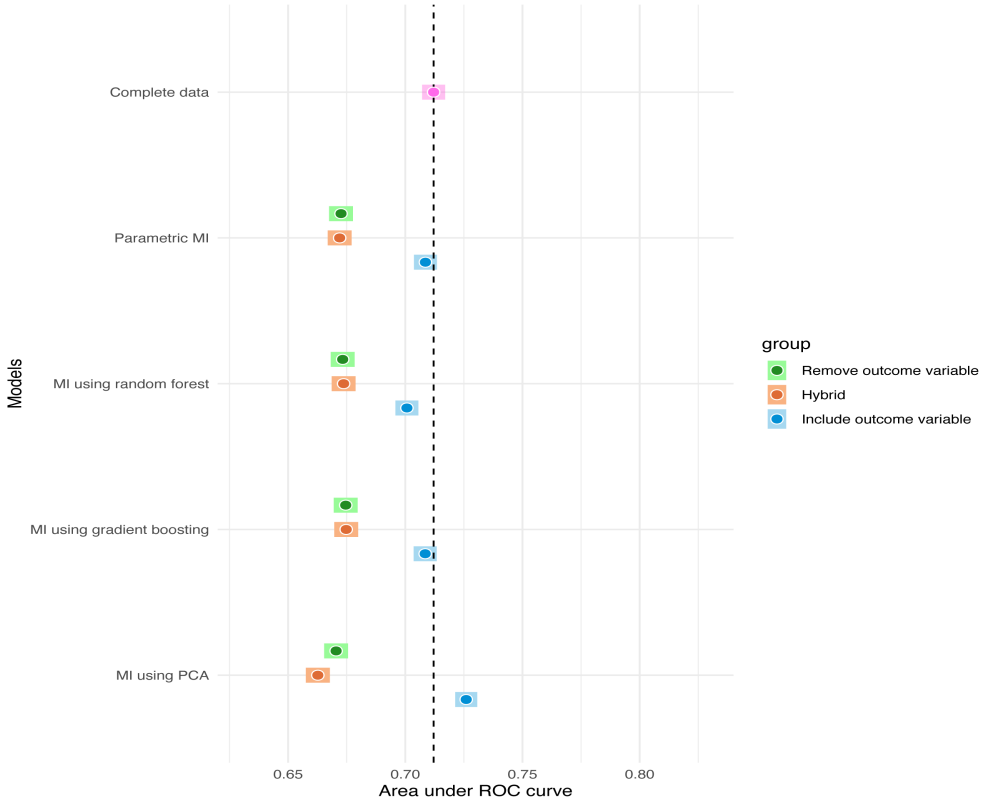


(e) High dimensional fitted with ridge regression (Rate of outcome: 10%)

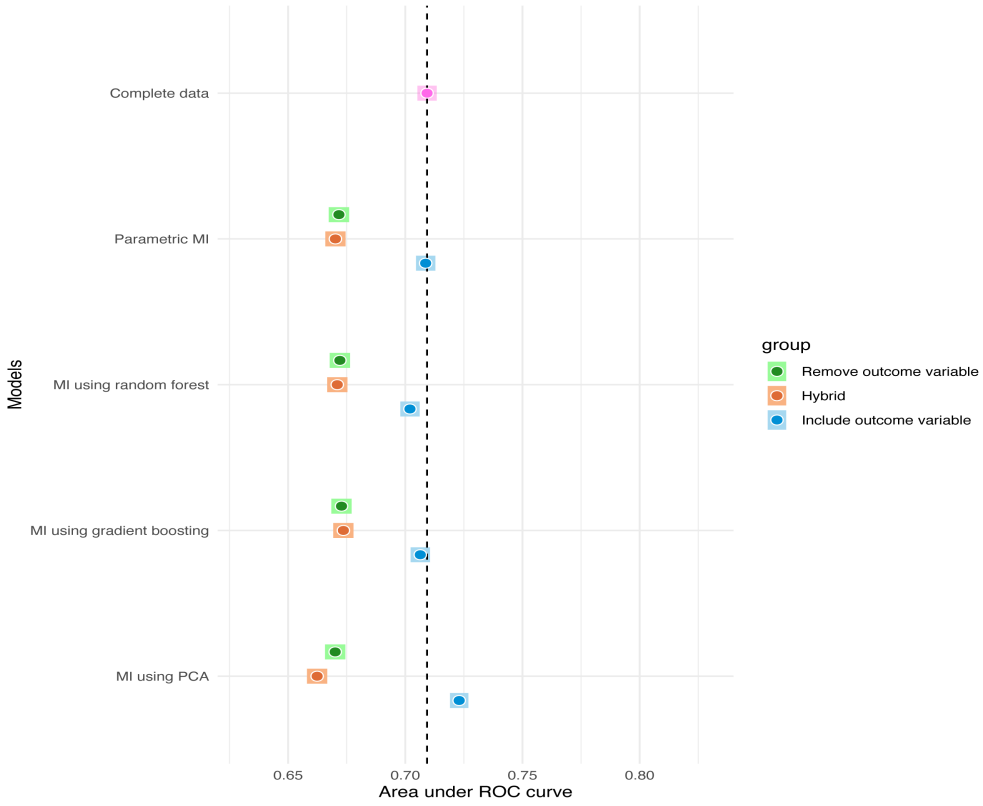


(f) High dimensional fitted with Random Forest (Rate of outcome: 10%)

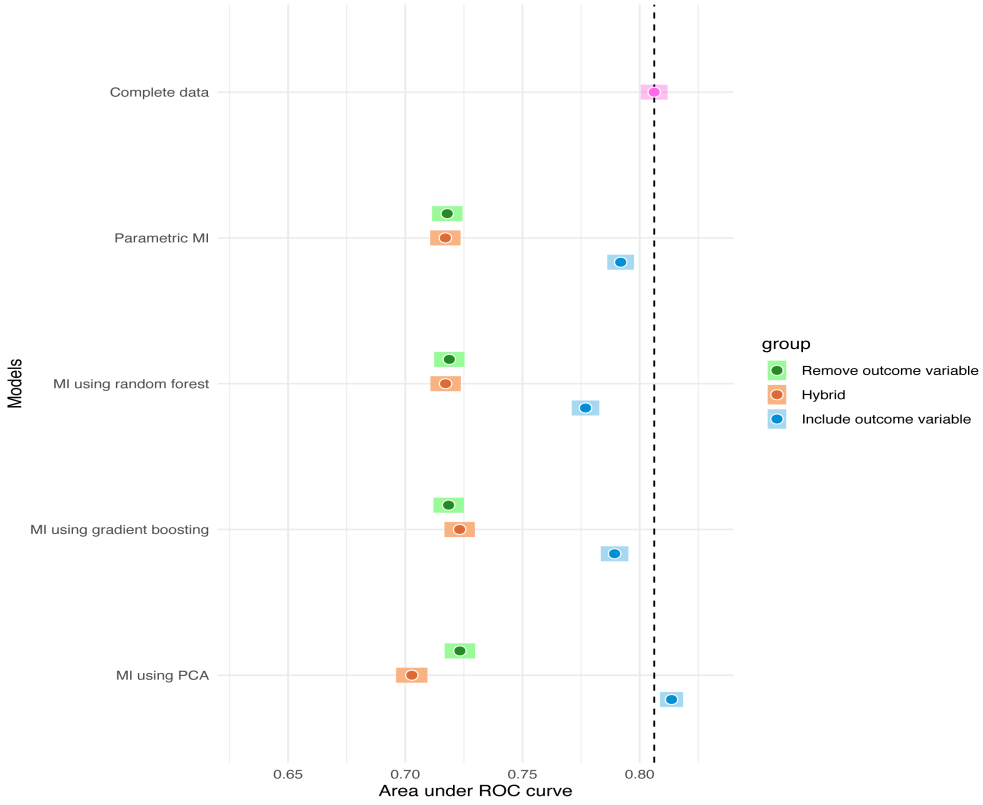
Figure 3.7: Calibration slope as well as the corresponding 95% CI for fitted models under different simulation scenarios. The dashed vertical lines showed the ideal value for calibration slope (1).



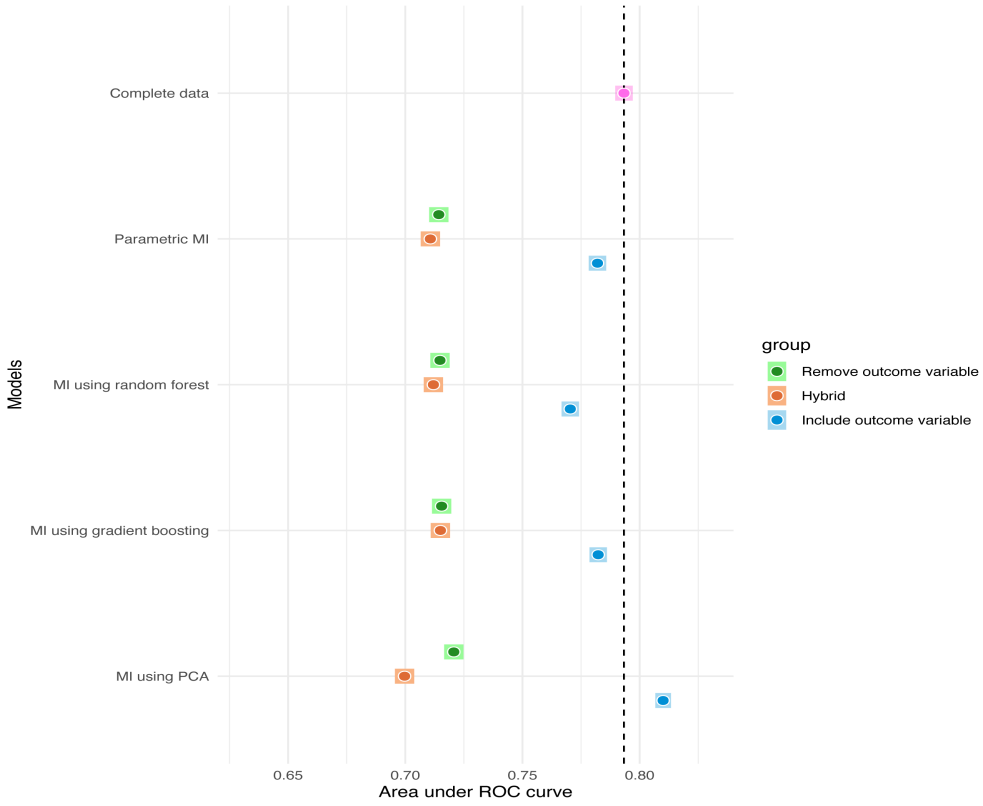
(a) Low dimensional (Rate of outcome: 6%)



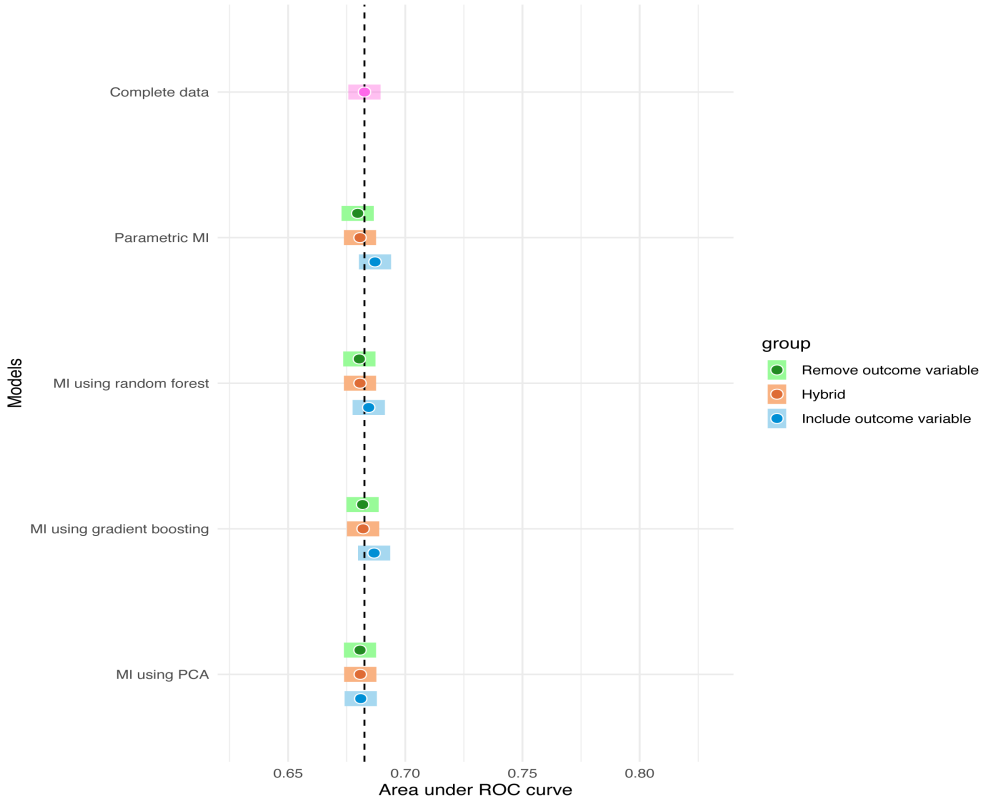
(b) Low dimensional (Rate of outcome: 10%)



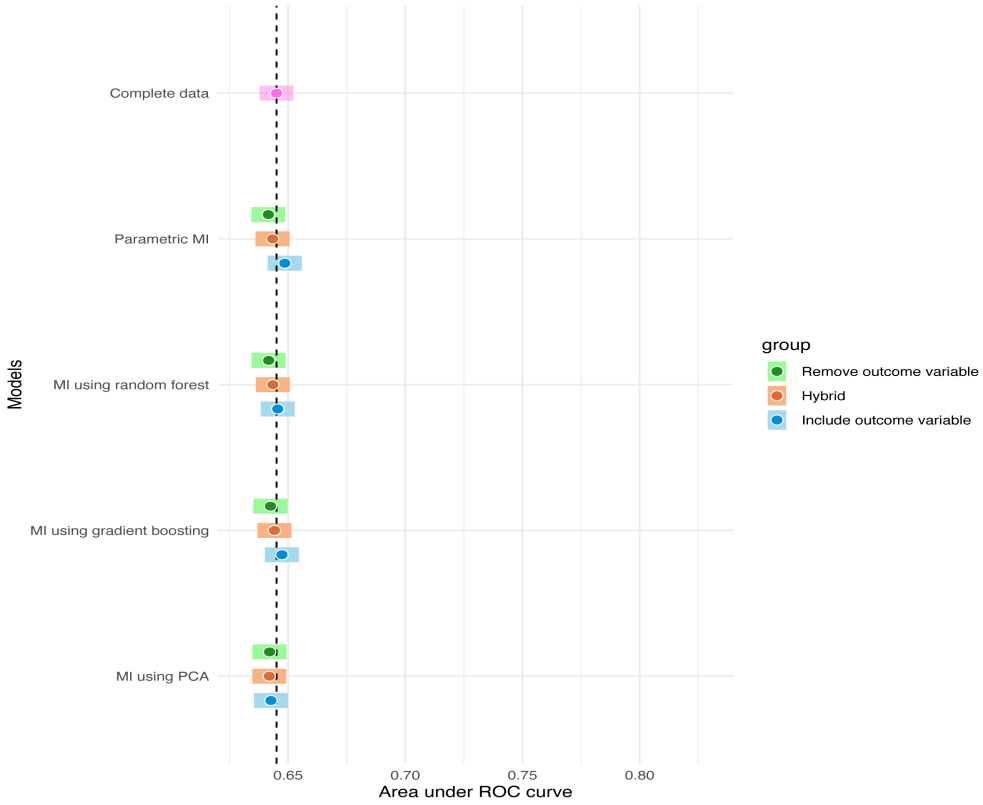
(c) Low dimensional with interactions (Rate of outcome: 6%)



(d) Low dimensional with interactions (Rate of outcome: 10%)



(e) High dimensional fitted with ridge regression (Rate of outcome: 10%)



(f) High dimensional fitted with Random Forest (Rate of outcome: 10%)

Figure 3.8: AUC as well as the corresponding 95% CI for fitted models under different simulation scenarios. The dashed vertical lines showed AUC of fitted model using the complete data.

or **no_y**). We didn't see meaningful differences in model performance among different approaches to imputing interaction terms (JAV, passive imputation or smcfcs) neither. Low dimensional settings with interaction terms in the underlying model and high dimensional setting demonstrated similar results. Evaluations of prediction models under different settings and scenarios were reported in Appendix C.

Chapter 4

APPLICATION TO BREAST CANCER SURVEILLANCE EXAMINATION DATA

We also applied methods studied to the Breast Cancer Surveillance Consortium (BCSC) dataset on modeling performance outcomes of surveillance mammogram examinations (mammography examinations performed for detecting asymptomatic subsequent breast cancer) in breast cancer survivors [7]. We used the BCSC surveillance cohort to serve the input for bootstrapping and a data illustration to compare these alternative MI approaches.

4.1 Study population

The data used focuses on surveillance mammogram examinations from three BCSC sites, namely New Hampshire, Vermont, and KP Washington. The study included women aged 18 years or older who were previously diagnosed with a breast cancer (AJCC 8th edition anatomic stage 0-III) from 1996 onward, based on the earliest date obtained from pathology and cancer registry files. Additionally, the women must have had breast imaging examinations occurring at least 6 months after their diagnosis date, allowing for serial diagnoses. We excluded women with incomplete diagnosis dates (only the year known) of primary breast cancer, missing age at diagnosis or stage at diagnosis, American Joint Commission on Cancer stage IV at diagnosis (8th edition) [2], no surgery or missing surgery data, bilateral mastectomy, or cancer diagnosis solely from biopsy data. Each BCSC registry and the Statistical Coordinating Center (SCC) have received institutional review board approval for all study procedures, including passive consenting processes (three registries) or a waiver of consent (three registries and the SCC), to enroll participants, link data, and perform analytic studies. All procedures are Health Insurance Portability and Accountability Act (HIPAA) compliant.

All registries and the SCC have received a Federal Certificate of Confidentiality and other protections for the identities of women, physicians, and facilities who are subjects of this research.

4.2 Outcome variables

The main outcome variables of interest in this study are performance measures quantifying surveillance harms: false-positive recall and false-positive biopsy. False-positive recall (hereafter referred to as FP recall) refers to cases where an initial end-of-day assessment of a mammogram indicated a positive finding, but no cancer was diagnosed within one year. False-positive biopsy (hereafter referred to as FP biopsy) refers to cases where a biopsy was recommended based on mammogram findings, but no cancer was diagnosed within one year.

4.3 Predictor variables

We considered all potential risk factors for surveillance harms (FP recall and FP biopsy), including:

1. **Women characteristics**: demographic characteristics included women's age at each surveillance exam, racial and ethnic group (Hispanic, Non-Hispanic Asian, Non-Hispanic Black, Non-Hispanic White and Others), body mass index (BMI) measured at each surveillance mammogram, menopausal status, and first-degree family history of breast cancer.

2. **Primary breast cancer characteristics**: pathological features including AJCC stage, histology, grade, and estrogen and progesterone receptor (ER/PR) status.

3. **Primary breast cancer treatment**: treatment included the type of surgery (breast-conserving surgery or mastectomy), use of radiation therapy, and the type of adjuvant therapy (Hormonal therapy, chemotherapy, both, or neither) based on all cancer registry records and pathology databases.

4. **Primary breast cancer diagnosis**: diagnosis information included women's age and calendar year at initial diagnosis, year since diagnosis (We intentionally included all women with a PHBC, regardless of time since diagnosis, since surveillance decisions are not limited to women with a recent diagnosis), and mode of detection (screening detected, interval detected or clinically detected).

5. **Exam characteristics**: imaging characteristics included surveillance mammography modality (film mammography, digital mammography, or digital breast tomosynthesis), BI-RADS breast density (almost entirely fatty, scattered fibroglandular, heterogeneously dense or extremely dense), number of previous false-positive surveillance exams, number of previous false-positive biopsies, time since previous mammogram, and time since previous surveillance mammogram (first surveillance, 3-8 months, 9-14 months, 15-23 months, or $\geq$ 24 months).

Among all predictor variables, continuous variables including age at each surveillance exam, age and calendar year at initial diagnosis, BMI were included in the models using natural cubic splines with four degrees of freedom. Time since previous mammogram was first log-transformed and then included in the models using natural cubic splines with three degrees of freedom. We also considered two-way interactions between age and all other predictors.

The analytical sample consisted of 100,149 surveillance mammograms from 16963 women, including 11235 (11.2%) FP recalls and 1658 (1.7%) FP biopsies. Risk factors for the two outcomes were summarized in Table 4.1 and 4.2. Most mammograms were from Non-Hispanic White women (91.4%). Median age at survellience mammograms was 65 years and median BMI was 26.6. There are several predictors that contain missing values and the missing rate ranged from 0.2% (type of surgery) to 23.5% (BMI). Mammograms were treated as independent observations, conditional on all observed predictors.

Table 4.1: Distribution of risk factors stratified by false-positive recall outcome

	False positive recall of mammogram: No (N=88914)	False positive recall of mammogram: Yes (N=11235)	Overall (N=100149)
Age at surveillance exam			
Mean (SD)	64.9 (11.2)	63.1 (11.8)	64.7 (11.3)
Median [Min, Max]	65.0 [24.0, 90.0]	63.0 [25.0, 90.0]	65.0 [24.0, 90.0]
Race and ethnicity			
Hispanic	1544 (1.7%)	189 (1.7%)	1733 (1.7%)
Non-Hispanic Asian	2286 (2.6%)	223 (2.0%)	2509 (2.5%)
Non-Hispanic Black	888 (1.0%)	121 (1.1%)	1009 (1.0%)
Non-Hispanic White	81317 (91.5%)	10246 (91.2%)	91563 (91.4%)
Others	1614 (1.8%)	225 (2.0%)	1839 (1.8%)
Missing	1265 (1.4%)	231 (2.1%)	1496 (1.5%)
Menopausal status			
Postmenopause	68762 (77.3%)	8160 (72.6%)	76922 (76.8%)
Peri/Premenopause	5619 (6.3%)	1174 (10.4%)	6793 (6.8%)
Missing	14533 (16.3%)	1901 (16.9%)	16434 (16.4%)
Family history of breast cancer			
No	64398 (72.4%)	8272 (73.6%)	72670 (72.6%)
Yes	23950 (26.9%)	2889 (25.7%)	26839 (26.8%)
Missing	566 (0.6%)	74 (0.7%)	640 (0.6%)
Body mass index (kg/m2)			
Mean (SD)	27.8 (6.27)	27.9 (6.30)	27.8 (6.27)
Median [Min, Max]	26.6 [15.1, 76.5]	26.6 [15.2, 82.4]	26.6 [15.1, 82.4]
Missing	20602 (23.2%)	2908 (25.9%)	23510 (23.5%)
Surveillance mammography modality			
Film mammography	29174 (32.8%)	5399 (48.1%)	34573 (34.5%)
Digital mammography	46356 (52.1%)	5047 (44.9%)	51403 (51.3%)
Digital breast tomosynthesis	13142 (14.8%)	746 (6.6%)	13888 (13.9%)
Missing	242 (0.3%)	43 (0.4%)	285 (0.3%)
BI-RADS breast density			
Almost entirely fatty	8551 (9.6%)	624 (5.6%)	9175 (9.2%)
Scattered fibroglandular	38665 (43.5%)	4285 (38.1%)	42950 (42.9%)
Heterogeneously dense	30798 (34.6%)	4550 (40.5%)	35348 (35.3%)
Extremely dense	4624 (5.2%)	499 (4.4%)	5123 (5.1%)
Missing	6276 (7.1%)	1277 (11.4%)	7553 (7.5%)
Months since last mammogram			
Mean (SD)	12.2 (5.74)	10.3 (5.91)	12.0 (5.79)
Median [Min, Max]	12.0 [3.00, 254]	11.0 [3.00, 153]	12.0 [3.00, 254]
Months since last surveillance mammogram			
1st surveillance mammogram	13171 (14.8%)	3125 (27.8%)	16296 (16.3%)
3-8	7334 (8.2%)	3048 (27.1%)	10382 (10.4%)
9-14	57377 (64.5%)	4100 (36.5%)	61477 (61.4%)
15-23	5781 (6.5%)	533 (4.7%)	6314 (6.3%)
24+	5251 (5.9%)	429 (3.8%)	5680 (5.7%)
Previous false positive recall of mammogram counts			
Mean (SD)	0.641 (1.18)	0.795 (1.22)	0.658 (1.18)
Median [Min, Max]	0 [0, 10.0]	0 [0, 9.00]	0 [0, 10.0]
Previous false positive biopsy counts			
Mean (SD)	0.0746 (0.293)	0.0849 (0.317)	0.0758 (0.295)
Median [Min, Max]	0 [0, 4.00]	0 [0, 4.00]	0 [0, 4.00]
Mode of detection of index breast cancer			
Screening detected	57406 (64.6%)	7310 (65.1%)	64716 (64.6%)
Interval detected	20211 (22.7%)	2541 (22.6%)	22752 (22.7%)
Clinically detected	7675 (8.6%)	973 (8.7%)	8648 (8.6%)
Missing	3622 (4.1%)	411 (3.7%)	4033 (4.0%)
Age at index breast cancer diagnosis			
Mean (SD)	59.2 (11.4)	59.2 (11.7)	59.2 (11.4)
Median [Min, Max]	59.0 [24.0, 90.0]	59.0 [24.0, 90.0]	59.0 [24.0, 90.0]
Years since diagnosis of index breast cancer			
1-2	23280 (26.2%)	4903 (43.6%)	28183 (28.1%)
<1	4676 (5.3%)	1559 (13.9%)	6235 (6.2%)
3–4	18348 (20.6%)	1704 (15.2%)	20052 (20.0%)
5–6	13770 (15.5%)	1068 (9.5%)	14838 (14.8%)
7-9	14147 (15.9%)	1103 (9.8%)	15250 (15.2%)
>10	14693 (16.5%)	898 (8.0%)	15591 (15.6%)
Calendar year of index breast cancer diagnosis			
Mean (SD)	2000 (5.33)	2000 (5.17)	2000 (5.32)
Median [Min, Max]	2000 [2000, 2020]	2000 [2000, 2020]	2000 [2000, 2020]
Histology of index breast cancer			
Ductal	56707 (63.8%)	7155 (63.7%)	63862 (63.8%)
Non-ductal	13341 (15.0%)	1586 (14.1%)	14927 (14.9%)
Missing	18866 (21.2%)	2494 (22.2%)	21360 (21.3%)
AJCC v8 stage of index breast cancer (anatomic)			
DCIS	19012 (21.4%)	2510 (22.3%)	21522 (21.5%)
Stage I	44537 (50.1%)	5709 (50.8%)	50246 (50.2%)
Stage IIA and II NOS	14875 (16.7%)	1858 (16.5%)	16733 (16.7%)
Stage IIB and above	10490 (11.8%)	1158 (10.3%)	11648 (11.6%)
Grade of index breast cancer			
1	20362 (22.9%)	2633 (23.4%)	22995 (23.0%)
2	35719 (40.2%)	4501 (40.1%)	40220 (40.2%)
3	26430 (29.7%)	3277 (29.2%)	29707 (29.7%)
Missing	6403 (7.2%)	824 (7.3%)	7227 (7.2%)
ER and PR status of index breast cancer			
ER+, PR+	55405 (62.3%)	6818 (60.7%)	62223 (62.1%)
ER+, PR-	6370 (7.2%)	761 (6.8%)	7131 (7.1%)
ER-, PR+	870 (1.0%)	112 (1.0%)	982 (1.0%)
ER-, PR-	9781 (11.0%)	1285 (11.4%)	11066 (11.0%)
Missing	16488 (18.5%)	2259 (20.1%)	18747 (18.7%)
Surgical treatment for index breast cancer			
Mastectomy	22839 (25.7%)	1703 (15.2%)	24542 (24.5%)
Breast conserving surgery	65945 (74.2%)	9504 (84.6%)	75449 (75.3%)
Missing	130 (0.1%)	28 (0.2%)	158 (0.2%)
Radiation treatment for index breast cancer			
With radiation	49840 (56.1%)	7171 (63.8%)	57011 (56.9%)
Without radiation	38384 (43.2%)	3976 (35.4%)	42360 (42.3%)
Missing	690 (0.8%)	88 (0.8%)	778 (0.8%)
Adjuvant therapy for index breast cancer			
None	30327 (34.1%)	4240 (37.7%)	34567 (34.5%)
Chemotherapy only	13055 (14.7%)	1613 (14.4%)	14668 (14.6%)
Hormonal therapy only	28489 (32.0%)	3376 (30.0%)	31865 (31.8%)
Both	11024 (12.4%)	1245 (11.1%)	12269 (12.3%)
Missing	6019 (6.8%)	761 (6.8%)	6780 (6.8%)

Table 4.2: Distribution of risk factors stratified by false-positive biopsy outcome

	False positive biopy: No (N=98491)	False positive biopy: Yes (N=1658)	Overall (N=100149)
Age at surveillance exam			
Mean (SD)	64.7 (11.3)	63.8 (11.3)	64.7 (11.3)
Median [Min, Max]	65.0 [24.0, 90.0]	64.0 [33.0, 90.0]	65.0 [24.0, 90.0]
Race and ethnicity			
Hispanic	1710 (1.7%)	23 (1.4%)	1733 (1.7%)
Non-Hispanic Asian	2482 (2.5%)	27 (1.6%)	2509 (2.5%)
Non-Hispanic Black	985 (1.0%)	24 (1.4%)	1009 (1.0%)
Non-Hispanic White	90070 (91.5%)	1493 (90.0%)	91563 (91.4%)
Others	1807 (1.8%)	32 (1.9%)	1839 (1.8%)
Missing	1437 (1.5%)	59 (3.6%)	1496 (1.5%)
Menopausal status			
Postmenopause	75699 (76.9%)	1223 (73.8%)	76922 (76.8%)
Peri/Premenopause	6652 (6.8%)	141 (8.5%)	6793 (6.8%)
Missing	16140 (16.4%)	294 (17.7%)	16434 (16.4%)
Family history of breast cancer			
No	71495 (72.6%)	1175 (70.9%)	72670 (72.6%)
Yes	26364 (26.8%)	475 (28.6%)	26839 (26.8%)
Missing	632 (0.6%)	8 (0.5%)	640 (0.6%)
Body mass index (kg/m2)			
Mean (SD)	27.8 (6.26)	28.7 (6.72)	27.8 (6.27)
Median [Min, Max]	26.6 [15.1, 82.4]	27.5 [15.7, 57.8]	26.6 [15.1, 82.4]
Missing	23051 (23.4%)	459 (27.7%)	23510 (23.5%)
Surveillance mammography modality			
Film mammography	33924 (34.4%)	649 (39.1%)	34573 (34.5%)
Digital mammography	50564 (51.3%)	839 (50.6%)	51403 (51.3%)
Digital breast tomosynthesis	13726 (13.9%)	162 (9.8%)	13888 (13.9%)
Missing	277 (0.3%)	8 (0.5%)	285 (0.3%)
BI-RADS breast density			
Almost entirely fatty	9051 (9.2%)	124 (7.5%)	9175 (9.2%)
Scattered fibroglandular	42324 (43.0%)	626 (37.8%)	42950 (42.9%)
Heterogeneously dense	34773 (35.3%)	575 (34.7%)	35348 (35.3%)
Extremely dense	5042 (5.1%)	81 (4.9%)	5123 (5.1%)
Missing	7301 (7.4%)	252 (15.2%)	7553 (7.5%)
Months since last mammogram			
Mean (SD)	12.0 (5.78)	11.6 (6.50)	12.0 (5.79)
Median [Min, Max]	12.0 [3.00, 254]	12.0 [3.00, 115]	12.0 [3.00, 254]
Months since last surveillance mammogram			
1st surveillance mammogram	15911 (16.2%)	385 (23.2%)	16296 (16.3%)
3-8	10122 (10.3%)	260 (15.7%)	10382 (10.4%)
9-14	60677 (61.6%)	800 (48.3%)	61477 (61.4%)
15-23	6202 (6.3%)	112 (6.8%)	6314 (6.3%)
24+	5579 (5.7%)	101 (6.1%)	5680 (5.7%)
Previous false positive recall of mammogram counts			
Mean (SD)	0.657 (1.18)	0.742 (1.25)	0.658 (1.18)
Median [Min, Max]	0 [0, 10.0]	0 [0, 8.00]	0 [0, 10.0]
Previous false positive biopsy counts			
Mean (SD)	0.0747 (0.292)	0.138 (0.438)	0.0758 (0.295)
Median [Min, Max]	0 [0, 4.00]	0 [0, 4.00]	0 [0, 4.00]
Mode of detection of index breast cancer			
Screening detected	63627 (64.6%)	1089 (65.7%)	64716 (64.6%)
Interval detected	22396 (22.7%)	356 (21.5%)	22752 (22.7%)
Clinically detected	8519 (8.6%)	129 (7.8%)	8648 (8.6%)
Missing	3949 (4.0%)	84 (5.1%)	4033 (4.0%)
Age at index breast cancer diagnosis			
Mean (SD)	59.2 (11.4)	59.3 (11.3)	59.2 (11.4)
Median [Min, Max]	59.0 [24.0, 90.0]	58.0 [25.0, 90.0]	59.0 [24.0, 90.0]
Years since diagnosis of index breast cancer			
1-2	27616 (28.0%)	567 (34.2%)	28183 (28.1%)
<1	6078 (6.2%)	157 (9.5%)	6235 (6.2%)
3-4	19714 (20.0%)	338 (20.4%)	20052 (20.0%)
5-6	14621 (14.8%)	217 (13.1%)	14838 (14.8%)
7-9	15034 (15.3%)	216 (13.0%)	15250 (15.2%)
>10	15428 (15.7%)	163 (9.8%)	15591 (15.6%)
Calendar year of index breast cancer diagnosis			
Mean (SD)	2000 (5.32)	2000 (5.32)	2000 (5.32)
Median [Min, Max]	2000 [2000, 2020]	2000 [2000, 2020]	2000 [2000, 2020]
Histology of index breast cancer			
Ductal	62812 (63.8%)	1050 (63.3%)	63862 (63.8%)
Non-ductal	14712 (14.9%)	215 (13.0%)	14927 (14.9%)
Missing	20967 (21.3%)	393 (23.7%)	21360 (21.3%)
AJCC v8 stage of index breast cancer (anatomic)			
DCIS	21127 (21.5%)	395 (23.8%)	21522 (21.5%)
Stage I	49459 (50.2%)	787 (47.5%)	50246 (50.2%)
Stage IIA and II NOS	16448 (16.7%)	285 (17.2%)	16733 (16.7%)
Stage IIB and above	11457 (11.6%)	191 (11.5%)	11648 (11.6%)
Grade of index breast cancer			
1	22627 (23.0%)	368 (22.2%)	22995 (23.0%)
2	39589 (40.2%)	631 (38.1%)	40220 (40.2%)
3	29190 (29.6%)	517 (31.2%)	29707 (29.7%)
Missing	7085 (7.2%)	142 (8.6%)	7227 (7.2%)
ER and PR status of index breast cancer			
ER+, PR+	61220 (62.2%)	1003 (60.5%)	62223 (62.1%)
ER+, PR-	7014 (7.1%)	117 (7.1%)	7131 (7.1%)
ER-, PR+	965 (1.0%)	17 (1.0%)	982 (1.0%)
ER-, PR-	10889 (11.1%)	177 (10.7%)	11066 (11.0%)
Missing	18403 (18.7%)	344 (20.7%)	18747 (18.7%)
Surgical treatment for index breast cancer			
Mastectomy	24241 (24.6%)	301 (18.2%)	24542 (24.5%)
Breast conserving surgery	74099 (75.2%)	1350 (81.4%)	75449 (75.3%)
Missing	151 (0.2%)	7 (0.4%)	158 (0.2%)
Radiation treatment for index breast cancer			
With radiation	56013 (56.9%)	998 (60.2%)	57011 (56.9%)
Without radiation	41718 (42.4%)	642 (38.7%)	42360 (42.3%)
Missing	760 (0.8%)	18 (1.1%)	778 (0.8%)
Adjuvant therapy for index breast cancer			
None	33928 (34.4%)	639 (38.5%)	34567 (34.5%)
Chemotherapy only	14451 (14.7%)	217 (13.1%)	14668 (14.6%)
Hormonal therapy only	31373 (31.9%)	492 (29.7%)	31865 (31.8%)
Both	12078 (12.3%)	191 (11.5%)	12269 (12.3%)
Missing	6661 (6.8%)	119 (7.2%)	6780 (6.8%)

4.4 Data imputation

We randomly split the data into training set (80%) and test set (20%) based on women's level and then applied the discussed methods to impute missing values of each predictor using all other candidate predictors in the training and test sets separately. For each MI method, we also imputed the missing values in three different scenarios: (1) **with_y**; (2) **hybrid**; (3) **no_y**. During simulation studies we conducted previously, we did not find any substantial differences between different approaches in imputing interaction terms. For the sake of simplicity, we imputed the interaction terms of BCSC data using the JAV approach, which directly imputes transformed values. All predictions and assessments were based on 25 imputed complete datasets.

4.5 Model fitting and evaluation

We conducted predictive modeling on each of the imputed data sets. We fitted the models on the training sets while evaluated and compared two different risk prediction models (Ridge Regression and Random Forest) on the test sets for each of the two outcomes (FP recall and FP biopsy) assessing the calibration (E/O ratio, calibration intercept, calibration slope) of predicted risks as well as discriminatory accuracy (AUC). For Ridge Regression, we used the function `cv.glmnet` in the R package `glmnet` to perform 5-fold cross-validation for tuning parameter selection. For Random Forest, we utilized the `caret` package to help us select two tuning parameters: `mtry` (the number of variables to randomly sampled at each split) and `ntree` (the number of trees to build in the Random Forest model). For `mtry`, we tried 7, 15, and 30, which corresponds to $0.5\sqrt{p}$, $\sqrt{p}$, $2\sqrt{p}$, respectively (p refers to the number of predictors in the model, we have 251 predictors); for `ntree`, we considered 500, 1000 and 1500. A 5-fold CV was used for tuning parameter selection.

For the data illustration, model performance measured by sensitivity and positive predictive value (PPV) at the 95^{th} percentile of the risk score distribution, was also evaluated for different approaches. Sensitivity refers to the percentage of exams with a predicted risk score

greater than the risk score threshold (95^{th} percentile of the risk score distribution) among all exams with FP recall/FP biopsy. PPV refers to the percentage of exams with FP recall/FP biopsy among all visits with a predicted risk score greater than the risk score threshold (95^{th} percentile of the risk score distribution).

4.6 Bootstrapping

We performed the bootstrapping process 100 times from the BCSC dataset. In each iteration, we randomly selected a sample of mammograms from a subset of 5000 women, without replacement, to mimic the original dataset's characteristics. Train-test split was based on the 5000 women's data, with 80% (4000 women) as the training set and 20% (1000 women) as the test set. We constructed 95% CIs of performance metrics in the testing set via 100 bootstrap samples.

Models fitted using Ridge Regression:

For FP recall, all MI approaches under three scenarios showed that the 95% CI for E/O ratios and calibration intercepts overlapped 1 and 0 respectively (Figure 4.1a, 4.1b), indicating no systematic biases. All MI approaches under three scenarios demonstrated that 95% CI for calibration slopes (Figure 4.1c) overlapped 1, suggesting no over/under-fitting. Moreover, we didn't observe meaningful differences in AUCs (0.71, 95% CI: [0.69, 0.73]) of all MI approaches (Figure 4.1d).

For FP biopsy, all MI approaches under three scenarios showed that the 95% CI for E/O ratios and calibration intercepts overlapped 1 and 0 respectively (Figure 4.2a, 4.2b), indicating no systematic biases. All MI approaches under three scenarios demonstrated that 95% CI for calibration slopes (Figure 4.2c) overlapped 1, suggesting no over/under-fitting. $\mathbf{MI_{GB}}$ (Figure 4.2d) had slightly higher AUCs (0.60, 95% CI: [0.54, 0.66]) than the other three MI approaches (0.58, 95% CI: [0.52, 0.64]).

Models fitted using Random Forest:

For FP recall, all MI approaches under three scenarios showed that the 95% CI for E/O ratios and calibration intercepts failed to overlap 1 and 0 respectively (Figure 4.3a,

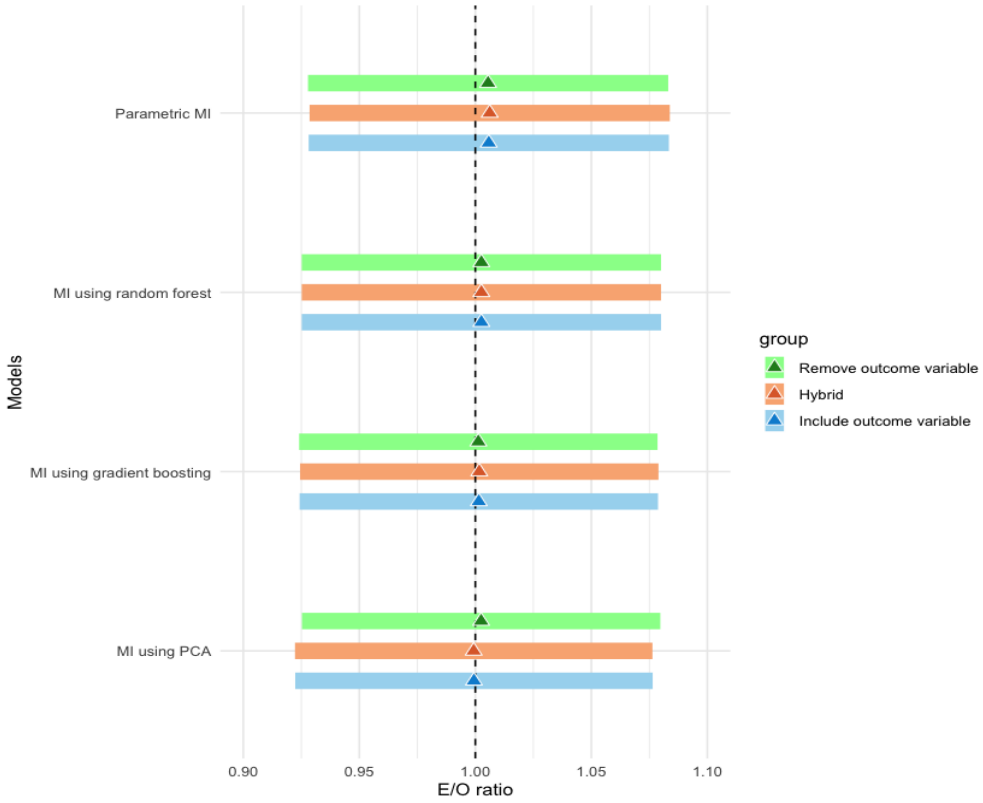
4.3b), indicating overestimation. All MI approaches under three scenarios had calibration slopes slightly greater than 1 (Figure 4.3c), but 95% CI for calibration slopes overlapped 1, suggesting no over/under-fitting. Moreover, we didn't observe meaningful differences in AUCs (0.70, 95% CI: [0.68, 0.72]) of all MI approaches (Figure 4.3d).

For FP biopsy, all MI approaches under three scenarios had E/O ratios greater than 1 and calibration intercepts lower than 1, although the 95% CI for E/O ratios and calibration intercepts overlapped 1 and 0 respectively (Figure 4.4a, 4.4b). All MI approaches under three scenarios had calibration slopes (Figure 4.4c) significantly less than 1, suggesting overfitting. $\mathbf{MI_{para}}$ and $\mathbf{MI_{RF}}$ (Figure 4.4d) had slightly higher AUCs (0.57, 95% CI: [0.51, 0.63]) than $\mathbf{MI_{GB}}$ and $\mathbf{MI_{PCA}}$ (0.56, 95% CI: [0.50, 0.62]).

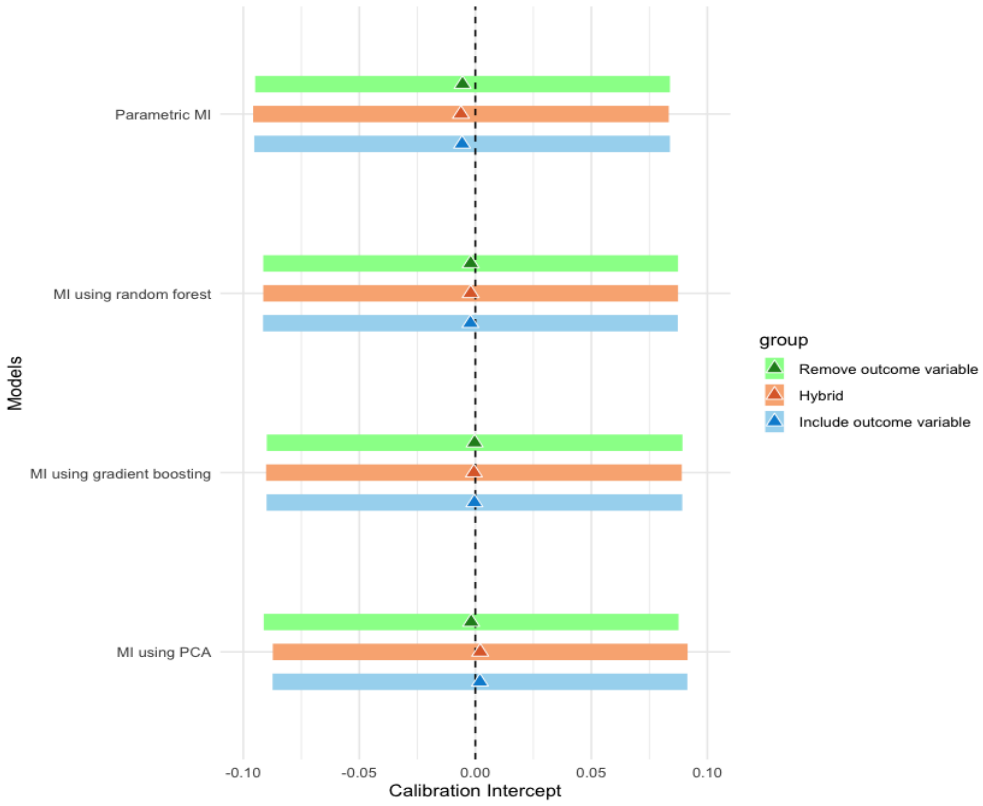
4.7 Data illustration

We also utilized the whole BCSC dataset to serve as a data example. Firstly, we conducted a 10-fold cross-validation to assess predictive performance using imputed data by different approaches. Samples were randomly partitioned into 10 groups at the woman level, among which 9 groups were used as the training set and the remaining one as the testing set in each cross-validation around. We constructed 95% CIs of performance metrics in the testing set via 10 estimates of standard errors from cross-validation. Furthermore, risk prediction models built using entire imputed data were evaluated using the sensitivity and PPV at the 95^{th} percentile of the risk score distribution. Ridge regression was employed to develop risk prediction models using each imputed dataset.

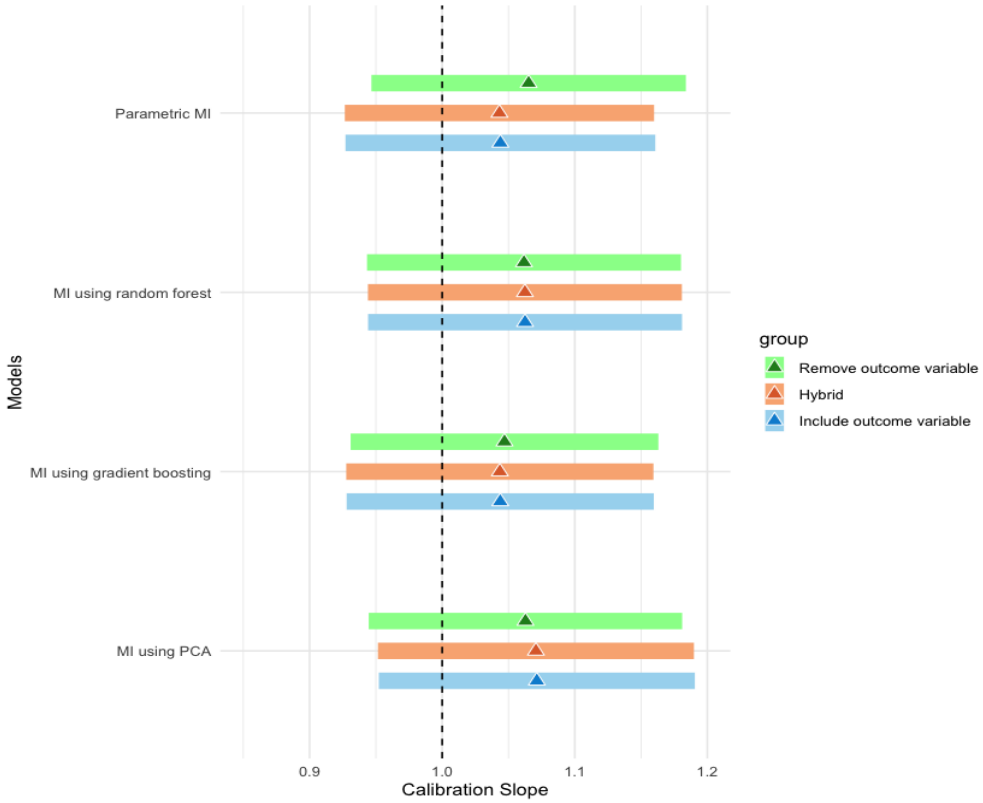
Figure 4.5 and 4.6 reported the calibration and discrimination for different MI approaches based on the 10-fold cross-validation. The results aligned with the bootstrapping we did previously. For both FP recall and FP biopsy, all MI approaches under three scenarios showed that the 95% CIs for E/O ratios, calibration intercepts, and calibration slopes overlapped 1, 0, and 1 respectively. For FP recall, there was no substantial differences in AUCs for all MI approaches (0.71, 95% CI: [0.70, 0.72]). For FP biopsy, $\mathbf{MI_{GB}}$ had slightly higher AUCs (0.62, 95% CI:[0.60, 0.64]) than the other three approaches (0.61, 95% CI:[0.59, 0.63]).



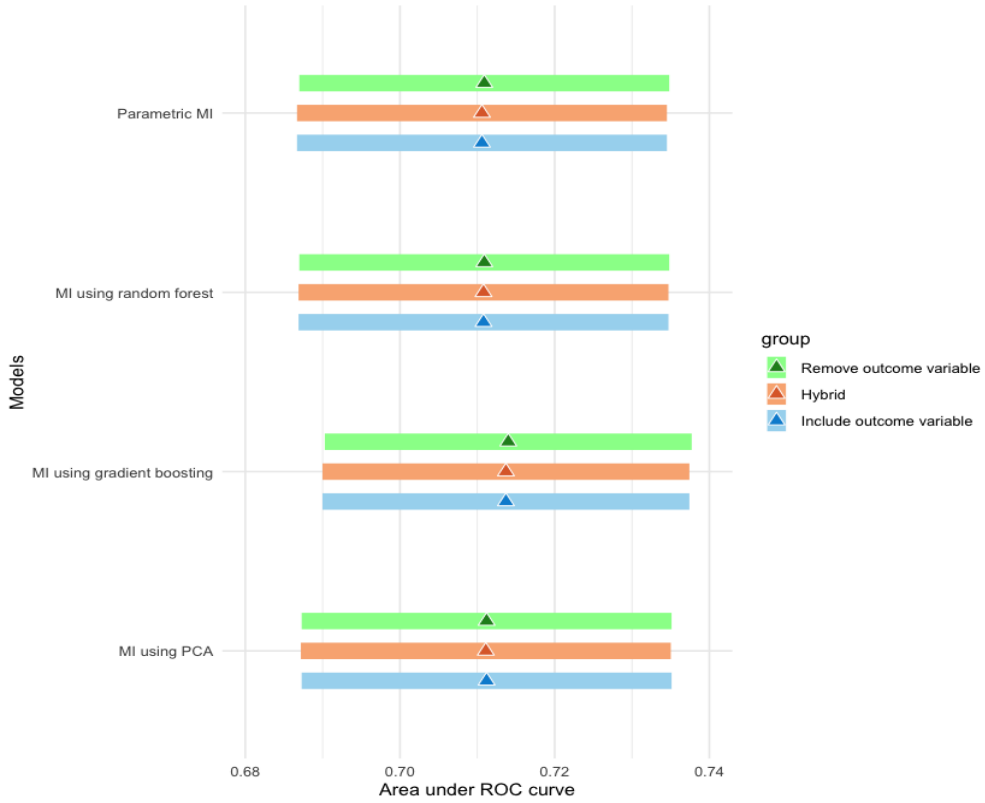
(a) E/O ratio (Outcome: FP recall)



(b) Calibration intercept (Outcome: FP recall)

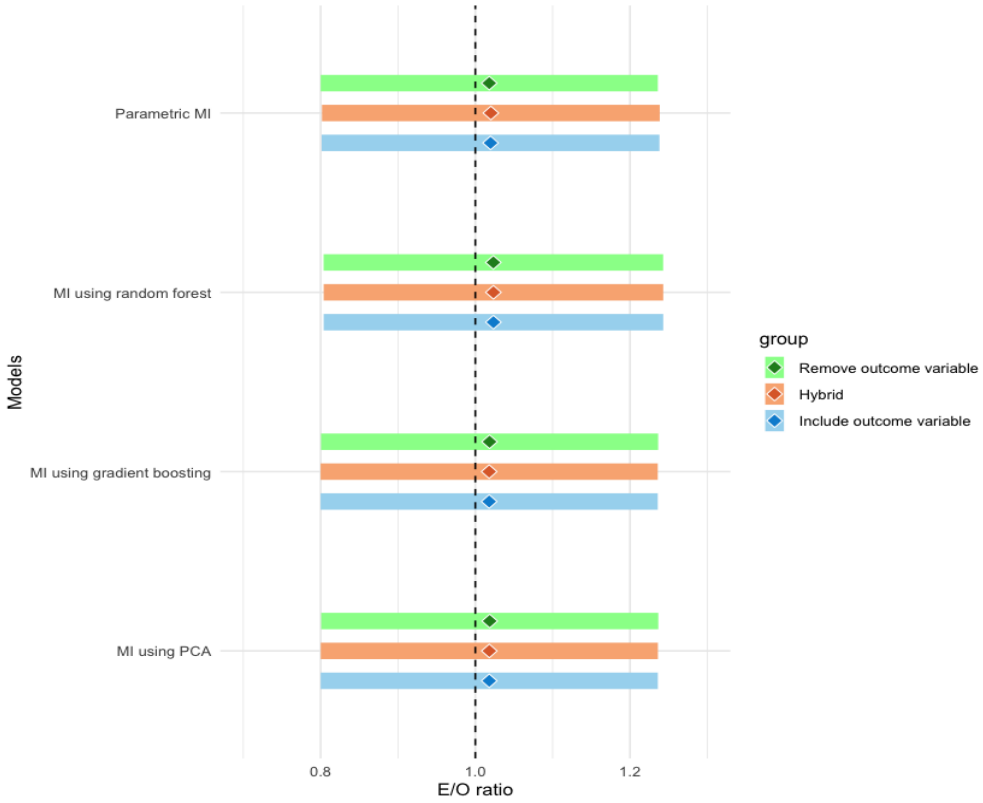


(c) Calibration slope (Outcome: FP recall)

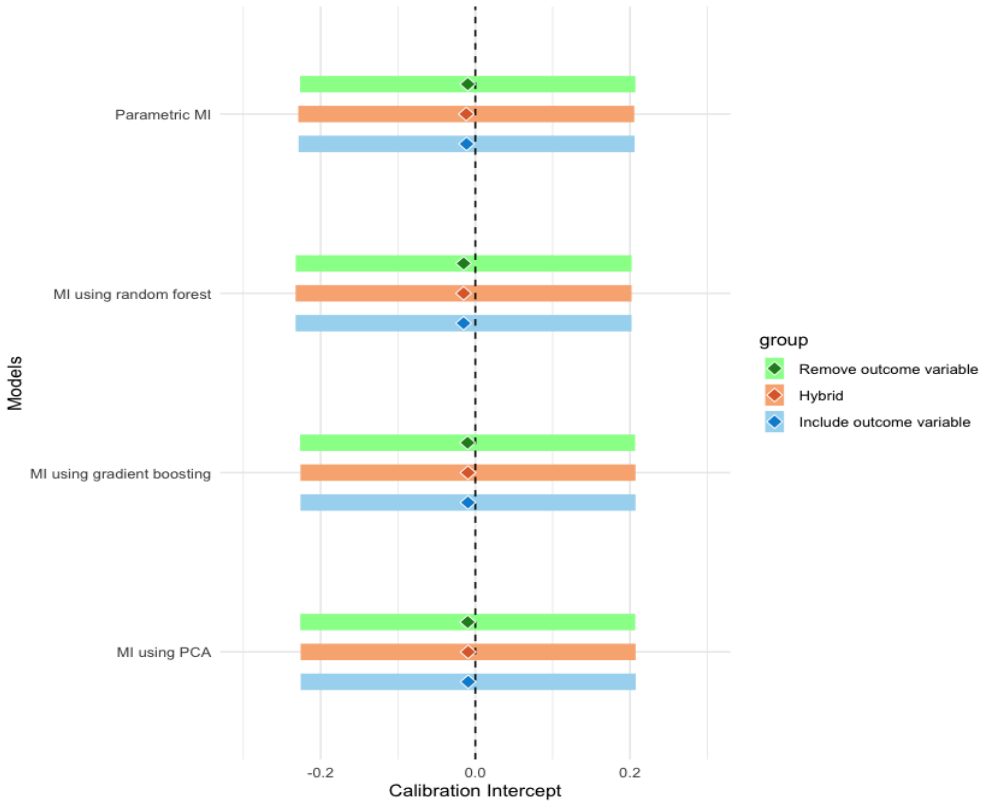


(d) AUC (Outcome: FP recall)

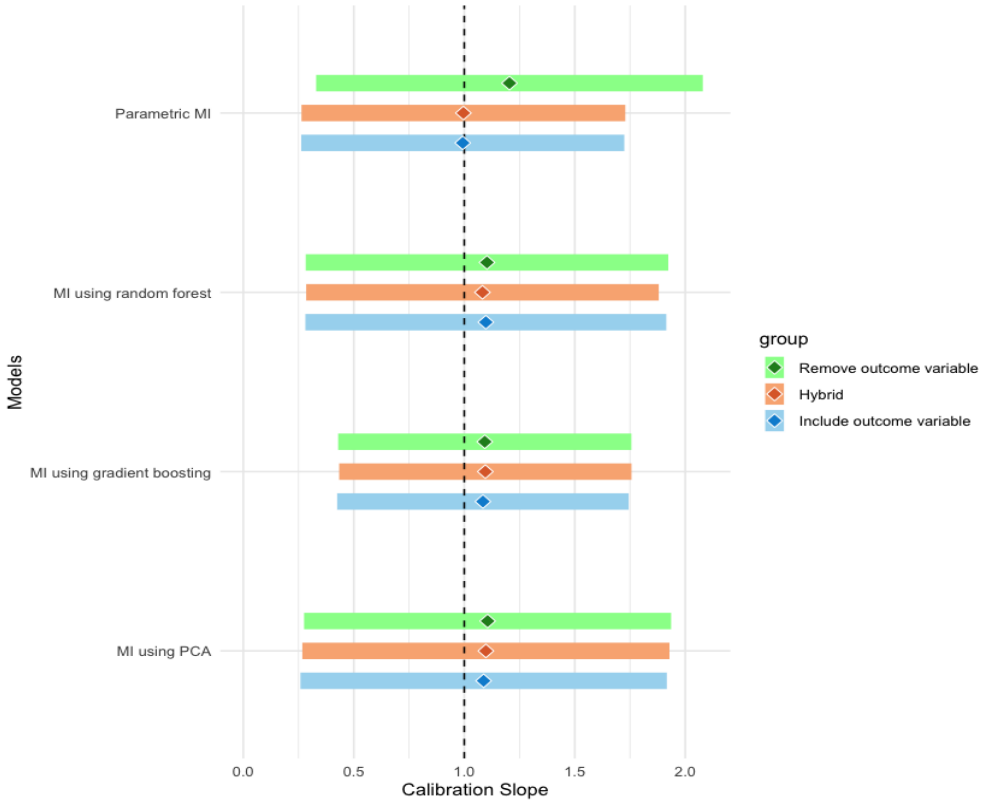
Figure 4.1: Assessment on overall model calibration and discrimination for FP recall built using Ridge Regression based on bootstrapping of BCSC data. The 95% CI for each metric was shown using the error bars. The dashed vertical lines showed the ideal value for each calibration metric (1 for E/O ratio, 0 for calibration intercept, 1 for calibration slope.



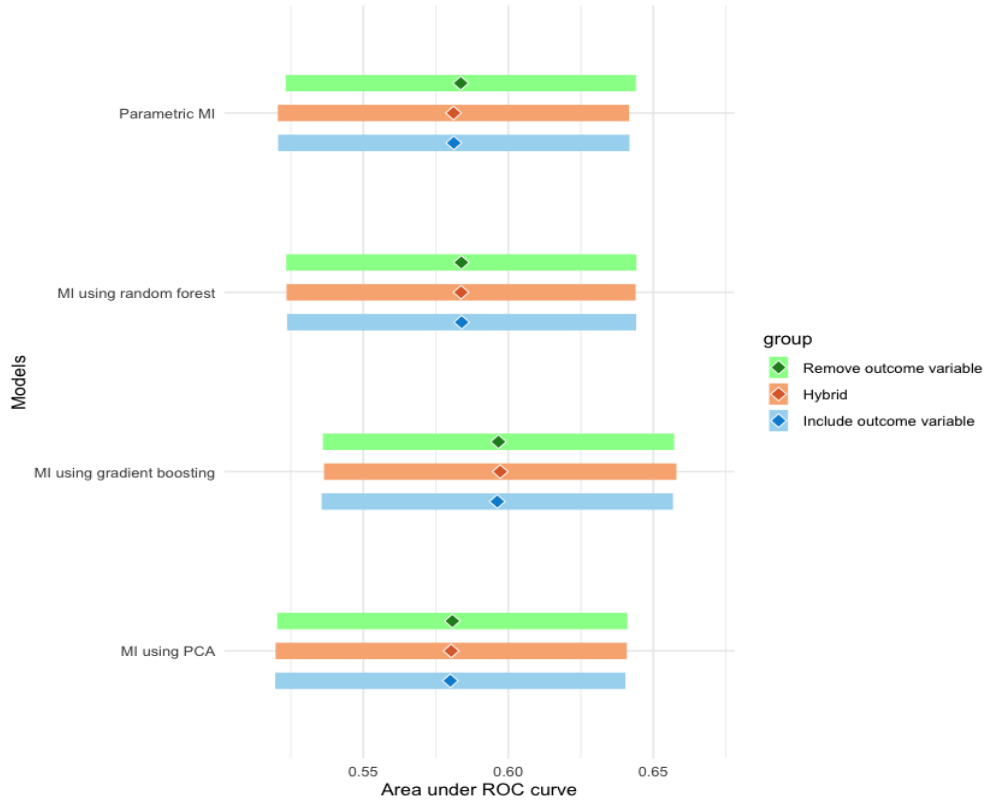
(a) E/O ratio (Outcome: FP biopsy)



(b) Calibration intercept (Outcome: FP biopsy)

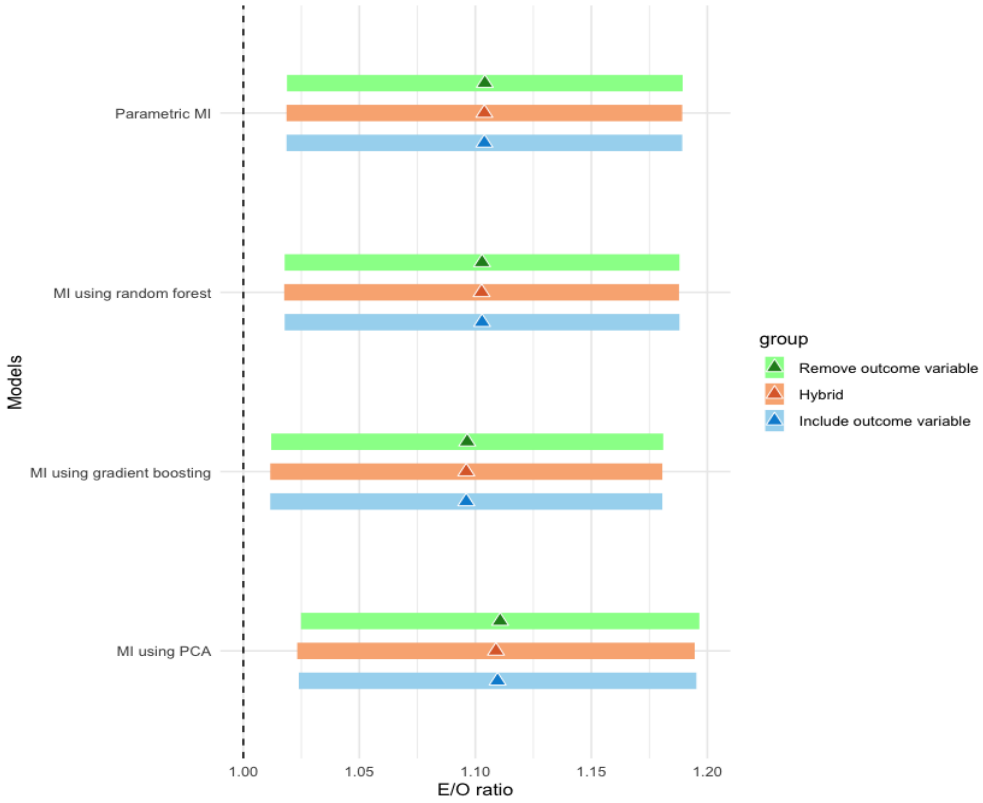


(c) Calibration slope (Outcome: FP biopsy)

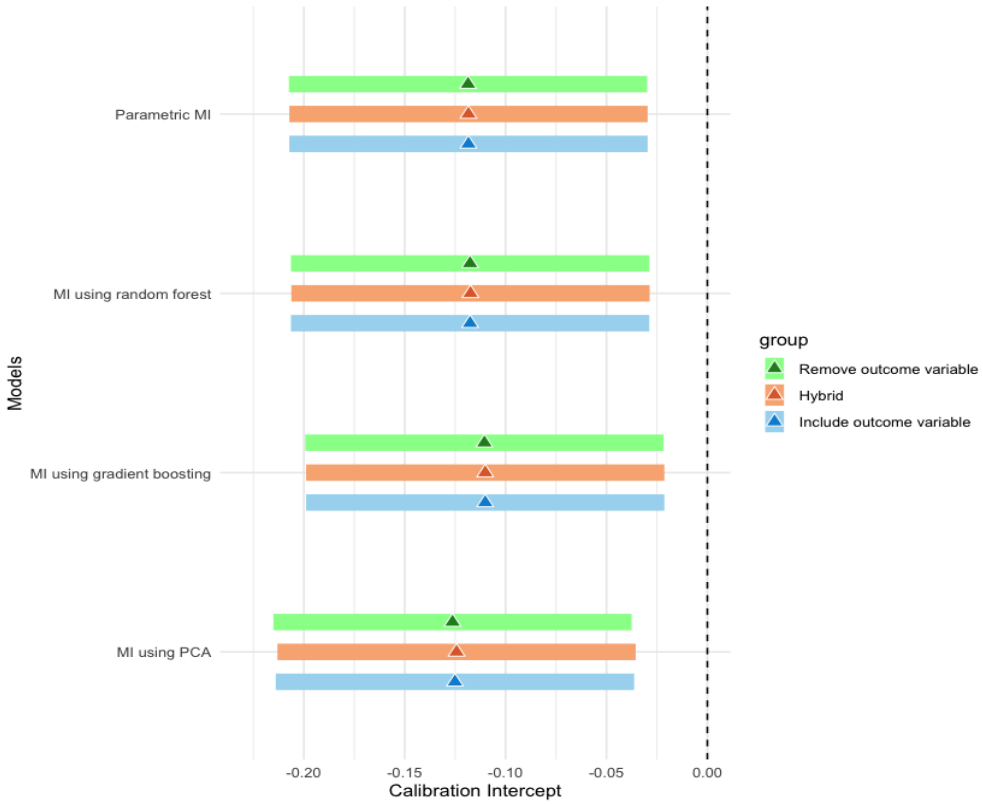


(d) AUC (Outcome: FP biopsy)

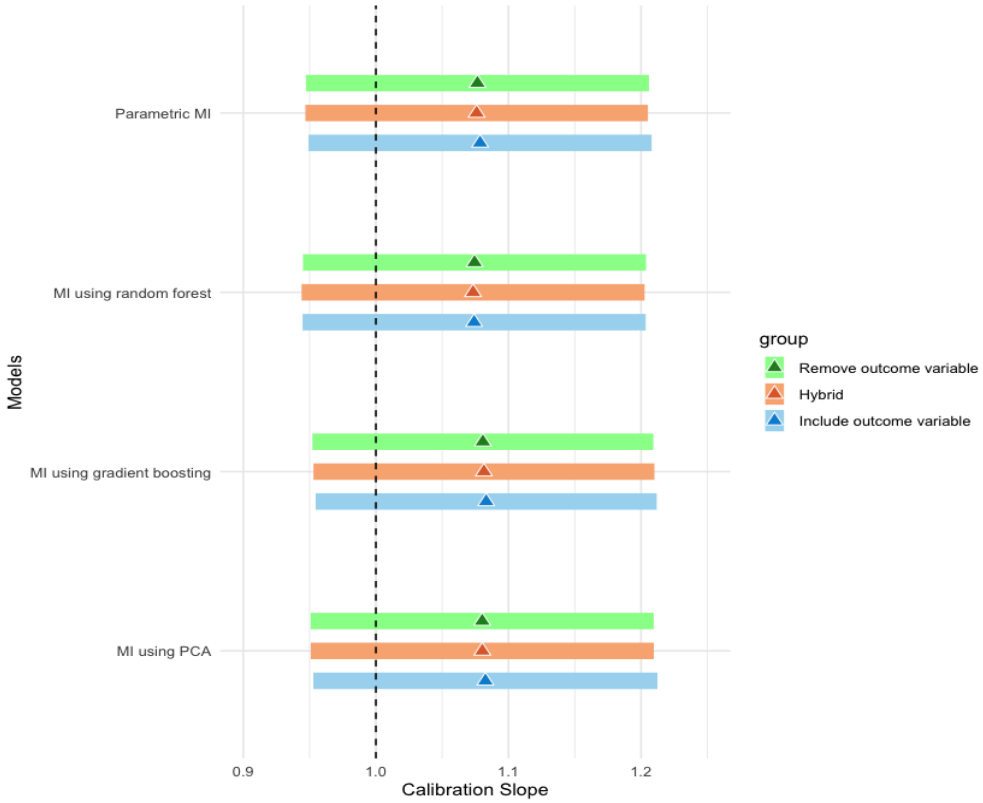
Figure 4.2: Assessment on overall model calibration and discrimination for FP biopsy built using Ridge Regression based on bootstrapping of BCSC data. The 95% CI for each metric was shown using the error bars. The dashed vertical lines showed the ideal value for each calibration metric (1 for E/O ratio, 0 for calibration intercept, 1 for calibration slope.



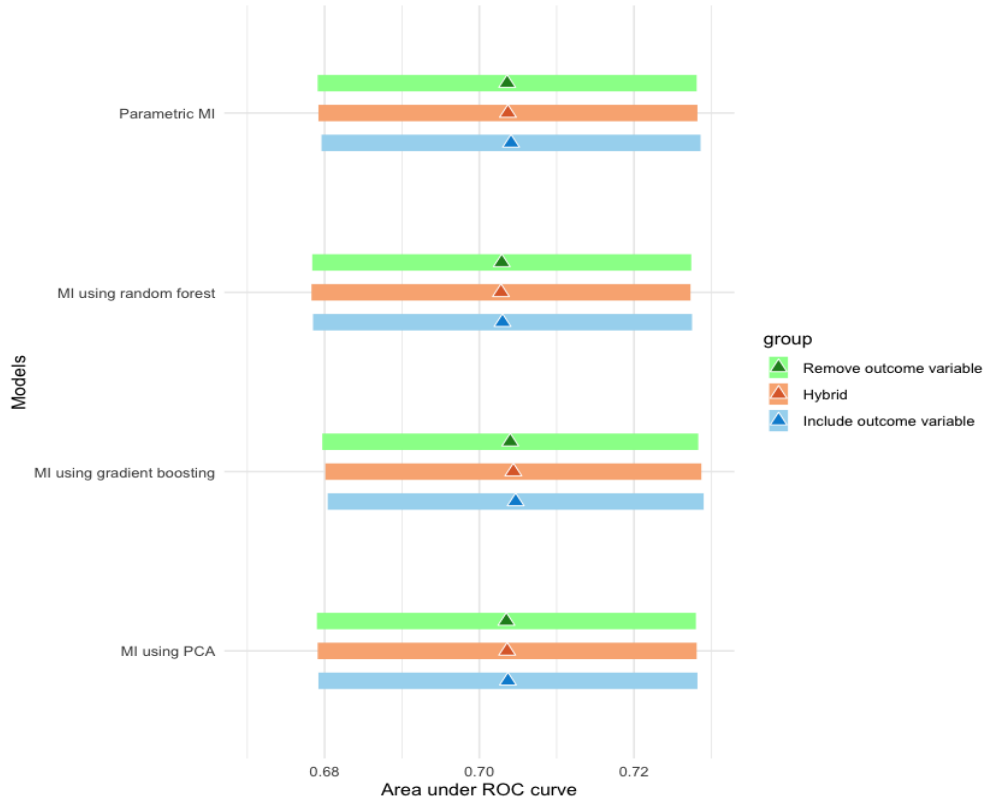
(a) E/O ratio (Outcome: FP recall)



(b) Calibration intercept (Outcome: FP recall)

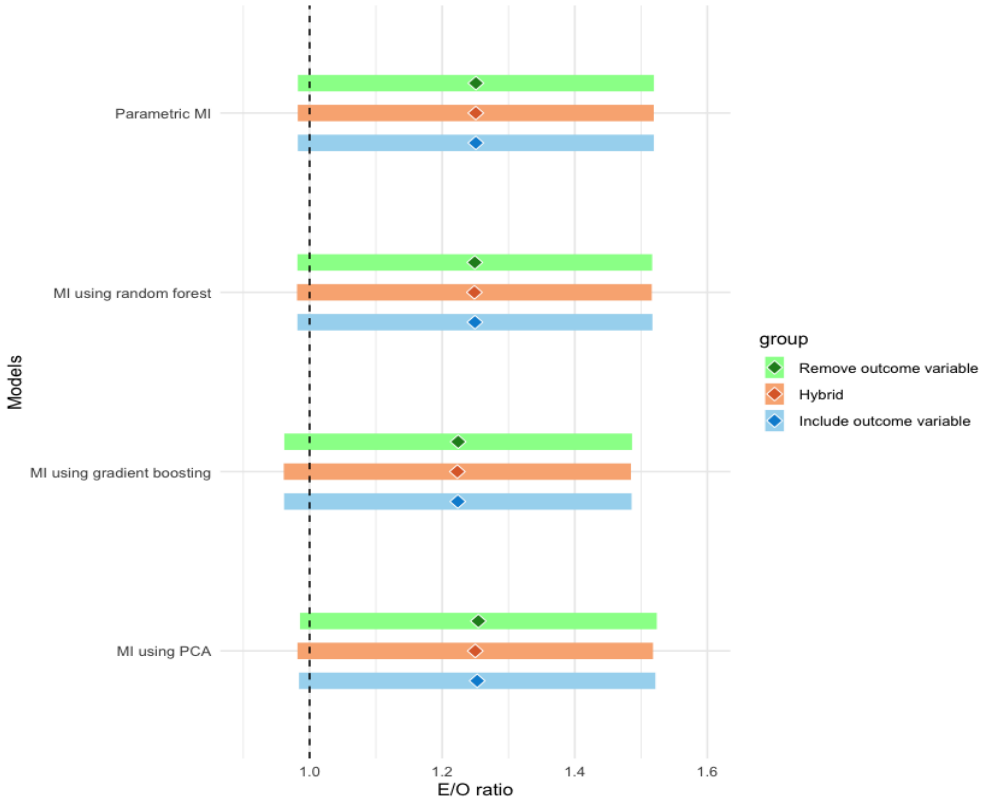


(c) Calibration slope (Outcome: FP recall)

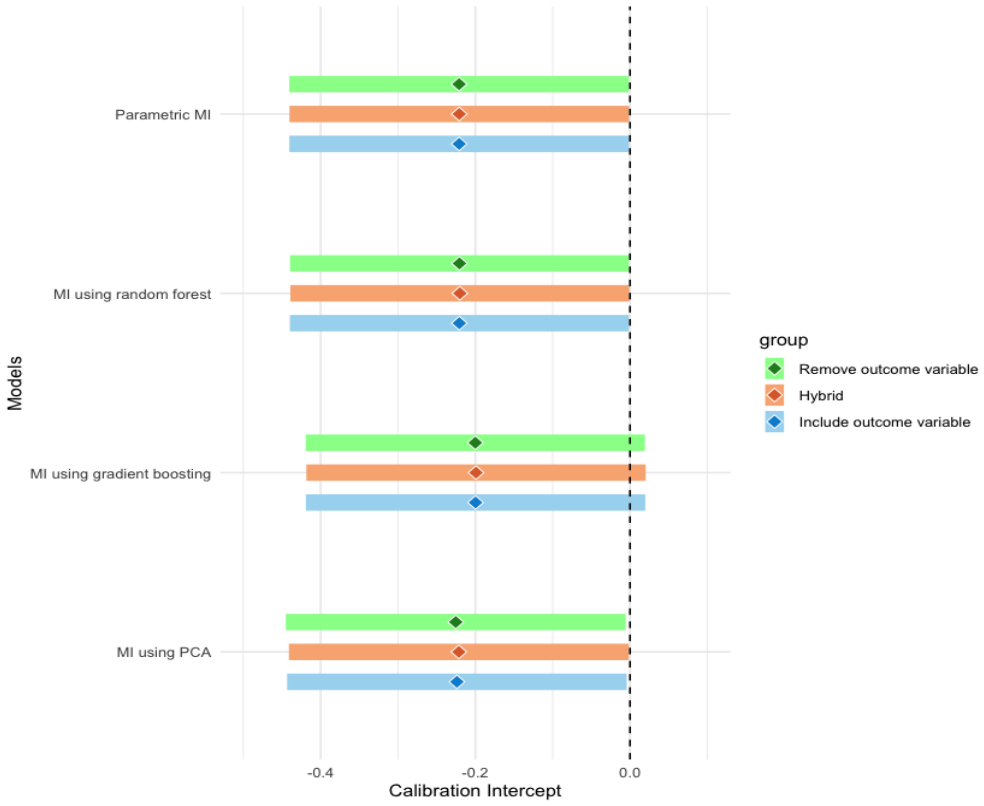


(d) AUC (Outcome: FP recall)

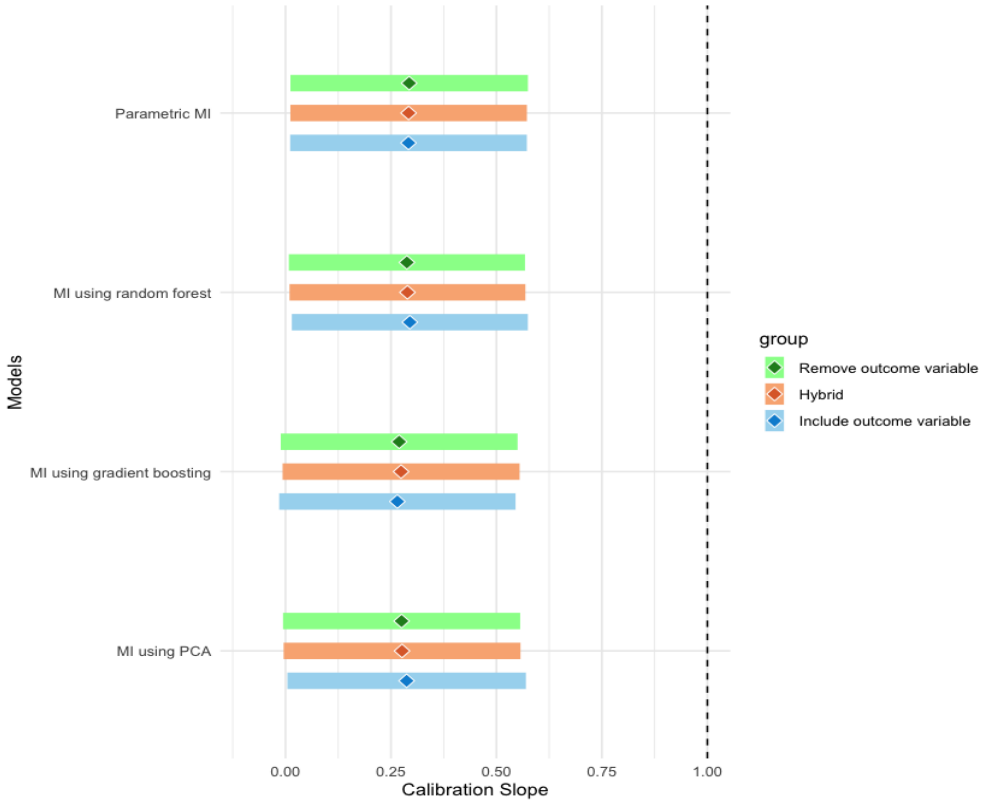
Figure 4.3: Assessment on overall model calibration and discrimination for FP recall built using Random Forest based on bootstrapping of BCSC data. The 95% CI for each metric was shown using the error bars. The dashed vertical lines showed the ideal value for each calibration metric (1 for E/O ratio, 0 for calibration intercept, 1 for calibration slope.



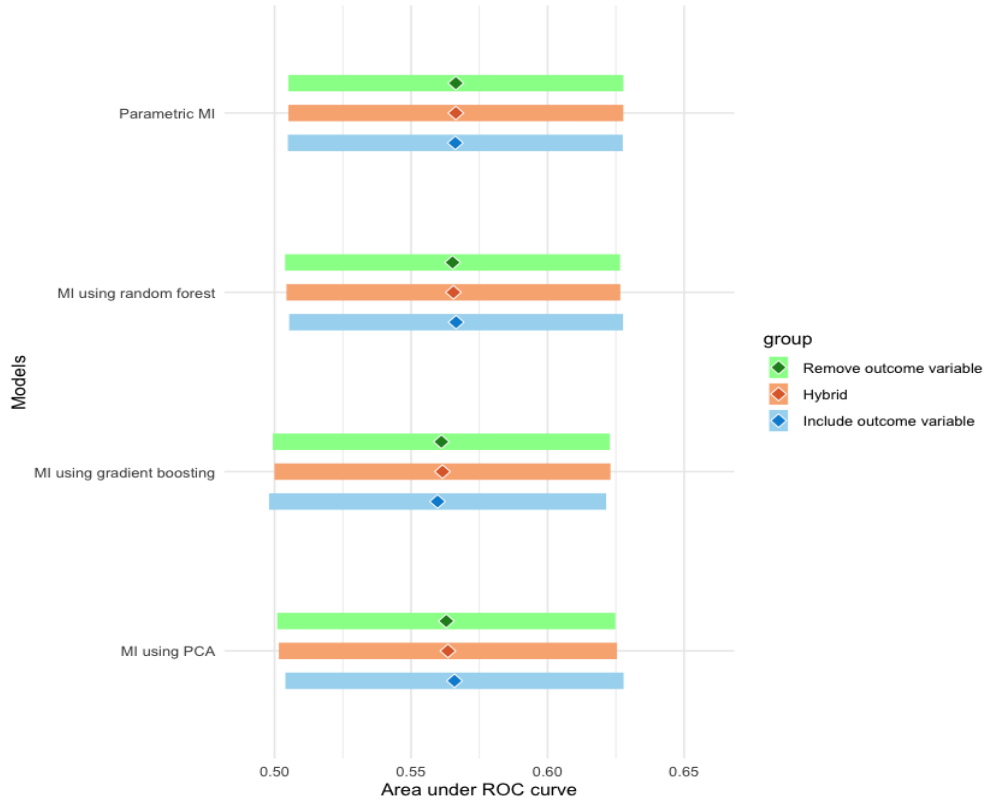
(a) E/O ratio (Outcome: FP biopsy)



(b) Calibration intercept (Outcome: FP biopsy)

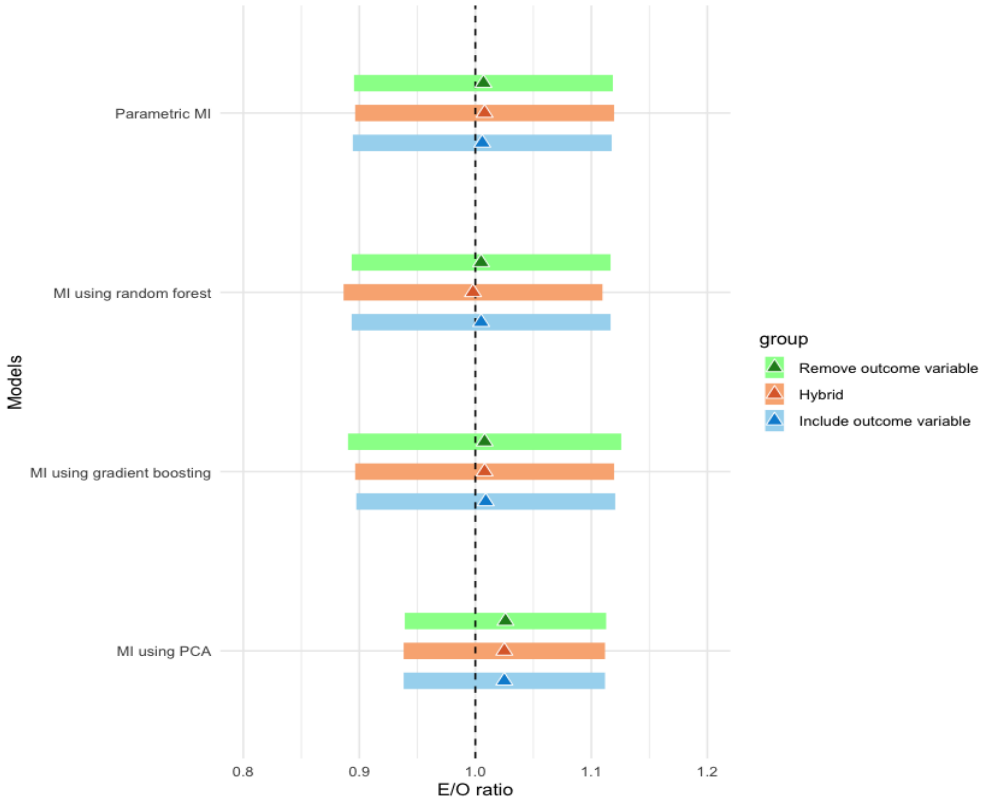


(c) Calibration slope (Outcome: FP biopsy)

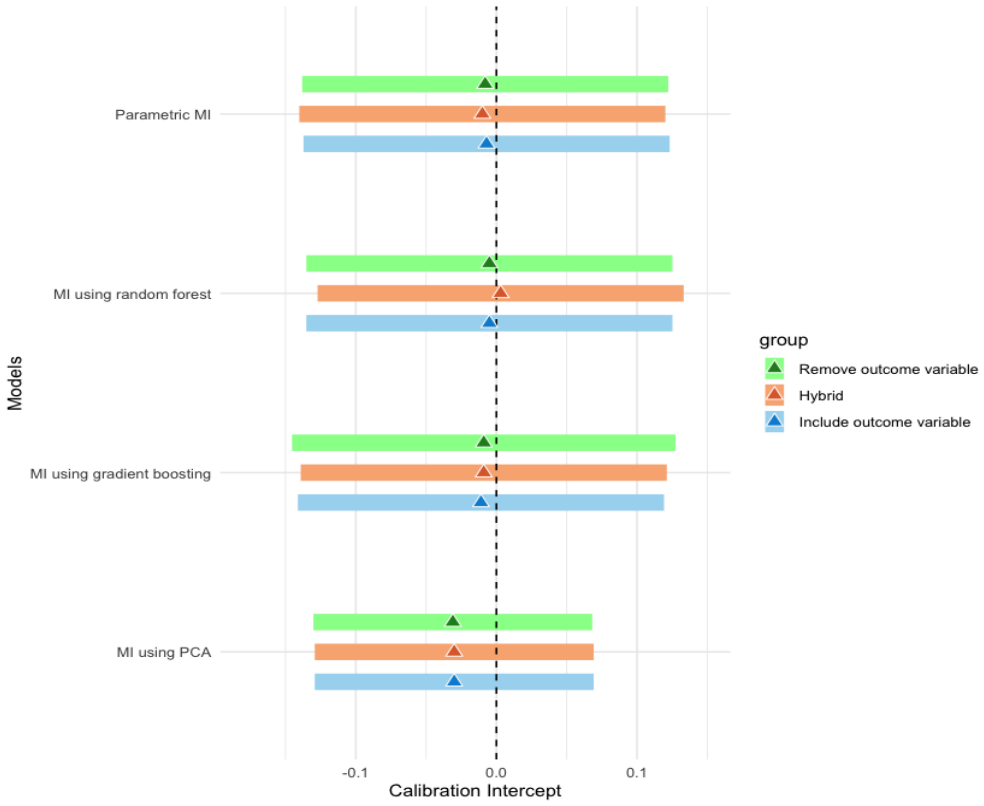


(d) AUC (Outcome: FP biopsy)

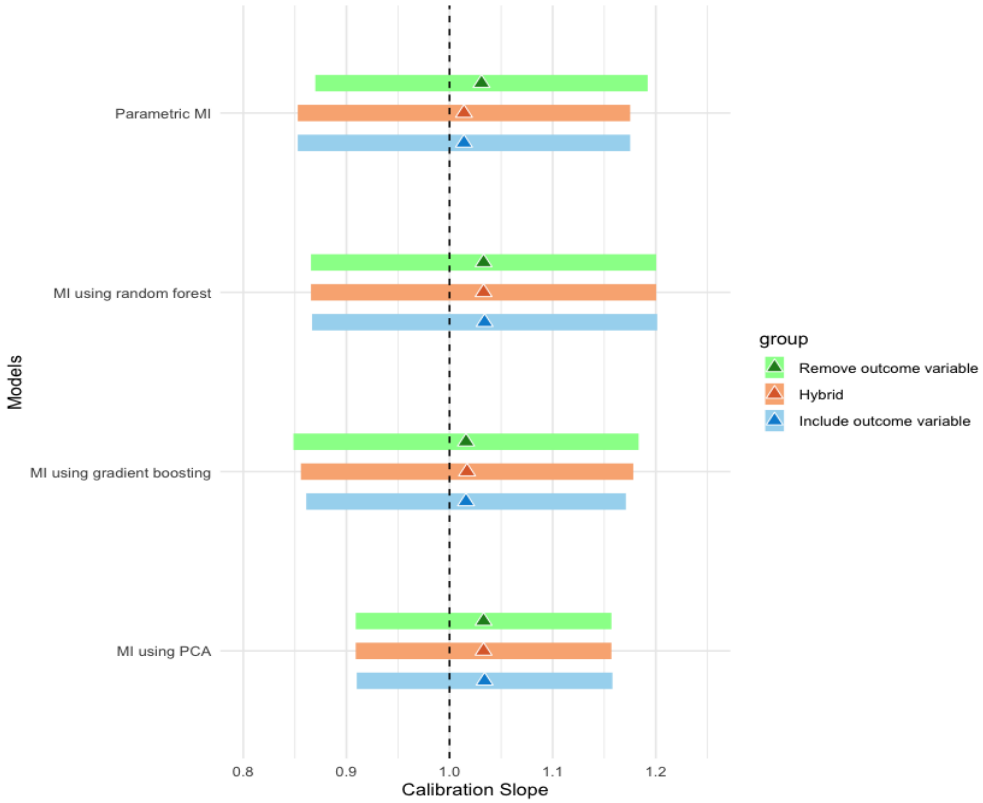
Figure 4.4: Assessment on overall model calibration and discrimination for FP biopsy built using Random Forest based on bootstrapping of BCSC data. The 95% CI for each metric was shown using the error bars. The dashed vertical lines showed the ideal value for each calibration metric (1 for E/O ratio, 0 for calibration intercept, 1 for calibration slope.



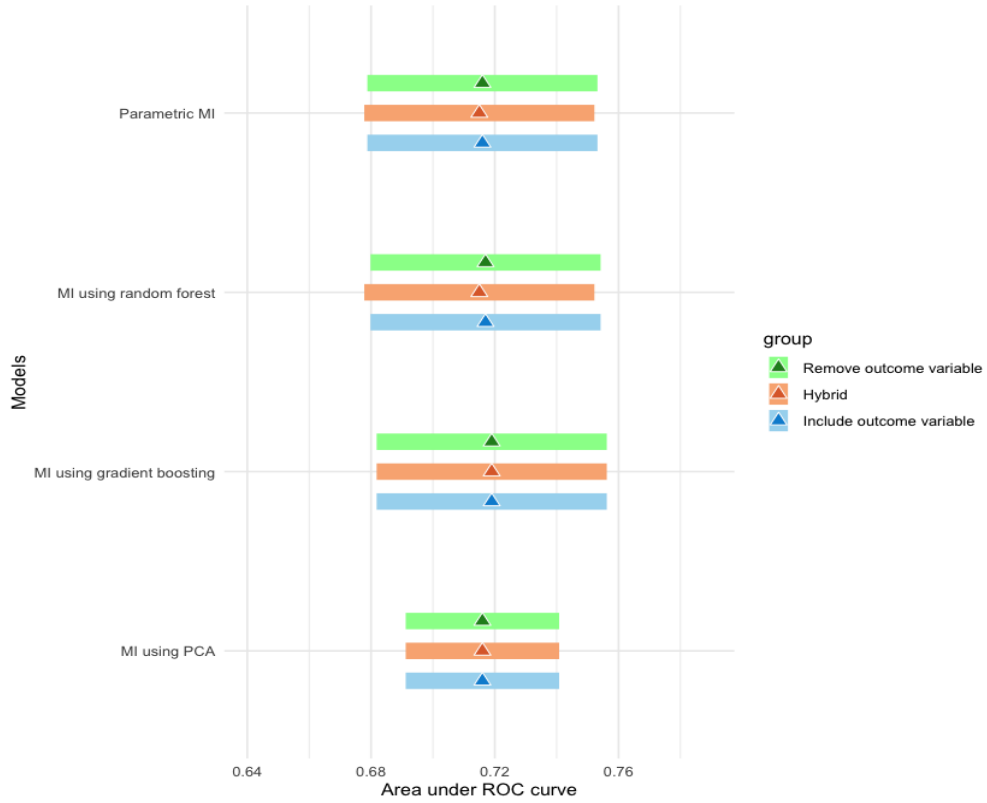
(a) E/O ratio (Outcome: FP recall)



(b) Calibration intercept (Outcome: FP recall)

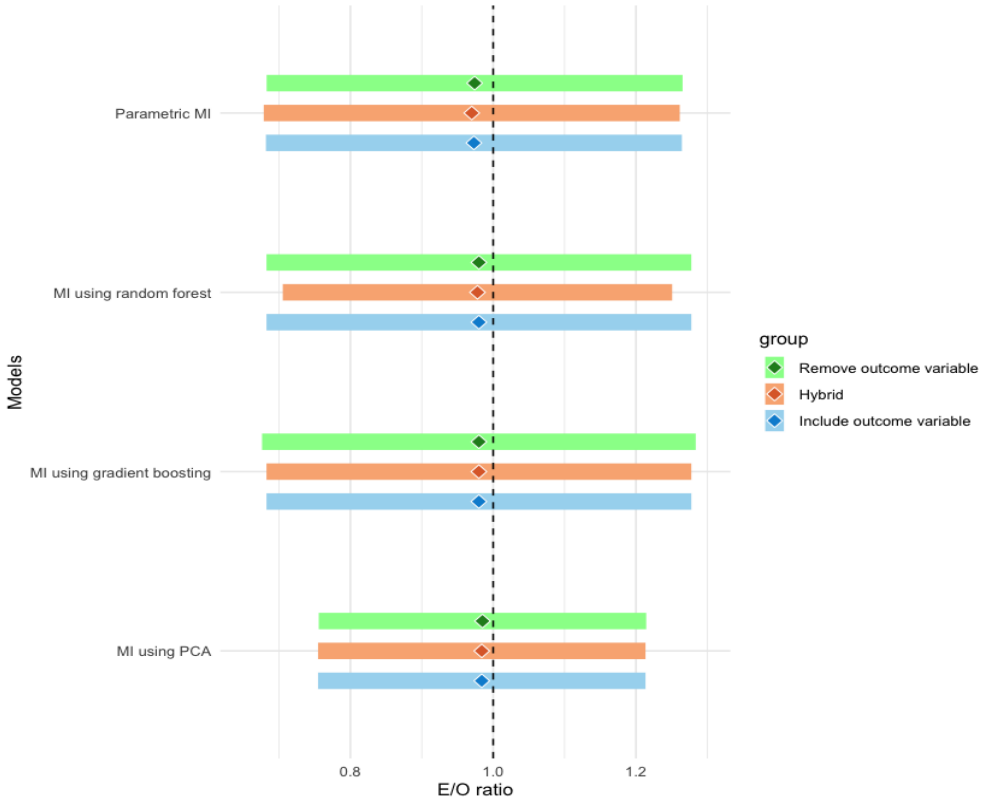


(c) Calibration slope (Outcome: FP recall)

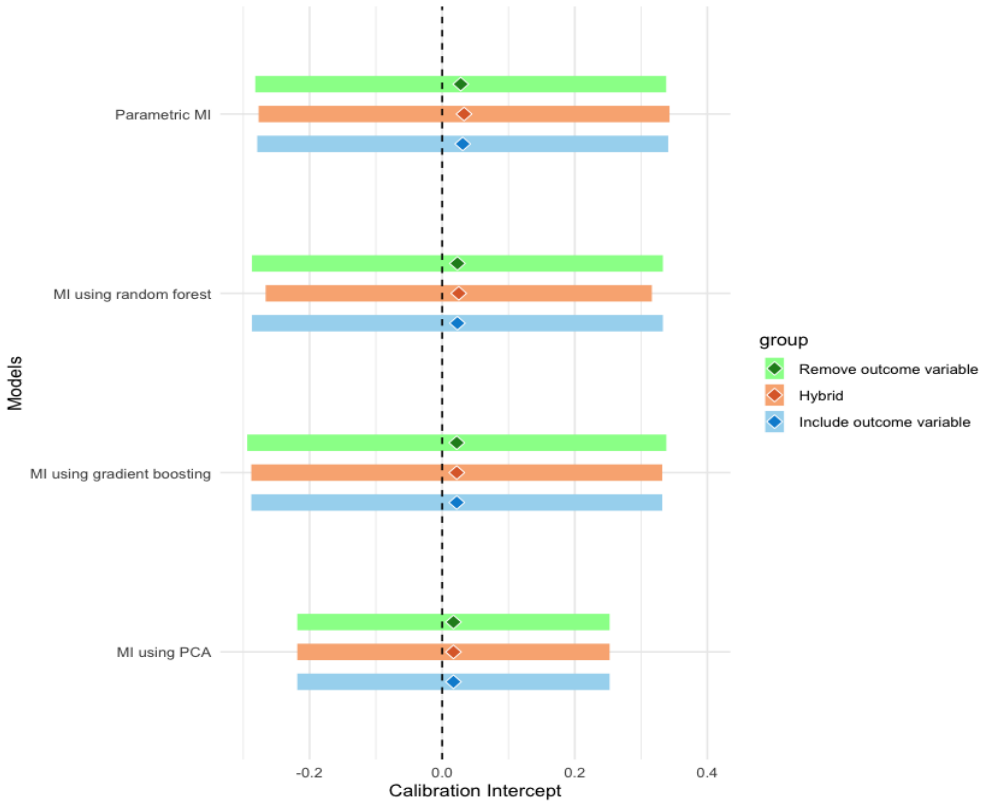


(d) AUC (Outcome: FP recall)

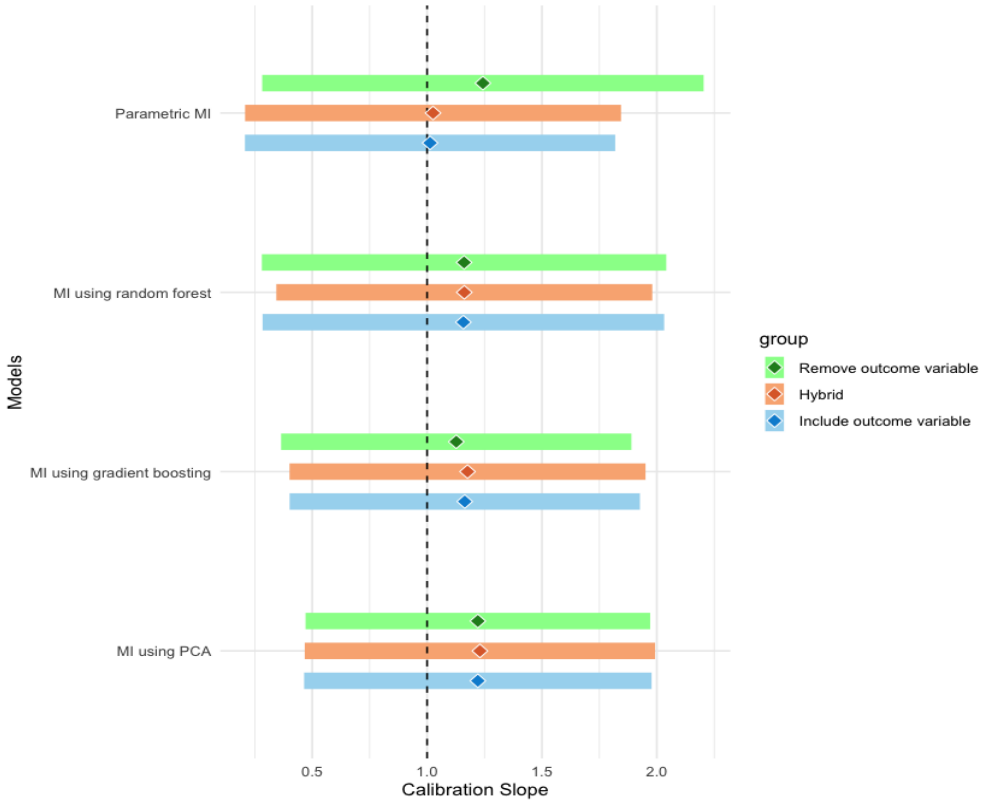
Figure 4.5: Assessment on overall model calibration and discrimination for FP recall built based on 10-fold CV. The 95% CI for each metric was shown using the error bars. The dashed vertical lines showed the ideal value for each calibration metric (1 for E/O ratio, 0 for calibration intercept, 1 for calibration slope.



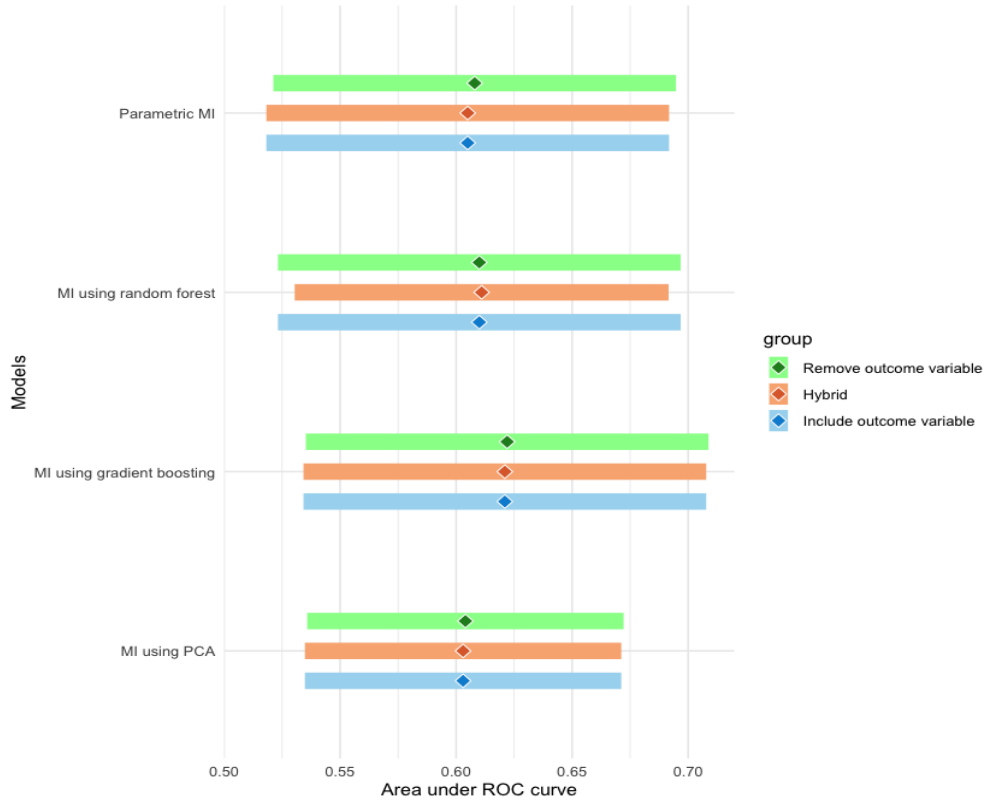
(a) E/O ratio (Outcome: FP biopsy)



(b) Calibration intercept (Outcome: FP biopsy)



(c) Calibration slope (Outcome: FP biopsy)



(d) AUC (Outcome: FP biopsy)

Figure 4.6: Assessment on overall model calibration and discrimination for FP biopsy built based on 10-fold CV. The 95% CI for each metric was shown using the error bars. The dashed vertical lines showed the ideal value for each calibration metric (1 for E/O ratio, 0 for calibration intercept, 1 for calibration slope.

Table 4.3 showed the estimated positive predictive value (PPV) and sensitivity at the 95^{th} percentile of the risk score distribution for different MI approaches under different scenarios. The results for different approaches or scenarios were not meaningfully different.

Table 4.3: Sensitivity and PPV at the 95^{th} percentile of the risk score distribution for different MI approaches under different scenarios

	Scenarios	Risk score threshold (%)*	Positive predictive value (PPV) (%)†	Sensitivity (%)‡
$\mathbf{MI_{para}}$	**no_y**	31.00	43.01	19.19
	hybrid	31.41	43.57	19.44
	with_y	31.62	43.33	19.34
$\mathbf{MI_{RF}}$	**no_y**	30.98	43.33	19.34
	hybrid	30.94	43.41	19.37
	with_y	31.01	43.33	19.34
$\mathbf{MI_{GB}}$	**no_y**	30.99	42.93	19.16
	hybrid	31.02	42.77	19.09
	with_y	31.10	42.77	19.09
$\mathbf{MI_{PCA}}$	**no_y**	31.08	43.10	18.07
	hybrid	31.04	42.89	18.68
	with_y	30.97	43.09	19.23

* Risk score threshold was determined at the 95^{th} percentile of the risk score distribution

† Positive predictive value at the 95^{th} percentile of the risk score distribution

‡ Sensitivity at the 95^{th} percentile of the risk score distribution

Chapter 5

DISCUSSION

In this study, we conducted a comprehensive comparison between parametric multiple imputation and multiple imputation methods employing machine learning techniques, such as random forest, gradient boosting, and principal component analysis via simulation studies and real-world breast cancer data. Our goal was to assess whether the use of machine learning methods for imputing missing data would yield an improved predictive performance in the risk prediction model constructed and evaluated using the imputed data. Our findings revealed that the adoption of machine learning-based imputation methods did not lead to superior model performance compared to traditional parametric imputation. Previous studies have highlighted the advantages of machine learning-based imputation methods, emphasizing their ability to capture complex relationships between variables and handle higher-order effects effectively. Some works even reported substantial predictive performance improvements when using machine learning methods for imputing missing data in contexts with not-so-rare outcomes, low-dimensional features and big sample sizes [26]. However, our study suggested that the gains in predictive performance achieved through imputation with machine learning approaches might not be as noticeable, at least in some clinical scenarios, as initially thought. It is crucial to consider that the success of machine learning techniques heavily relies on data characteristics, sample size, and the complexity of the underlying relationships in the dataset. In the context of risk prediction models, the impact of imputation methods on predictive performance might be further complex, influenced by the nature of the risk factors and their interactions in the specific domain. In settings with moderate sample sizes, modest number of predictors, and infrequent outcomes, the predictive performance of a risk prediction model may not benefit from using machine-learning MI compared to parametric

MI.

In addition to comparing parametric and machine-learning imputation methods, we investigated the impact of including the outcome variable during MI in the training and test sets on the predictive performance of the risk prediction model. For discriminatory accuracy, our findings indicated that including the outcome variable in the test set during the imputation process resulted in higher AUCs for the risk prediction model compared to removing the outcome variable in the test set, especially for data with only a few predictors. However, we observed that some AUCs derived from the imputation with outcomes in the test sets were even greater than the AUCs obtained from the complete dataset. This observation raises concerns about potential overly optimistic AUC estimates when incorporating outcomes in MI for the test set. Including the outcome variable in the training set can help capture relationships between the predictor variables and the outcome, leading to more accurate imputed values [32, 43]. However, missing data can occur at any stage of the clinical prediction modeling pipeline, including model development, validation, or deployment. If data is allowed to be missing at model deployment in clinical practice, it is recommended that the outcome variable should be omitted from the imputation model during the development phase [39] to avoid potential biases and overly optimistic AUC estimates.

Foe calibration, all MI methods under three scenarios resulted in well-calibrated predictions in high-dimensional settings (95% CIs for E/O ratios, calibration intercepts, and slopes overlapping 1, 0, and 1, respectively). However, in low-dimensional settings, we observed that using the hybrid approach (where we included the outcome variable in the imputation process in the training set, but omitted it in the test set) led to poor calibration slopes (significantly less than 1) for all MI methods, indicating overfitting. One possible reason for this may be model overdispersion, especially when we only have a few predictors. We fitted prediction models on the imputed training set (outcome variable was included in the imputation model) and the imputed test set (outcome variable was omitted in the imputation model) separately and then compared the variance of predicted risks obtained from the two models. We got a higher variation in the predicted risks when we used the training set

to fit the model, compared to that fitted using the test set. However, we did not observe meaningful differences of variance in the high dimensional settings. These observations may suggest that the hybrid approach may not be suitable in low-dimensional settings.

We also investigated the impacts of model misspecification of higher-order effects during MI on predictive performance. We did not observe substantial changes in predictive performance when we misspecified any interaction terms in our simulations. However, it is essential to exercise caution in drawing general conclusions from these findings, as the simulations we conducted represent a relatively simple scenario. More complex real-world data may exhibit different patterns and sensitivities to model misspecification. Therefore, while our initial results suggest that misspecification of higher-order effects may not have an appreciable impact on predictive performance in this specific setting, further investigation with diverse and more intricate datasets is necessary to gain a comprehensive understanding of the potential consequences of model misspecification during MI in risk prediction modeling.

Although my thesis is focused on comparing different multiple imputation approaches, we also found that different selected risk prediction models could impact model performance, as suggested in some prior works [45, 20, 53, 10]. Our findings reveal intriguing differences between Random Forest and Ridge Regression utilized to build risk models. Random Forest exhibited E/O ratios significantly greater than 1, indicating a tendency to overestimate risks, and calibration intercepts significantly less than 1, further reinforcing the notion of overestimation. Additionally, the calibration slopes significantly less than 1 suggest that the Random Forest model was prone to overfitting the data, leading to suboptimal predictions. Machine learning approaches are sensitive to the choice of hyperparameters. In this case, we tried to expand the search for tuning parameters in the simulation studies by (1) expand the range of `mtry` values (we also considered $0.25\sqrt{p}$ and $4\sqrt{p}$ apart from originally $0.5\sqrt{p}$, $\sqrt{p}$ and $2\sqrt{p}$, where p is the number of predictors); (2) add another hyperparameter to tune: the minimal size of terminal nodes (we considered 5, 10, 20 and 30). However, we did not get any meaningful improvement in model performance. To our knowledge, we followed best practices for tuning hyperparameters using cross-validation, although results could vary if

alternative hyperparameters were selected. Another possible reason for Random Forest not performing well can be that it suffers from class imbalance in our data (e.g. the rate of FP biopsy was 1%). Some resampling techniques (e.g. over-sampling, under-sampling, SMOTE) [16] could be applied in future work.

In conclusion, machine learning-based MI didn't outperform traditional parametric MI in terms of predictive performance. We recommend whether to use machine learning-based MI should be carefully evaluated based on the specific context and characteristics of the data. Moreover, we recommend against including the outcome variable in the imputation model for the test set. Although including outcome in the test set during MI led to higher AUCs but raised concerns of over-optimistic predictive performance. Finally, we recommend being cautious of using Random Forest as the risk prediction model for similar prediction modeling settings.

BIBLIOGRAPHY

[1] Todd A Alonzo. Clinical prediction models: a practical approach to development, validation, and updating: by ewout w. steyerberg, 2009.

[2] Mahul B Amin, Stephen B Edge, Frederick L Greene, David R Byrd, Robert K Brookland, Mary Kay Washington, Jeffrey E Gershenwald, Carolyn C Compton, Kenneth R Hess, Daniel C Sullivan, et al. *AJCC cancer staging manual*, volume 1024. Springer, 2017.

[3] Vincent Audigier, François Husson, and Julie Josse. Multiple imputation for continuous variables using a bayesian principal component analysis. *Journal of statistical computation and simulation*, 86(11):2140–2156, 2016.

[4] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

[5] Jonathan W Bartlett and Tim P Morris. Multiple imputation of covariates by substantive-model compatible fully conditional specification. *The Stata Journal*, 15(2):437–456, 2015.

[6] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[7] Diana SM Buist. Factors to consider in developing breast cancer risk models to implement into clinical care. *Current epidemiology reports*, 7:113–116, 2020.

[8] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[10] Evangelia Christodoulou, Jie Ma, Gary S Collins, Ewout W Steyerberg, Jan Y Verbakel, and Ben Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, 110:12–22, 2019.

[11] Serkalem Demissie, Michael P LaValley, Nicholas J Horton, Robert J Glynn, and L Adrienne Cupples. Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. *Statistics in medicine*, 22(4):545–557, 2003.

[12] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[13] Yongshi Deng and Thomas Lumley. Multiple imputation through xgboost. *arXiv preprint arXiv:2106.01574*, 2021.

[14] James D Dziura, Lori A Post, Qing Zhao, Zhixuan Fu, and Peter Peduzzi. Strategies for dealing with missing data in clinical trials: from design to analysis. *The Yale journal of biology and medicine*, 86(3):343, 2013.

[15] Craig K Enders. *Applied missing data analysis*. Guilford Publications, 2022.

[16] Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20(1):18–36, 2004.

[17] Jessica M Franklin, Sebastian Schneeweiss, Jennifer M Polinski, and Jeremy A Rassen. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational statistics & data analysis*, 72:219–226, 2014.

[18] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[19] Jelle J Goeman. L1 penalized estimation in the cox proportional hazards model. *Biometrical journal*, 52(1):70–84, 2010.

[20] Benjamin Y Gravesteijn, Daan Nieboer, Ari Ercole, Hester F Lingsma, David Nelson, Ben Van Calster, Ewout W Steyerberg, Cecilia Åkerlund, Krisztina Amrein, Nada Andelic, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *Journal of clinical epidemiology*, 122:95–107, 2020.

[21] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[22] Jorgen Hilden, J Dik F Habbema, and Beth Bjerregaard. The measurement of performance in probabilistic diagnosis. *Methods of information in medicine*, 17(04):227–237, 1978.

[23] Shangzhi Hong and Henry S Lynn. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC medical research methodology*, 20(1):1–12, 2020.

[24] Bahram Jafrasteh, Daniel Hernández-Lobato, Simón Pedro Lubián-López, and Isabel Benavente-Fernández. Gaussian processes for missing value imputation. *arXiv preprint arXiv:2204.04648*, 2022.

[25] Mortaza Jamshidian and Peter M Bentler. Ml estimation of mean and covariance structures with missing data using complete data routines. *Journal of Educational and behavioral Statistics*, 24(1):21–24, 1999.

[26] José M Jerez, Ignacio Molina, Pedro J García-Laencina, Emilio Alba, Nuria Ribelles, Miguel Martín, and Leonardo Franco. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, 50(2):105–115, 2010.

[27] Julie Josse and François Husson. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 153(2):79–99, 2012.

[28] Julie Josse, Jérôme Pagès, and François Husson. Multiple imputation in principal component analysis. *Advances in data analysis and classification*, 5:231–246, 2011.

[29] Hyun Kang. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5):402–406, 2013.

[30] Peng Li, Elizabeth A Stuart, and David B Allison. Multiple imputation: a flexible tool for handling missing data. *Jama*, 314(18):1966–1967, 2015.

[31] Roderick JA Little. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287–296, 1988.

[32] Karel GM Moons, Rogier ART Donders, Theo Stijnen, and Frank E Harrell Jr. Using the outcome for imputation of missing predictor values was preferred. *Journal of clinical epidemiology*, 59(10):1092–1101, 2006.

[33] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, 2013.

[34] Tapani Raiko, Alexander Ilin, and Juha Karhunen. Principal component analysis for large scale problems with lots of missing values. In *European Conference on Machine Learning*, pages 691–698. Springer, 2007.

[35] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[36] Donald B Rubin. Multiple imputation for survey nonresponse, 1987.

[37] Anoop D Shah, Jonathan W Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway. Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study. *American journal of epidemiology*, 179(6):764–774, 2014.

[38] Rose Sisk, Matthew Sperrin, Niels Peek, Maarten van Smeden, and Glen P Martin. Imputation and missing indicators for handling missing data in the development and implementation of clinical prediction models: a simulation study. *arXiv preprint arXiv:2206.12295*, 2022.

[39] Rose Sisk, Matthew Sperrin, Niels Peek, Maarten van Smeden, and Glen Philip Martin. Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: A simulation study¡? pag\vspace*{-6pt}?¿. *Statistical Methods in Medical Research*, page 09622802231165001, 2023.

[40] Alexander J Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans. Gaussian processes and svm: Mean field and leave-one-out. 2000.

[41] N Solaro, A Barbiero, G Manzi, and PA Ferrari. A simulation comparison of imputation methods for quantitative data in the presence of multiple data patterns. *Journal of Statistical Computation and Simulation*, 88(18):3588–3619, 2018.

[42] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[43] Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338, 2009.

[44] Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010.

[45] Yu-Ru Su, Diana SM Buist, Janie M Lee, Laura Ichikawa, Diana L Miglioretti, Erin J Aiello Bowles, Karen J Wernli, Karla Kerlikowske, Anna Tosteson, Kathryn P Lowry,

et al. Performance of statistical and machine learning risk prediction models for surveillance benefits and failures in breast cancer survivors. *Cancer Epidemiology, Biomarkers & Prevention*, 32(4):561–571, 2023.

[46] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

[47] Stef Van Buuren and Catharina GM Oudshoorn. Multivariate imputation by chained equations, 2000.

[48] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[49] Paul T Von Hippel. 8. how to impute interactions, squares, and other transformed variables. *Sociological methodology*, 39(1):265–291, 2009.

[50] Levi Waldron, Melania Pintilie, Ming-Sound Tsao, Frances A Shepherd, Curtis Huttenhower, and Igor Jurisica. Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics*, 27(24):3399–3406, 2011.

[51] Akbar K Waljee, Ashin Mukherjee, Amit G Singal, Yiwei Zhang, Jeffrey Warren, Ulysses Balis, Jorge Marrero, Ji Zhu, and Peter DR Higgins. Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*, 3(8):e002847, 2013.

[52] Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399, 2011.

[53] Annemieke Witteveen, Gabriela F Nane, Ingrid MH Vliegen, Sabine Siesling, and Maarten J IJzerman. Comparison of logistic regression and bayesian networks for risk prediction of breast cancer recurrence. *Medical decision making*, 38(7):822–833, 2018.

[54] Muhan Zhou, Yulei He, Mandi Yu, and Chiu-Hsieh Hsu. A nonparametric multiple imputation approach for missing categorical data. *BMC medical research methodology*, 17(1):1–12, 2017.

Appendix A

VERSION FOR R PACKAGES

- R (4.2.0)
- ROCR (1.0-11)
- glmnet (4.1-6)
- pensim (1.3.6)
- tidyr (1.2.1)
- dplyr (1.0.10)
- ggplot2 (3.4.0)
- caret (6.0-94)
- VIM (6.2.2)
- mice (3.15.0)
- randomForest (4.7-1.1)
- mixgb (0.1.0)
- missMDA (1.18)
- sandwich (3.0-2)

Appendix B

SAMPLE CODE

```
# Code for simulating high dimensional datasets
library(pensim)
x <- create.data(
    nvars = c(20, 16, 8, 20),
    cors = c(0, 0.7, 0.7, 0),
    associations = c(0.5, 0.6, 0.3, 0.2),
    firstonly = c(TRUE, TRUE, TRUE, FALSE),
    nsamples = 1000,
    response = "binary",
    logisticintercept = -2.8)
data_5 <- x$data
table(data_5$outcome)
outcome = data_5$outcome

# Induce missing data in predictors during simulation studies
# (take high dimensional setting as an example)
data_mar <- data.frame(cbind(ampute(data[1:10], mech="MAR", prop = 0.1,
↪  bycases = FALSE)$amp, data[11:29], ampute(data[30:45], mech="MAR", prop
↪  = 0.2, bycases = TRUE)$amp, data[46:65]))


# Different approaches to imputing missing data in predictors
```

```r
# (take high dimensional setting as an example)
library(mice)
data_para_mar_no_y <- mice(data_mar[1:64], m=5) # parametric MI, no outcome
data_para_mar <- mice(data_mar, m=5) # parametric MI, with outcome

library(mice)
data_rf_mar_no_y <- mice(data_mar[1:64], m=5, meth="rf") # MI using random
↪ forest, no outcome
data_rf_mar <- mice(data_mar, m=5, meth="rf") # MI using random forest,
↪ with outcome

library(mixgb)
data_imp_gb_mar_no_y <- mixgb(data_mar[1:64], m=5) # MI using gradient
↪ boosting, no outcome
data_imp_gb_mar <- mixgb(data_mar, m=5) # MI using gradient boosting, with
↪ outcome

library(missMDA)
## MI using PCA, no outcome
cv.nb_data_mar_no_y <- estim_ncpFAMD(data_mar, ncp.min=1, ncp.max=5,
↪ method.cv="kfold")
data_imp_pca_mar_no_y <- MIPCA(data_mar[1:64], ncp=cv.nb_data_mar_no_y$ncp,
↪ nboot=5)
## MI using PCA, with outcome
cv.nb_data_mar <- estim_ncpFAMD(data_mar, ncp.min=1, ncp.max=5,
↪ method.cv="kfold")
data_imp_pca_mar <- MIFAMD(data_mar,ncp=cv.nb_data_mar$ncp, nboot=5)
```

```r
# Functions to evaluate risk model using imputed datasets
# (take low dimensional setting as an example)
predict_fun_cat <- function(x, fits) {
  predict(fits, newdata=x, type="response")
}

calibrationMetrics = function(data,risk_var){
  print(dim(data))
  data$risk_interest = data[,risk_var]

  ## EO ratio:
  E = mean(data$risk_interest)
  O = mean(data$outcome)
  EORatio = E/O
  library(sandwich)
  glm_fit = glm(formula = outcome~1, family = binomial(link="logit"), data =
  ↪  data)
  summary(glm_fit)$coefficients
  beta_hat = summary(glm_fit)$coefficients["(Intercept)","Estimate"]
  cov <- vcovHC(glm_fit, type="HC1")
  var_beta_hat = diag(cov)
  se_EORatio = sqrt(E^2/exp(beta_hat)^2 * var_beta_hat)
  EORatio_95CI = c(EORatio - 1.96*se_EORatio, EORatio + 1.96*se_EORatio)
  EORatio_res = paste0(round(EORatio,digits=4), " (",
  ↪  round(EORatio_95CI[1],digits=4), ", ",
  ↪  round(EORatio_95CI[2],digits=4), ")")
```

```
## Calibration intercept:
library(sandwich)
glm_fit = glm(formula =
↪  outcome~1+offset(I(log(risk_interest/(1-risk_interest)))), family =
↪  binomial(link="logit"), data = data)
cox_intercept = summary(glm_fit)$coefficient["(Intercept)","Estimate"]
cov <- vcovHC(glm_fit, type="HC1")
cox_intercept_se = sqrt(diag(cov))
cox_intercept_95CI = c(cox_intercept-1.96*cox_intercept_se,
↪  cox_intercept+1.96*cox_intercept_se)
cox_intercept_res = paste0(round(cox_intercept,digits=4), " (",
↪  round(cox_intercept_95CI[1],digits=4), ", ",
↪  round(cox_intercept_95CI[2],digits=4), ")")

## Calibration slope:
library(sandwich)
glm_fit = glm(formula = outcome~I(log(risk_interest/(1-risk_interest))),
↪  family = binomial(link="logit"), data = data)
cox_slope = summary(glm_fit)$coefficient["I(log(risk_interest/(1 -
↪  risk_interest)))","Estimate"]
cov <- vcovHC(glm_fit, type="HC1")
cox_slope_se = sqrt(diag(cov))["I(log(risk_interest/(1 -
↪  risk_interest)))"]
cox_slope_95CI = c(cox_slope-1.96*cox_slope_se,
↪  cox_slope+1.96*cox_slope_se)
```

```
  cox_slope_res = paste0(round(cox_slope,digits=4), " (",
  ↪  round(cox_slope_95CI[1],digits=4), ", ",
  ↪  round(cox_slope_95CI[2],digits=4), ")")

  return(list(EORatio=c(EORatio, se_EORatio), cox_intercept=c(cox_intercept,
  ↪  cox_intercept_se), cox_slope=c(cox_slope, cox_slope_se)))
}

evaluation_cat <- function(imp.data, seed=77) {
  set.seed(seed)
  training <- list()
  test <- list()
  training.samples <- imp.data[[1]][[7]] %>% createDataPartition(p = 0.8,
  ↪  list = FALSE)
  for (df in imp.data) {
    train.data  <- df[training.samples, ]
    test.data <- df[-training.samples, ]
    training <- append(training, list(train.data), after=1)
    test <- append(test, list(test.data), after=1)
  }

  fits <- lapply(training, glm, formula="outcome~x1+x2+x3+x4+x5+x6",
  ↪  family="binomial")

  ## predictions in different imputation datasets
  predicts <- mapply(predict_fun_cat, test, fits)
```

```
## AUC
predictions <- apply(predicts, MARGIN=2, FUN=function(pred)
↪  prediction(pred, test[[1]][, 7], label.ordering = NULL))
for (i in 1:5) {
  test[[i]]$pred <- predicts[, i]
  test[[i]]$outcome <- as.numeric(as.character(test[[i]]$outcome))
}
auc_ROCR <- lapply(predictions, FUN=function(x) performance(x, measure =
↪  "auc"))
auc_ROCR <- lapply(auc_ROCR, FUN=function(auc) round(auc@y.values[[1]],
↪  6))
se_auc <- apply(predicts, MARGIN=2, FUN=function(pred)
↪  sqrt(var(roc(test[[1]][, 7], pred))))

# Calibration metrics
cali <- lapply(test, FUN=function(x) calibrationMetrics(as.data.frame(x),
↪  "pred"))
EO <- lapply(cali, FUN=function(x) x$EORatio[1])
se_EO <- lapply(cali, FUN=function(x) x$EORatio[2])
cali_int <- lapply(cali, FUN=function(x) x$cox_intercept[1])
se_cali_int <- lapply(cali, FUN=function(x) x$cox_intercept[2])
cali_slope <- lapply(cali, FUN=function(x) x$cox_slope[1])
se_cali_slope <- lapply(cali, FUN=function(x) x$cox_slope[2])

## pooled estimates
pooled <- summary(pool(fits), conf.int=TRUE)
```

```
  results <- list(pooled, auc_ROCR, se_auc, EO, se_EO, cali_int,
  ↪  se_cali_int, cali_slope, se_cali_slope)
  return(results)
}

# Calculate sensitivity and PPV at the 95th percentile of risk score
↪  distribution
predicts <- mapply(predict_fun_cat, test, fits)
df <- data.frame(predict = predicts, actual = test)
95_percentile <- tail(head(arrange(data,desc(predict)), n = nrow(data) *
↪  0.05), 1)$predict
data$95_percentile <- ifelse(df$predict >= 95_percentile, 1, 0)
sensitivity <- mean((df %>% filter(actual == 1))$95_percentile)
ppv <- mean((df %>% filter(95_percentile == 1))$actual)
```

Appendix C

ADDITIONAL TABLES

Table C.1: Low dimensional setting where the rate of outcome was 0.06: evaluation of prediction models (logistic regression) built for imputed data sets

	Not specify interaction terms	Just Another Variable (JAV)	Passive Imputation	SMCFCS
Remove outcome during MI	E/O ratio: 1.00(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.97(0.02) AUC: 0.68(0.00)	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 1.00(0.02) AUC: 0.68(0.00)	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.96(0.02) AUC: 0.67(0.00)	-
Hybrid	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.79(0.01) AUC: 0.67(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.78(0.01) AUC: 0.67(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.78(0.01) AUC: 0.67(0.00)	-
Include outcome during MI	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.98(0.01) AUC: 0.71(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.97(0.01) AUC: 0.71(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.96(0.01) AUC: 0.71(0.00)	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.97(0.01) AUC: 0.71 (0.00)

* Original complete data set: 1)E/O ratio: 1.00(0.01); 2) Calibration intercept: -0.01(0.01); 3) Calibration slope: 1.00(0.01); 4) AUC: 0.71(0.00).

Table C.2: Low dimensional setting with interactions where the rate of outcome was 0.06: evaluation of prediction models (logistic regression) built for imputed data sets

	Not specify interaction terms	Just Another Variable (JAV)	Passive Imputation	SMCFCS
Remove outcome during MI	E/O ratio: 1.00(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.97(0.02) AUC: 0.72(0.00)	E/O ratio: 1.01(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.96(0.02) AUC: 0.72(0.00)	E/O ratio: 1.01(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.95(0.02) AUC: 0.72(0.00)	-
Hybrid	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.66(0.01) AUC: 0.72(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: 0.01(0.01) Calibration slope: 0.67(0.01) AUC: 0.71(0.00)	E/O ratio: 0.99(0.01) Calibration intercept: 0.02(0.01) Calibration slope: 0.66(0.01) AUC: 0.71(0.00)	-
Include outcome during MI	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.98(0.01) AUC: 0.79(0.00)	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.97(0.01) AUC: 0.79(0.00)	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.96(0.01) AUC: 0.79(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: 0.01(0.01) Calibration slope: 0.99(0.01) AUC: 0.80 (0.00)

* Original complete data set: 1)E/O ratio: 1.00(0.01); 2) Calibration intercept: 0.00(0.01); 3) Calibration slope: 1.00(0.01); 4) AUC: 0.81(0.00).

Table C.3: Low dimensional setting where the rate of outcome was 0.10: evaluation of prediction models (logistic regression) built for imputed data sets

	Not specify interaction terms	Just Another Variable (JAV)	Passive Imputation	SMCFCS
Remove outcome during MI	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.98(0.01) AUC: 0.67(0.00)	E/O ratio: 1.02(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.97(0.01) AUC: 0.67(0.00)	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.98(0.01) AUC: 0.67(0.00)	-
Hybrid	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.78(0.01) AUC: 0.67(0.00)	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.79(0.01) AUC: 0.67(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.80(0.01) AUC: 0.67(0.00)	-
Include outcome during MI	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.99(0.01) AUC: 0.71(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.98(0.01) AUC: 0.71(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.99(0.01) AUC: 0.71(0.00)	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.98(0.01) AUC: 0.71 (0.00)

* Original complete data set: 1)E/O ratio: 1.00(0.01); 2) Calibration intercept: 0.00(0.01); 3) Calibration slope: 1.00(0.01); 4) AUC: 0.71(0.00).

Table C.4: Low dimensional setting with interactions where the rate of outcome was 0.10: evaluation of prediction models (logistic regression) built for imputed data sets

	Not specify interaction terms	Just Another Variable (JAV)	Passive Imputation	SMCFCS
Remove outcome during MI	E/O ratio: 1.00(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.98(0.01) AUC: 0.71(0.00)	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.97(0.01) AUC: 0.71(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: 0.01(0.01) Calibration slope: 0.99(0.01) AUC: 0.71(0.00)	-
Hybrid	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.67(0.01) AUC: 0.71(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.69(0.01) AUC: 0.71(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.67(0.01) AUC: 0.71(0.00)	-
Include outcome during MI	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.99(0.01) AUC: 0.78(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.99(0.01) AUC: 0.78(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.98(0.01) AUC: 0.78(0.00)	E/O ratio: 1.01(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 0.99(0.01) AUC: 0.79 (0.00)

* Original complete data set: 1)E/O ratio: 1.00(0.01); 2) Calibration intercept: 0.00(0.01); 3) Calibration slope: 1.01(0.01); 4) AUC: 0.79(0.00).

Table C.5: High dimensional setting where the rate of outcome was 0.10: evaluation of prediction models (ridge regression) built for imputed data sets

	Not specify interaction terms	Just Another Variable (JAV)	Passive Imputation	SMCFCS
Remove outcome during MI	E/O ratio: 1.00(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 1.02(0.02) AUC: 0.68(0.00)	E/O ratio: 1.03(0.01) Calibration intercept: -0.02(0.01) Calibration slope: 0.98(0.02) AUC: 0.67(0.00)	E/O ratio: 1.02(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 1.03(0.02) AUC: 0.67(0.00)	-
Hybrid	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.99(0.02) AUC: 0.68(0.00)	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 0.99(0.02) AUC: 0.67(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 1.01(0.02) AUC: 0.67(0.00)	-
Include outcome during MI	E/O ratio: 1.01(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 1.00(0.02) AUC: 0.69(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 1.01(0.02) AUC: 0.68(0.00)	E/O ratio: 1.00(0.01) Calibration intercept: -0.01(0.01) Calibration slope: 1.01(0.02) AUC: 0.68(0.00)	E/O ratio: 1.01(0.01) Calibration intercept: 0.00(0.01) Calibration slope: 1.01(0.02) AUC: 0.68 (0.00)

* Original complete data set: 1)E/O ratio: 1.00(0.01); 2) Calibration intercept: -0.01(0.01); 3) Calibration slope: 1.00(0.02); 4) AUC: 0.68(0.00).

Table C.6: High dimensional setting where the rate of outcome was 0.10: evaluation of prediction models (random forest) built for imputed data sets

	Not specify interaction terms	Just Another Variable (JAV)	Passive Imputation	SMCFCS
Remove outcome during MI	E/O ratio: 1.13(0.01) Calibration intercept: -0.14(0.01) Calibration slope: 0.86(0.02) AUC: 0.64(0.00)	E/O ratio: 1.15(0.01) Calibration intercept: -0.15(0.01) Calibration slope: 0.85(0.02) AUC: 0.64(0.00)	E/O ratio: 1.14(0.01) Calibration intercept: -0.14(0.01) Calibration slope: 0.85(0.02) AUC: 0.64(0.00)	-
Hybrid	E/O ratio: 1.13(0.01) Calibration intercept: -0.14(0.01) Calibration slope: 0.87(0.02) AUC: 0.64(0.00)	E/O ratio: 1.13(0.01) Calibration intercept: -0.15(0.01) Calibration slope: 0.81(0.02) AUC: 0.64(0.00)	E/O ratio: 1.13(0.01) Calibration intercept: -0.14(0.01) Calibration slope: 0.82(0.02) AUC: 0.64(0.00)	-
Include outcome during MI	E/O ratio: 1.13(0.01) Calibration intercept: -0.14(0.01) Calibration slope: 0.90(0.02) AUC: 0.65(0.00)	E/O ratio: 1.13(0.01) Calibration intercept: -0.14(0.01) Calibration slope: 0.84(0.02) AUC: 0.64(0.00)	E/O ratio: 1.13(0.01) Calibration intercept: -0.14(0.01) Calibration slope: 0.85(0.02) AUC: 0.64(0.00)	E/O ratio: 1.13(0.01) Calibration intercept: -0.13(0.01) Calibration slope: 0.85(0.02) AUC: 0.64 (0.00)

* Original complete data set: 1)E/O ratio: 1.12(0.01); 2) Calibration intercept: -0.14(0.01); 3) Calibration slope: 0.87(0.02); 4) AUC: 0.65(0.00).