Evaluating Existing Risk Prediction Models in the Presence of Missing Covariates

Jeremy M G Taylor, Pin Li, Matthew Schipper
Department of Biostatistics, University of Michigan

February 28, 2020

Prediction Models

- Binary outcome variable Y.
- Baseline covariates X.
- A prediction model is an equation or algorithm that gives $P(Y = 1|X)$
- Three different settings
  - Building a new model
  - Applying an existing model
  - Evaluating an existing model

Different settings for using prediction models

- Three different settings
  - Build a new model, have data (Y_i, X_i), i=1,...,N
  - Apply an existing model to an individual, have data (X), for one subject
  - **Evaluating an existing model, have new data (Y_i, X_i), i=1,...,N**

Main Goal

- The assessment of an existing prediction models when the data have missing covariate values.
- Metrics for assessing predictions models
  - Discrimination - AUC = area under the ROC curve.
  - Calibration - Brier score (BS) = MSE of predicted probabilities.
- Question, how to estimate AUC and BS when there are missing X's? .

Methods for handling missing data

- Two general methods for analyzing datasets with missing values.
  - Multiple imputation (MI).
  - Inverse Probability Weighting (IPW).
- Both MI and IPW involve models
  - A model for the value of the missing variable
  - A model for the probability of missingness
- Question, Should these models involve Y?
- How do MI and IPW based methods compare with simply throwing away any observations with missing X's, i.e. Complete Case analysis?

Should Y be included in the imputation model?

- In general for multiple imputation it is well known that the outcome should be included as a covariates when imputing missing X's.
- When building a new prediction model Moons et al (2008) showed that the outcome variable should be included in the multiple imputation approach.
- People are suspicious of this and don't want to believe it.

Comparing MI and IPW

- How do MI, IPW and Augmented IPW (AIPW) compare in terms of:
  - Bias
  - Efficiency
  - Robustness to model mispecification
  - Reason for missingness:
    - missing complete at random (MCAR),
    - missing at random (MAR),
    - Missing Not at Random (MNAR).
- The ideal estimators is - what the estimate would be if there were no missing data

AUC and BS will likely differ between studies

- Loosely speaking, a prediction model is said to be "validated" if a new dataset gives a similar AUC and BS as the original study.
- AUC and BS can differ between studies because
  - Because the distributions of $[Y|X]$ differ between the populations
  - or because the distribution of $[X]$ differs between the populations

Prostate cancer example

- The Cancer of the Prostate Risk Assessment (CAPRA) score published in 2005 as external model.

Details of the CAPRA score

Table: CAPRA score

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

Prostate cancer example

- Patients from Mayo Clinic 2008-2012 (n=1268), 90% missing in Percent positive biopsy.
- Use 3-year PFS as binary outcome. Compare outcome vs CAPRA score to get AUC. Compare outcome vs CAPRA rate for each score to get Brier score.
- Use PSA, Gleason Score, T-stage, Age (and outcome) to build the weight model and/or imputation model in IPW, AIPW and MI.

CAPRA score distribution and predicted probabilities

Table: Patient distribution and Kaplan-Meier analysis for the CAPRA database (n=1439)

CAPRA Score	% Patients	3-Yr % RFS(95%CI)
0-1	27.9	91(85-95)
2	30.0	89(83-94)
3	20.6	81(73-87)
4	10.8	81(69-89)
5	5.8	69(51-82)
6	3.0	54(27-75)
7 or Greater	2.0	24(9-43)

Estimates of AUC and BS

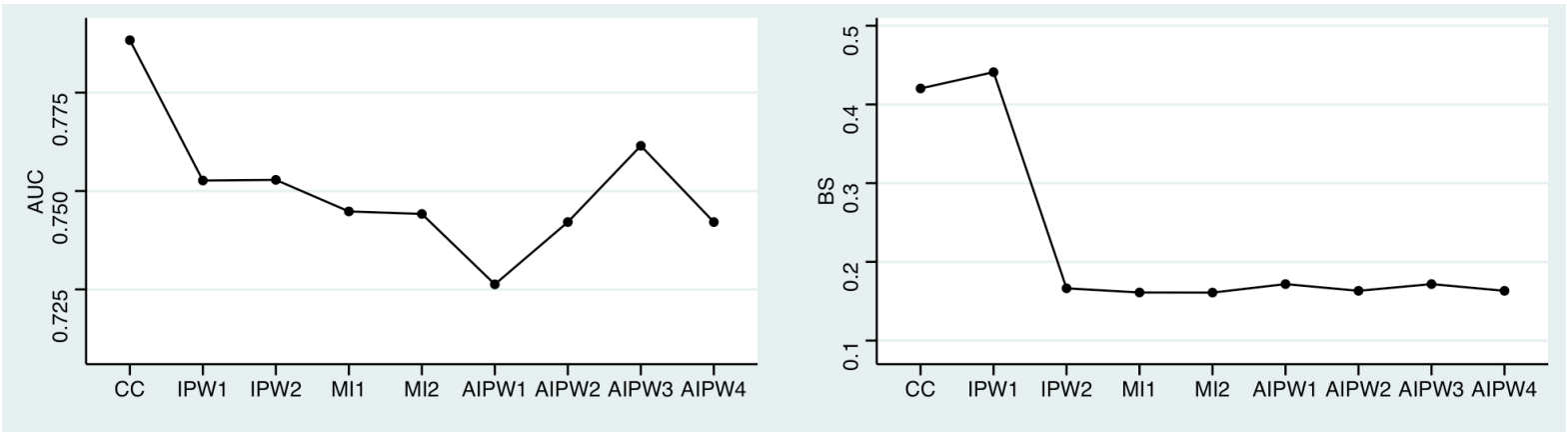


Figure: Varying estimates of AUC and Brier Scores for Prostate Cancer example, based on how missing data are handled

Different estimates of AUC and BS, depending on which method is used. Which is best?

Definition of Brier Score

A model was built from external data.

- $F_E(X)$ and $F_E(Y|X)$ denote the true distributions for the external data
- The existing model $\hat{p}(Y = 1|X)$ is an approximation to $F_E(Y|X)$.

- $F_I(X)$ and $F_I(Y|X)$ denote the true distributions for the internal data
- dataset of size N , binary outcome Y and p-dimensional vector of covariates X

The BS is given by

$$BS = \sum_{i=1}^{N} (Y_i - \hat{p}_i)^2 / N \tag{2.1}$$

Given the distribution of $F_I(X)$ and $F_I(Y|X)$,

$$TrueBrier_I(\hat{p}) = \sum_{Y} \int_{X} (Y - \hat{p})^2 F_I(Y|X) F_I(X) dX \tag{2.2}$$

Definition of AUC and C-index

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

$$AUC/C-index = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} I(\beta X_i > \beta X_j) I(Y_i > Y_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} I(Y_i > Y_j)} \tag{2.3}$$

Let X_1 denote the covariates in cases and X_0 denote the covariates in controls. Their distributions are $F_I(X_1) = F_I(X|Y=1)$ and $F_I(X_0) = F_I(X|Y=0)$, respectively.

$$TrueAUC_I(\hat{p}) = \int_{X_1}\int_{X_0} I(\beta X_1 > \beta X_0) F_I(X_1) F_I(X_0) dX_1 dX_0 \tag{2.4}$$

Method

Some X values are missing in the internal data.

ID	X_1	X_2	X_3	Y	R
1	4	2	-2	1	1
2	NA	-2	1	0	0
3	NA	-3	2	1	0
4	2	-3	4	1	1
...					

How do we get **good** estimates of $TrueBrier_I(\hat{p})$ and $TrueAUC_I(\hat{p})$?
- Small bias, low variability, robust to model misspecification.

Complete case analysis

Using only complete case (i.e $R_i = 1$) the simplest estimates are

$$BS_{CC} = \frac{\sum_{i=1}^{N}(Y_i - \hat{p}_i)^2 R_i}{\sum_{i=1}^{N} R_i} \tag{2.5}$$

$$C - index_{CC} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} I(\beta X_i > \beta X_j) I(Y_i > Y_j) R_i R_j}{\sum_{i=1}^{N}\sum_{j=1}^{N} I(Y_i > Y_j) R_i R_j} \tag{2.6}$$

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

Multiple Imputation

- Build imputation models: $F(X_{mis}|X_{obs}, Y)$ or for $F(X_{mis}|X_{obs})$
- Draw a value of X_{mis} from the model, then apply the external model on the imputed complete data to get the predictions of Y and calculate BS and AUC
- Repeat 2nd step for M times (M=5), average the predicted BS and AUC from the multiple imputed datasets using Rubin's rule
- When there is more than one covariate with missing values, a chained equation approach is used to impute the missing values sequentially

Inverse Probability Weighting

The weight (W_i) is the inverse probabilty of observation i being complete ($R_i = 1$), i.e. $W_i = 1/\Pr(R_i = 1)$.
Build the weight model of either $\Pr(R_i = 1|X_i, Y_i)$ or $\Pr(R_i = 1|X_i)$ by logistic regression.

$$BS_{IPW} = \frac{\sum_{i=1}^{N}(Y_i - \hat{p}_i)^2 R_i W_i}{\sum_{i=1}^{N} R_i W_i} \tag{2.7}$$

$$C - index_{IPW} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} I(\beta X_i > \beta X_j) I(Y_i > Y_j) R_i W_i R_j W_j}{\sum_{i=1}^{N}\sum_{j=1}^{N} I(Y_i > Y_j) R_i W_i R_j W_j} \tag{2.8}$$

Augmented Inverse Probability Weighting

Augmented Inverse Probability Weighting (AIPW) also include information from subjects with missing data.

- Predicted mean X^*, i.e., $E(X_{mis}|X_{obs}, Y)$ or $E(X_{mis}|X_{obs})$, from imputation model, created for all subjects
- Weight W_i from weight model

$$BS_{AIPW} = \frac{\sum_{i=1}^{N}(Y_i - \hat{p}_i)^2 R_i W_i + (Y_i - \hat{p}_i{}^*)^2(1 - R_i W_i)}{N} \tag{2.9}$$

- Missing subject $R_i = 0$: $(Y_i - \hat{p}_i{}^*)^2$.
- Complete subject $R_i = 1$: $(Y_i - \hat{p}_i)^2 W_i + (Y_i - \hat{p}_i{}^*)^2(1 - W_i)$.

Augmented Inverse Probability Weighting

For C-index,

$$C-index_{AIPW} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} I(Y_i > Y_j)\{I(\beta X_i > \beta X_j)R_i W_i R_j W_j + I(\beta X_i^* > \beta X_j^*)(1 - R_i W_i R_j W_j)\}}{\sum_{i=1}^{N}\sum_{j=1}^{N} I(Y_i > Y_j)} \tag{2.10}$$

Consistency of AIPW estimators

Double robustness property of AIPW method.

If either the weight model or the model for missing X is correctly specified then the AIPW estimators are consistent.

Simulation

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

X_1, X_2, X_3 are sampled from $N(0, 1)$ and about 40% of X_1 is missing. The covariates can be independent, or correlated with $cor(X_1, X_3) = -0.5$.

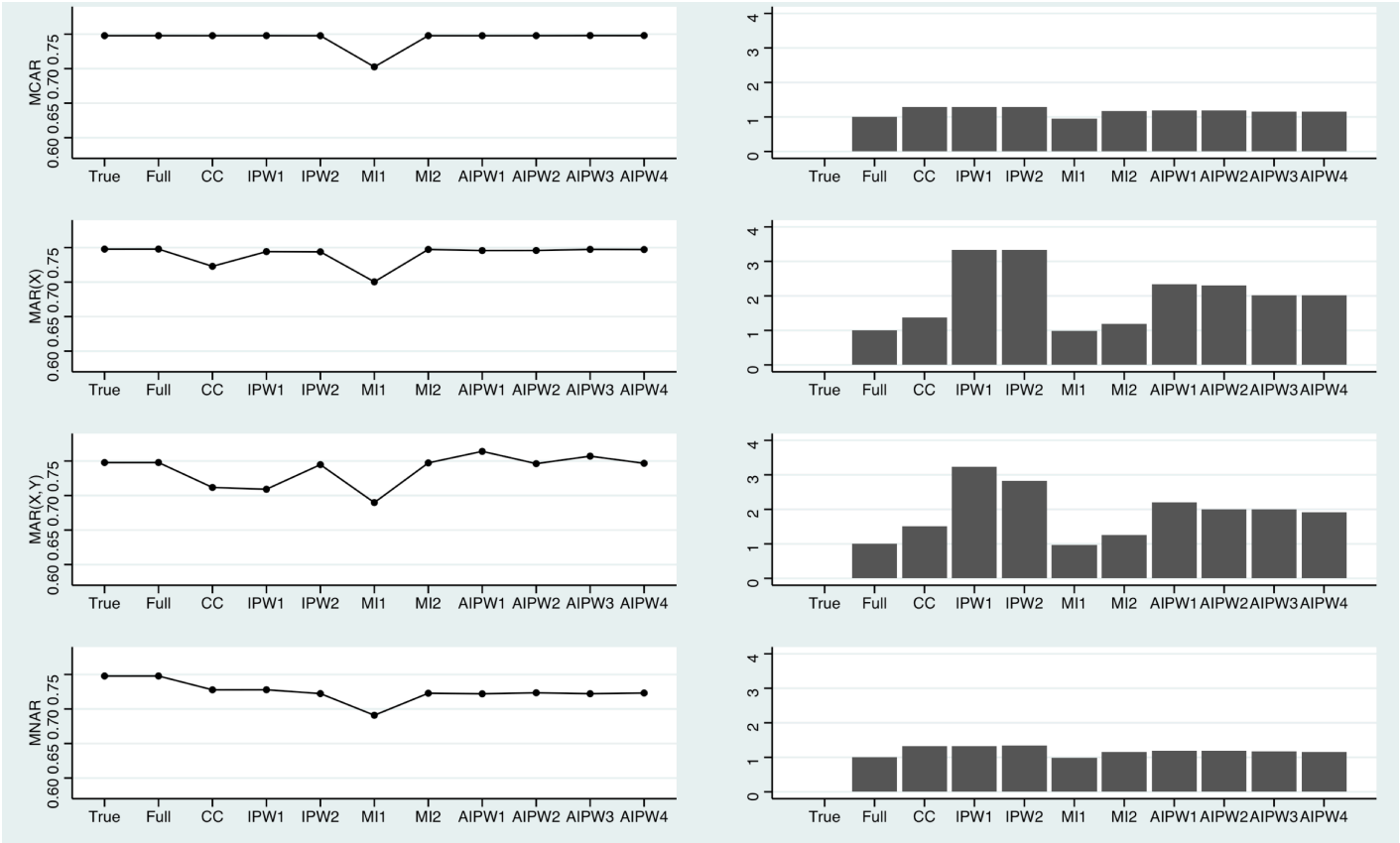
Simulation results of mean and relative SD of AUC and BS from 1000 iterations.
MCAR: Pr(X_1 is missing)=0.4.
MAR(X_2, X_3): Pr(X_1 is missing)= $invlogit(-0.5 + 2X_2 - 2X_3)$.
MAR(X_2, Y): Pr(X_1 is missing)= $invlogit(-0.5 + 2X_2 + Y)$.
MNAR(X_1): Pr(X_1 is missing)= $invlogit(-0.5 + 3X_1)$.

Simulation

We compared the validation of external models on full internal data:

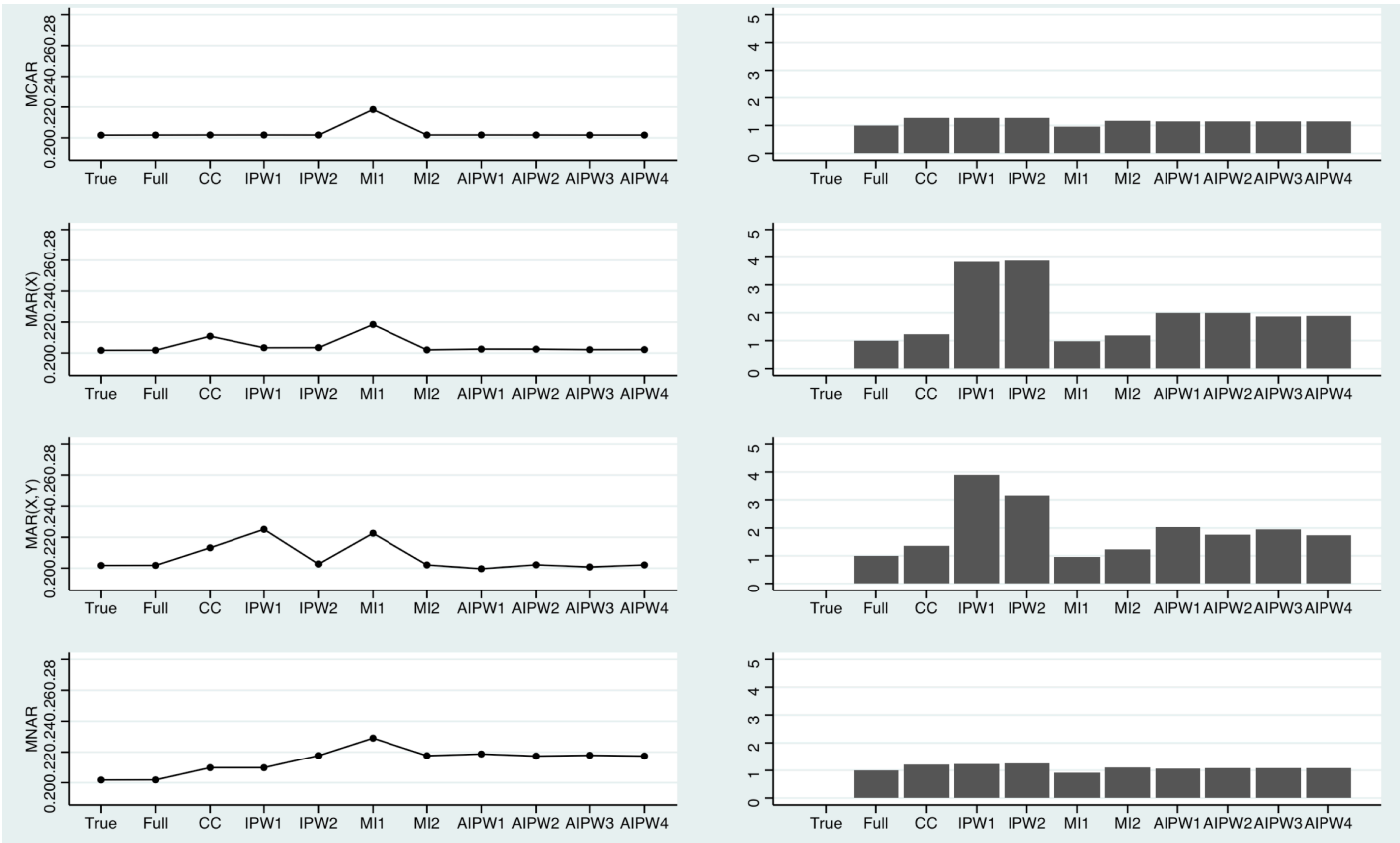
- True: true value based on internal data distribution
- Full: data without missing (target value)
- CC: complete cases anaylsis
- IPW1: weight model uses X
- IPW2: weight model uses X & Y
- MI1: imputation model uses X
- MI2: imputation model uses X & Y
- AIPW1: weight model uses X, imputation model uses X
- AIPW2: weight model uses X & Y, imputation model uses X
- AIPW3: weight model uses X, imputation model uses X & Y
- AIPW4: weight model uses X & Y, imputation model uses X & Y

Simulation results - AUC



*The model used to impute the missing X in MI2 and create X^* in AIPW3 and AIPW4 is slightly misspecified.

Simulation results - Brier Score



Summary of Simulation Results

- When there are missing observations in the internal data, MI and IPW can be used to get unbiased BS and AUC if the imputation model or weight model is correctly specified.
- AIPW can improve the efficiency of IPW, and also get the double robustness from mis-specification of weight model or imputing model.
- MI can be more efficient than IPW and AIPW
- The outcome variable should be included in the multiple imputation under all scenarios.
- If IPW or AIPW methods are used and missingness does not depend on Y, then it does not appear to be necessary to include Y in either the weight model or the missing variable model.

Data analysis

- The Cancer of the Prostate Risk Assessment (CAPRA) score published in 2005 as external model.
- Patients from Mayo Clinic 2008-2012 (n=1268), 90% missing in Percent positive biopsy.
- Use 3-year PFS as binary outcome. Compare outcome vs CAPRA score to get AUC. Compare outcome vs CAPRA rate for each score to get Brier score.
- Use PSA, Gleason Score, T-stage, Age (and outcome) to build the weight model and/or imputation model in IPW, AIPW, and MI.

Data analysis

Table: CAPRA score

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

Data analysis

Table: Patient distribution and Kaplan-Meier analysis for the CAPRA database (n=1439)

CAPRA Score	% Patients	3-Yr % RFS(95%CI)
0-1	27.9	91(85-95)
2	30.0	89(83-94)
3	20.6	81(73-87)
4	10.8	81(69-89)
5	5.8	69(51-82)
6	3.0	54(27-75)
7 or Greater	2.0	24(9-43)

Data analysis

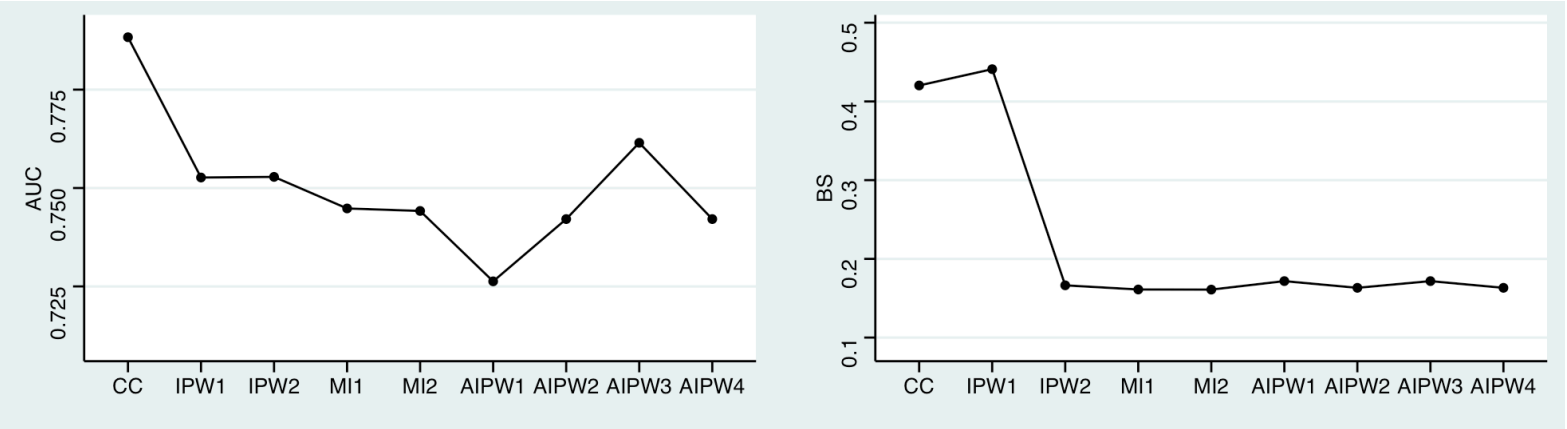


Figure: Varying estimates of AUC and Brier Scores for Prostate Cancer example, based on how missing data are handled

MI2 and AIPW4 are the best to use.

Summary

- The estimands for AUC and Brier Score do depend on the joint distribution of X and Y
- The approach to handle missing data can result in quite different estimates.
- For MI when there are missing observations you should include Y in the model for imputing missing X's.
- For IPW methods when there are missing observations you should include Y in the required models, unless missingness does not depend on Y.
- MI can be more efficient than IPW and AIPW

For further reading

Long Q, Zhang X, Johnson BA. (2011)
Robust estimation of area under ROC curve using auxiliary variables in the presence of missing biomarker values.
Biometrics, 67(2): 559–567.

Moons KG, Donders RA, Stijnen T, Harrell Jr FE. (2006)
Using the outcome for imputation of missing predictor values was preferred.
Journal of clinical epidemiology, 59(10): 1092–1101.

White IR, Royston P, Wood AM. (2011)
Multiple imputation using chained equations: issues and guidance for practice.
Statistics in medicine, 30(4): 377–399.

Bang H, Robins JM. (2005)
Doubly robust estimation in missing data and causal inference models.
Biometrics, 61(4): 962–973.