Variable selection and Binary prediction with incomplete data: Balance between fairness and precision

He Ren[1] and Chun Wang[1]
[1] College of Education, University of Washington

UNIVERSITY *of* WASHINGTON | COLLEGE OF EDUCATION

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION FOUNDED 1916

Outline

> Research question background

> Methods
  - Bootstrap Imputation-Stability Selection (BI-SS)
  - Stacked Elastic Net (SENET)

> Simulations

> Discussions

Background - Motivation

> "Project HOPE: Achieving home discharge for institutionally bound patients with PROMs, AI, and EHR"
>   - **Ho**me Post acute care (PAC) potential (HoPe)

> **Aim 2**: Develop a machine learning algorithm to prioritize actionable HoPe barriers (clinical, functional, and social factors) and estimate the degree of change in these factors required for safe home discharge.

Challenges & Solutions

> 76 predictors, N=134, 807

> Missing data

> Subgroups
> - Sex and racial groups

> Penalized regression

> Multiple Imputation

> Equal AUCs across groups, Calibration across groups, and the Equal F1 across groups

Procedures

Split the data: training (75%) & Test data (25%)

Multiple imputation (MI)

- We conducted MI on training data
- Re-use the training data MI model on the test data

[Training] **Variable selection**

[Training] Fit logistic regression model on each imputed data sets

[Test] Compute AUC

Methods

> Bootstrap Imputation-Stability Selection (BI-SS) (Long & Johnson, 2015)
  - Generate bootstrap dataset and impute the bootstrap dataset
  - Variable selection via randomized Lasso on each imputed bootstrap data set
  - Combine selection results via stability selection

> Stacked Elastic Net method (SENET) (Du, Boss, Han, et al., 2022)
  - Optimize pooled objective function jointly over all imputed data sets

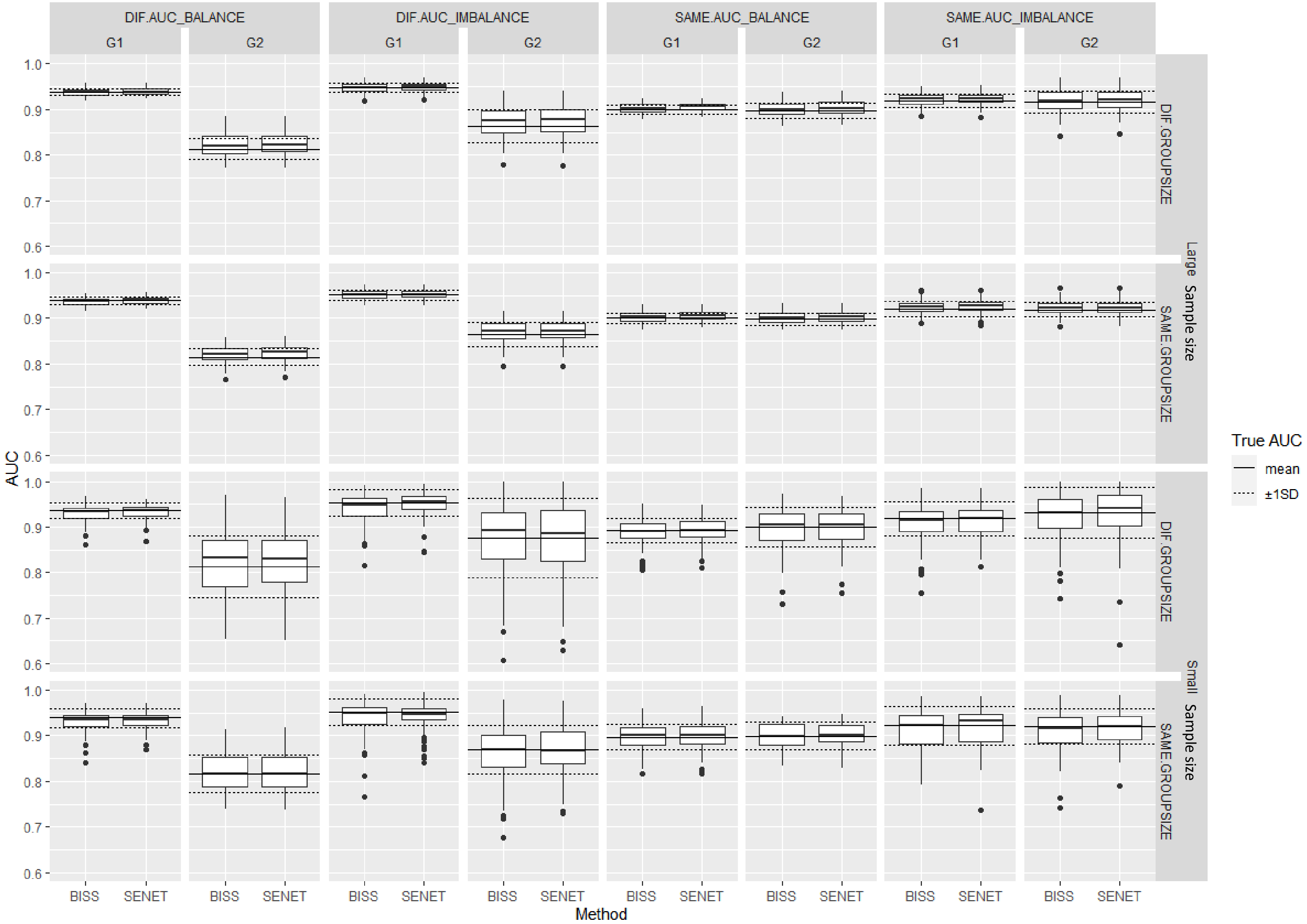
Simulation design

> 25% significant predictors
> Correlations among predictors: 0.5

> Overall sample size: N=800 vs. N=4800
> Overall Base rate: balanced (~50%) vs. unbalanced (~10%)
> Group sample size: $N_{g1}: N_{g2} = 1:1$ vs. $N_{g1}: N_{g2} = 3:1$
> Group AUC: same vs. different across groups

- $logit\big(P(Y = 1)\big) = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$
  > Varying β_0: base rate
  > Varying β_p's: true AUC

Simulation results (Estimated AUC results of study 2)



UNIVERSITY of WASHINGTON

Algorithm bias

- Algorithm bias metric I: Equal error rate
- Algorithm bias metric II: Calibration
  - The agreement between observed outcomes and predictions. A fair algorithm should have the same good calibration across subgroups. (Barocas & Selbst, 2016, Kleinberg et al., 2016; Chouldechova, 2017)
- When the "base rate" is unequal in subpopulations, attempts to satisfy criterion of calibration (i.e., prediction parity) will result in different AUCs (Chouldechova, 2017).

Summary

> BI-SS vs. SENET

> Data vs. algorithmic bias

Thanks!
heren@uw.edu

Acknowledgement: The project is supported by NIH R01AG077706-02 awarded to Mayo Clinic (PI: Dr. Andrea Cheville) and subcontracted with the University of Washington.

UNIVERSITY *of* WASHINGTON | COLLEGE OF EDUCATION

BI-SS

(Long & Johnson, 2015)

From original data set, Z_O, draw K Bootstrap data sets

Perform imputation on the K data sets separately, $Z_{BI}^{(k)}$

For each k, perform randomized Lasso

$$\hat{\boldsymbol{\beta}}_\lambda^{(k)} = \operatorname*{argmin}_{\boldsymbol{\beta}} \left(\left\| \boldsymbol{y}^{(k)} - \boldsymbol{X}^{(k)}\boldsymbol{\beta} \right\|_2^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{w_j^{(k)}} \right)$$

The set of selected variables is

$$\hat{S}_\lambda^{(k)} = \left\{ j \colon \hat{\boldsymbol{\beta}}_\lambda^{(k)} \neq 0 \right\}$$

Combine variables from K data sets

$$\hat{S}_\pi = \left\{ j \colon \max_{\lambda \in \Lambda} \left(\Pi_j^\lambda \right) \geq \pi \right\}$$

where $\Pi_j^\lambda = (1/K) \sum_k I \left(j \in \hat{S}_\lambda^{(k)} \right)$

Use cross-validation to select optimal (π, λ)

Given $\hat{S}_\pi$, fit a classical regression model on each $Z_{BI}^{(k)}$, take an average

$$\hat{\beta}_{\hat{S}_\pi} = 1/K \sum_k \hat{\beta}_{\hat{S}_\pi}^{(k)}$$

Re-generate another M imputation datasets from Z_0 and use Rubin's rule (1987) to combine parameter estimates.

SENET (Du et al., 2022)

> Joint loss function

- $\hat{\beta} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(-\frac{1}{n} \sum_{m=1}^{M} \sum_{i=1}^{n} r_i \log L \left(\boldsymbol{\beta} \middle| Z_{OI}^{(m)} \right) + \lambda P_\alpha(\boldsymbol{\beta}) \right)$

Weight: $1/\ M$,
or f_i/M

m^{th} imputed data set

$P_\alpha(\boldsymbol{\beta}) = \alpha \Sigma_j |\beta_j| + (1-\alpha)\Sigma_j \beta_j^2$

Real data analysis

- N= 134, 807
  - 11,761 SNF discharge (8.724%)
- 76 predictors of mixed metric
  - Demographics
  - Social determinants of health
    - Food insecurity, housing stability, intimate partner, social connections, nutrition, etc.
    - (absolute) Correlations among them are in the range of: (0.0002, 0.9601)
  - Missingness
    - Item level: the missing proportion is in the range of (0, 0.678)
    - Person level : the missing proportion is in the range of (0, 0.776)

Real data results

Item content	Domain	
How often do you have a drink containing alcohol?	Alcohol	
Do you have a regular dentist that you see at least once a year for a check-up?	Dental	
Choose your current employment status	Employment	BI-SS
Do you use extra virgin olive oil as your main source of fat in your diet?	Nutrition	
On average, how many minutes do you engage in exercise at this level?	Physical activates	
On average, how many days per week do you engage in moderate to strenuous exercise (like walking fast, running, jogging, dancing, swimming, biking, or other activities that cause a light or heavy sweat)?	Physical activates	
In a typical week, how many times do you talk on the telephone with family, friends, or neighbors?	Social connections	BI-SS
In the past 12 months, has lack of transportation kept you from meetings, work, or getting things needed for daily living?	Transportation needs	
Marital status	Demographic	
Severity and Age Weighted Sum of Diseases	Demographic	
Age	Demographic	
Difference between date of discharge and SDOH completion date	Demographic	

FOUNDED 1916 ...IATION

Real data analysis

> Overall out-of-sample prediction accuracy

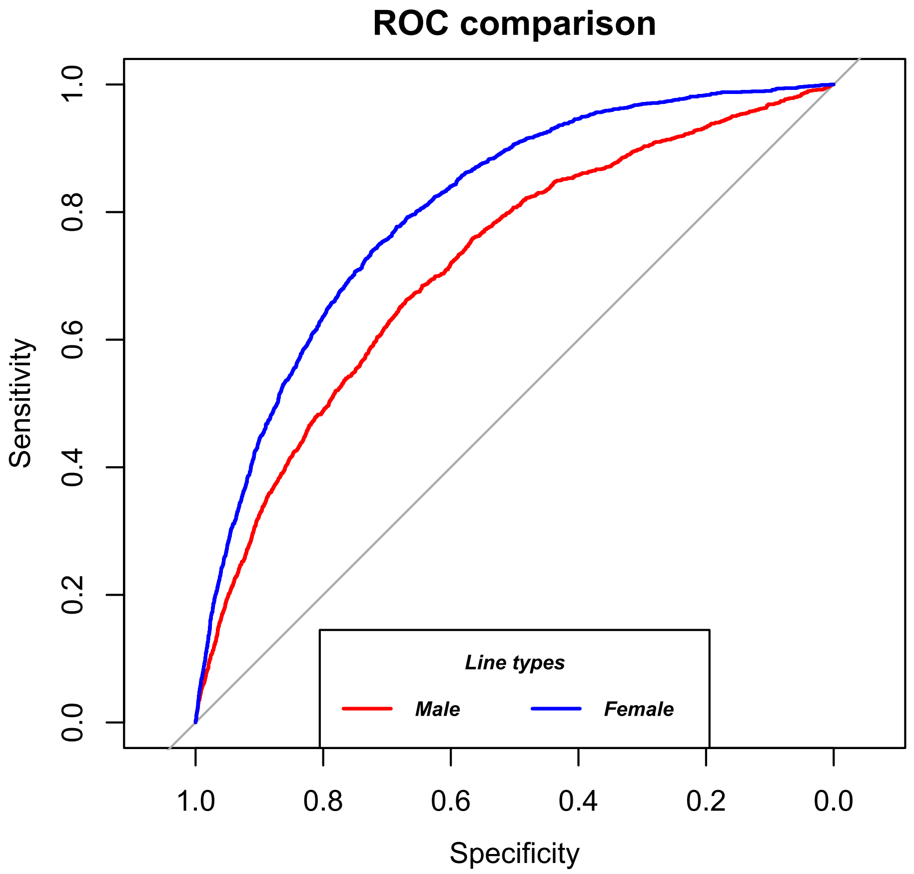
Method	BI-SS	SENET	All Variables
AUC	0.767	0.765	0.769

Real data analysis

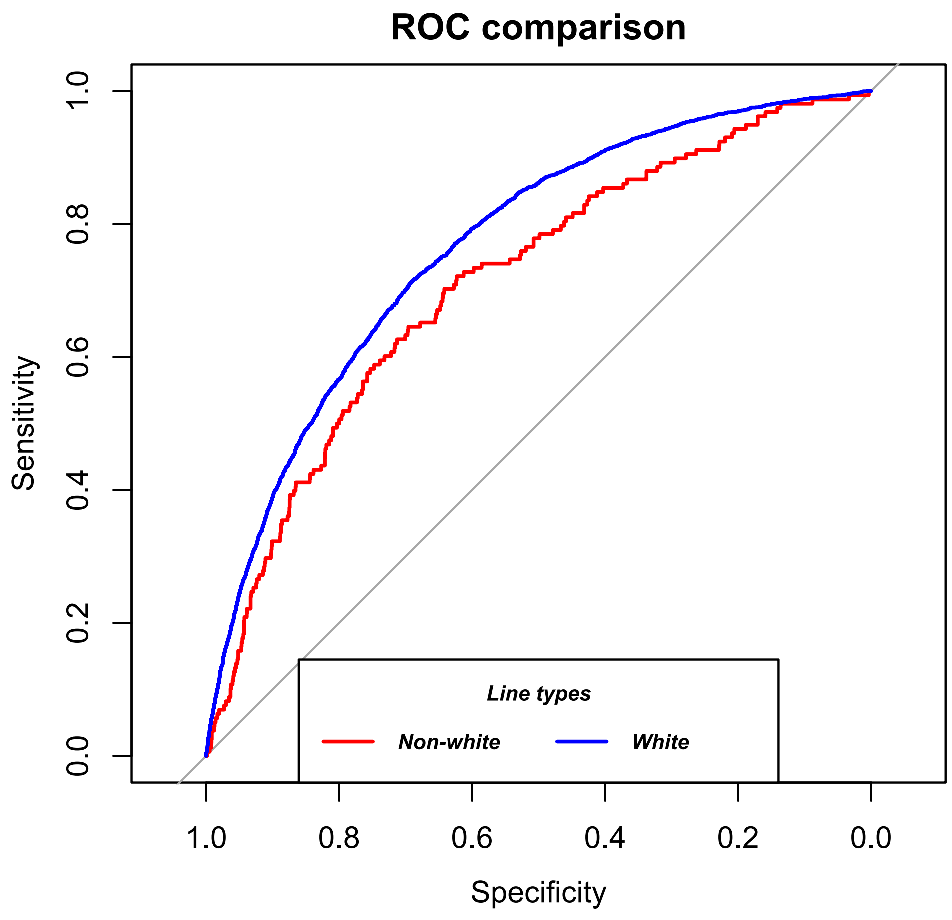
- Subgroup out-of-sample prediction accuracy
- Algorithmic bias ?

	Category	AUC	N	Home	Proportion
Sex	Male	0.716	15272	13917	91.13%
	Female	0.803	18428	16843	91.40%
Race	Black or African American	0.686	1089	1028	94.40%
	American Indian or Alaska Native	0.764	216	197	91.20%
	Asian	0.715	612	579	94.61%
	Native Hawaiian or Other Islander	0.607	67	63	94.03%
	White	0.767	30956	28174	91.01%
	Other	0.760	472	442	93.64%
	Refused	0.712	289	278	96.19%
Race	Non-White	0.712	2745	2587	94.24%
	White	0.767	30956	28174	91.01%
Ethnicity	Hispanic or Latino	0.696	1454	1395	95.94%
	Not Hispanic or Latino	0.765	31715	28877	91.05%
	Patient refused/Decline	0.786	532	489	91.92%

Algorithm bias metric I: Equal error rate



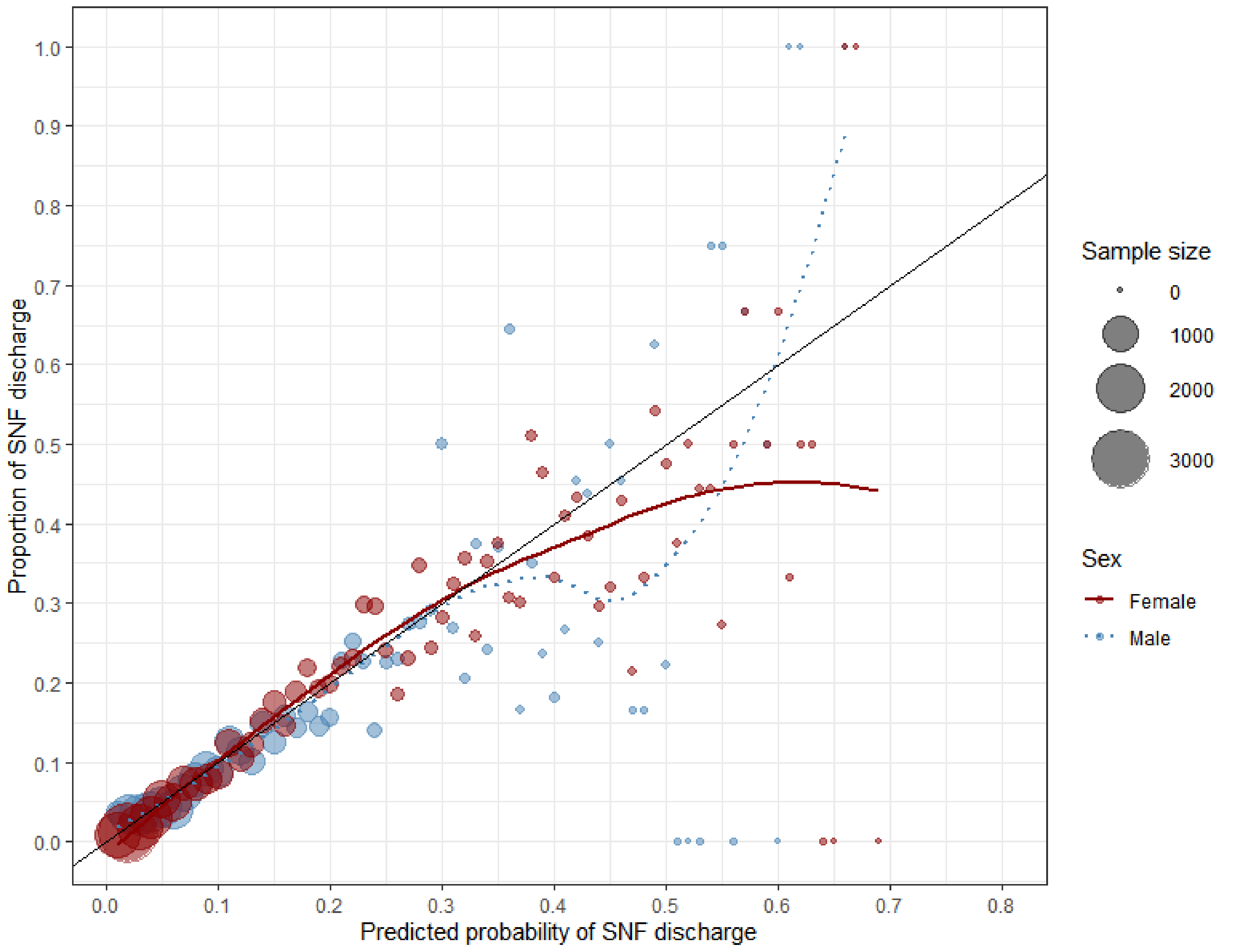
Algorithm bias metric I: Equal error rate



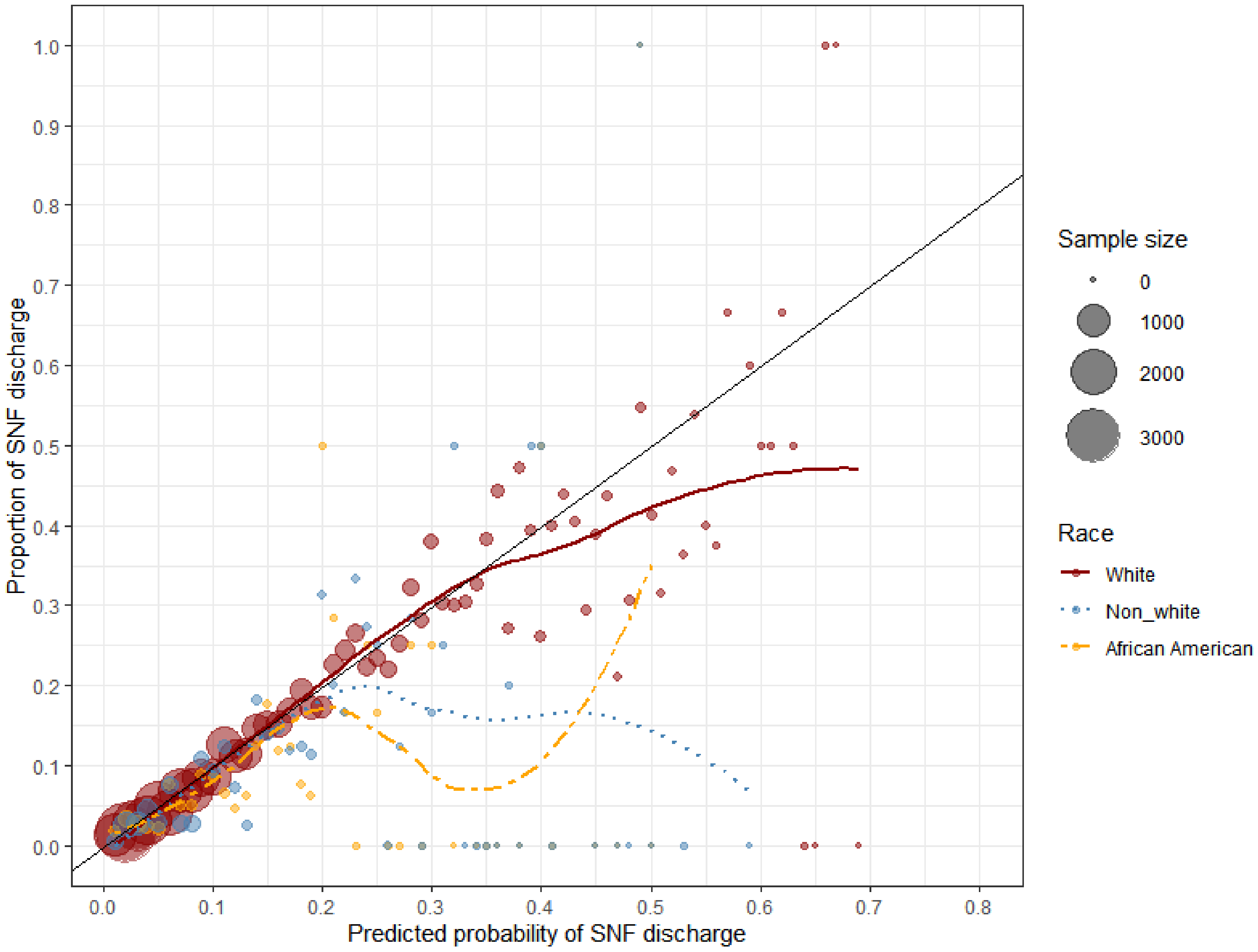
Algorithm bias metric II

> Calibration: The agreement between observed outcomes and predictions. A fair algorithm should have the same good calibration across subgroups. (Barocas & Selbst, 2016, Kleinberg et al., 2016; Chouldechova, 2017)

- The predicted probability of SNF discharge should match the overall proportion of actual SNF discharge in the same group

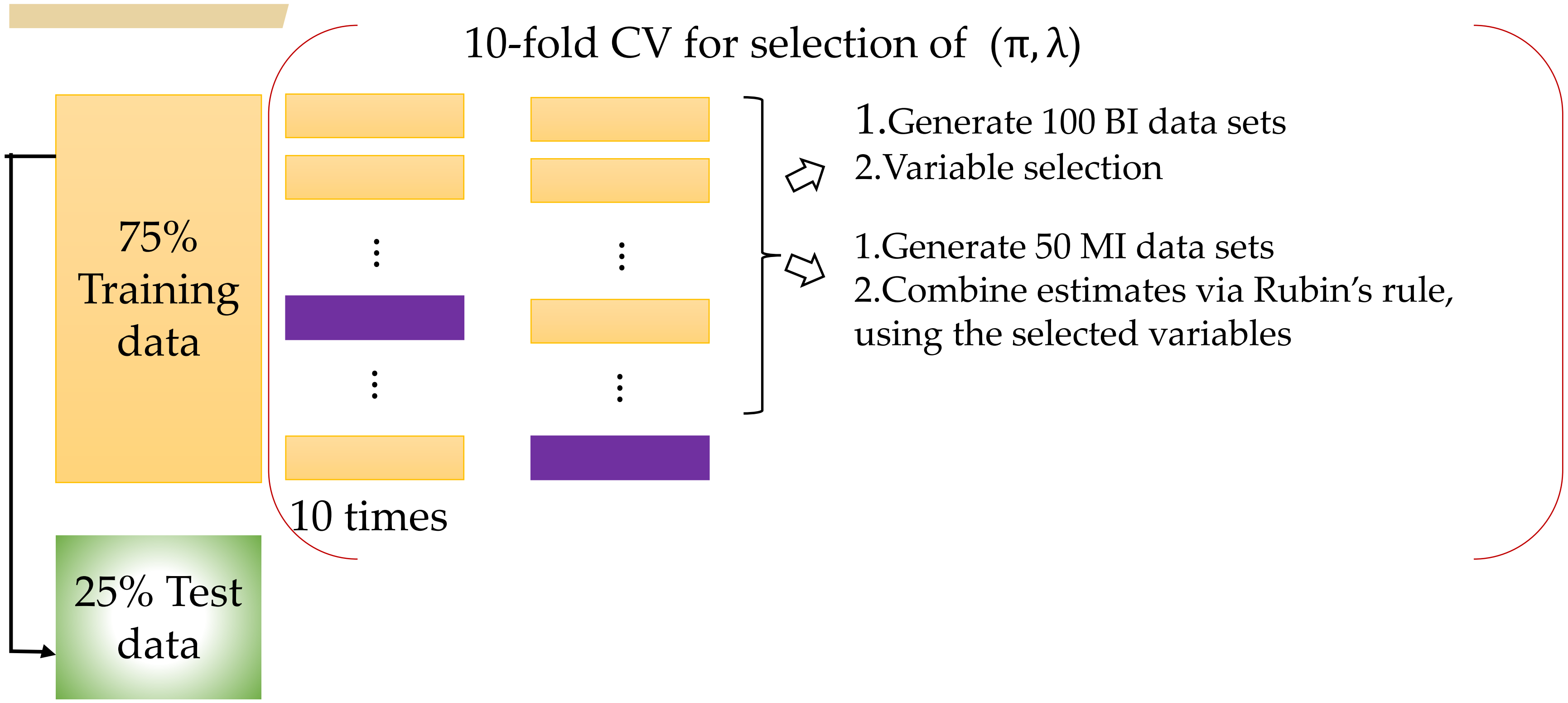


FOUNDED 1916
AMERICAN EDUCATIONAL RESEARCH ASSOCIATION

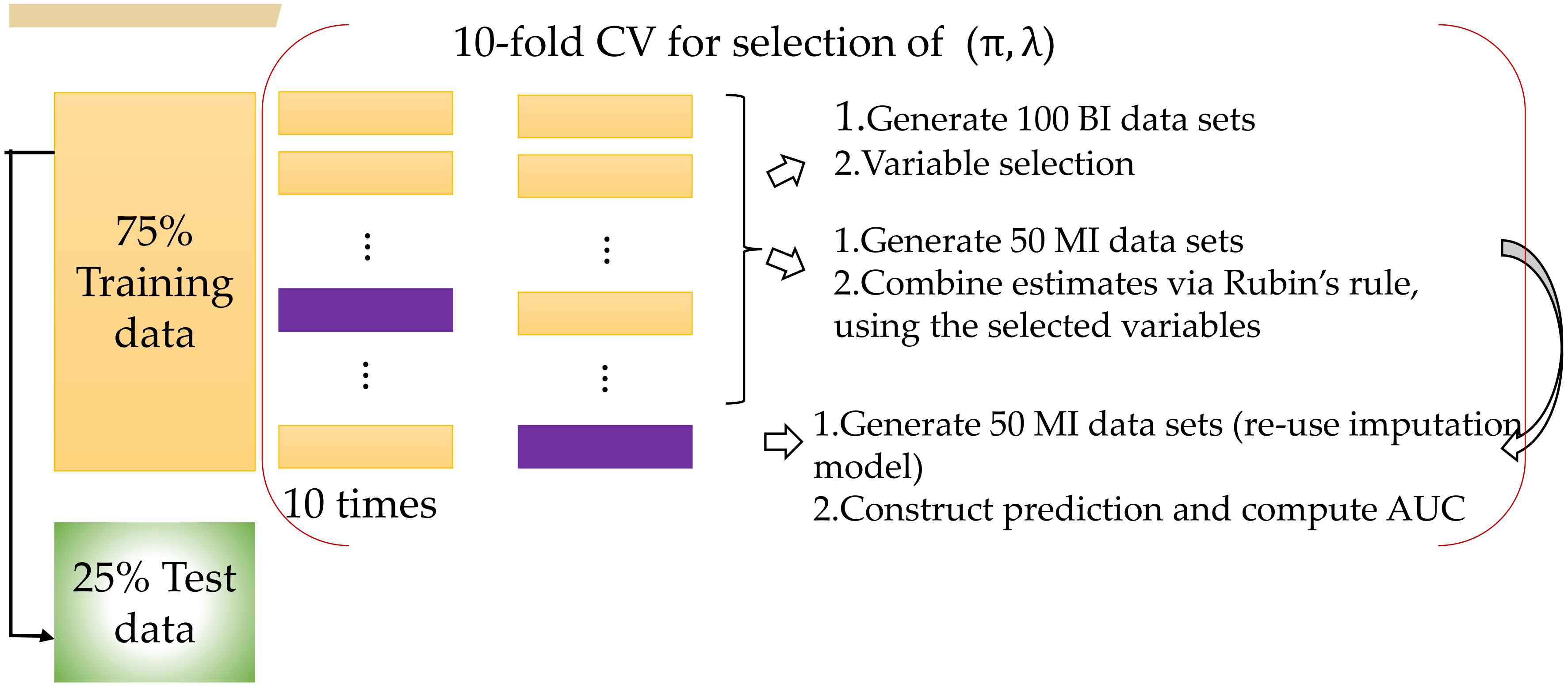


Proportion of SNF discharge
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Predicted probability of SNF discharge
Sample size
0
1000
2000
3000
Race
White
Non_white
African American

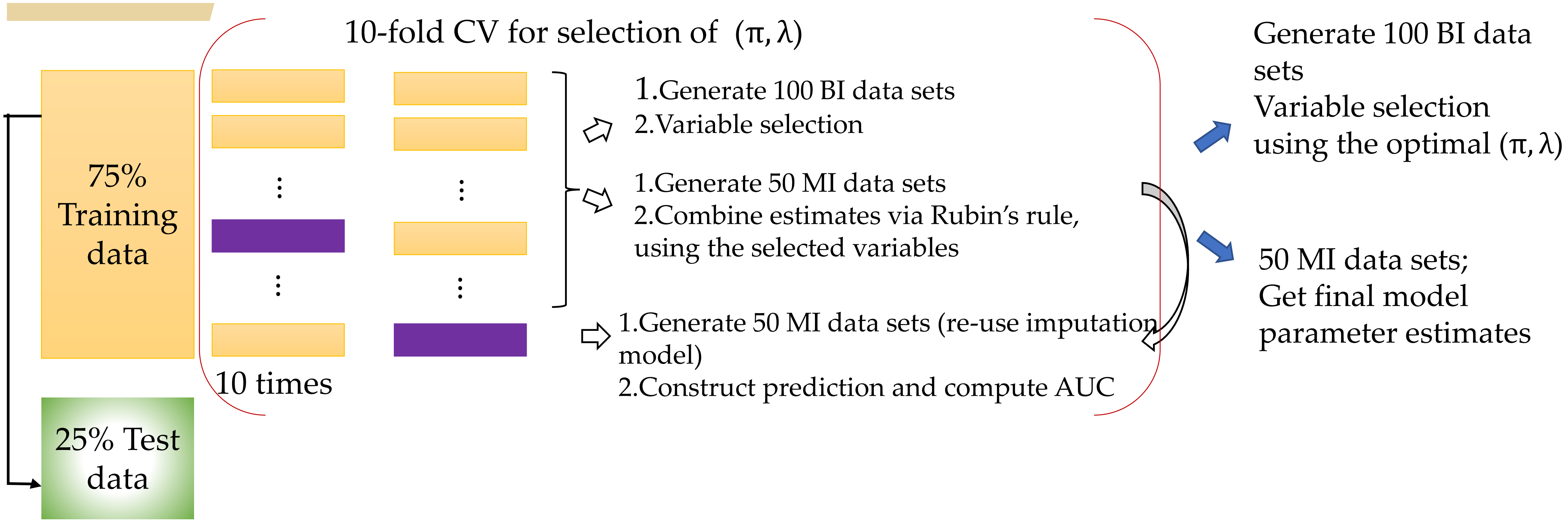
Selection of (π, λ)



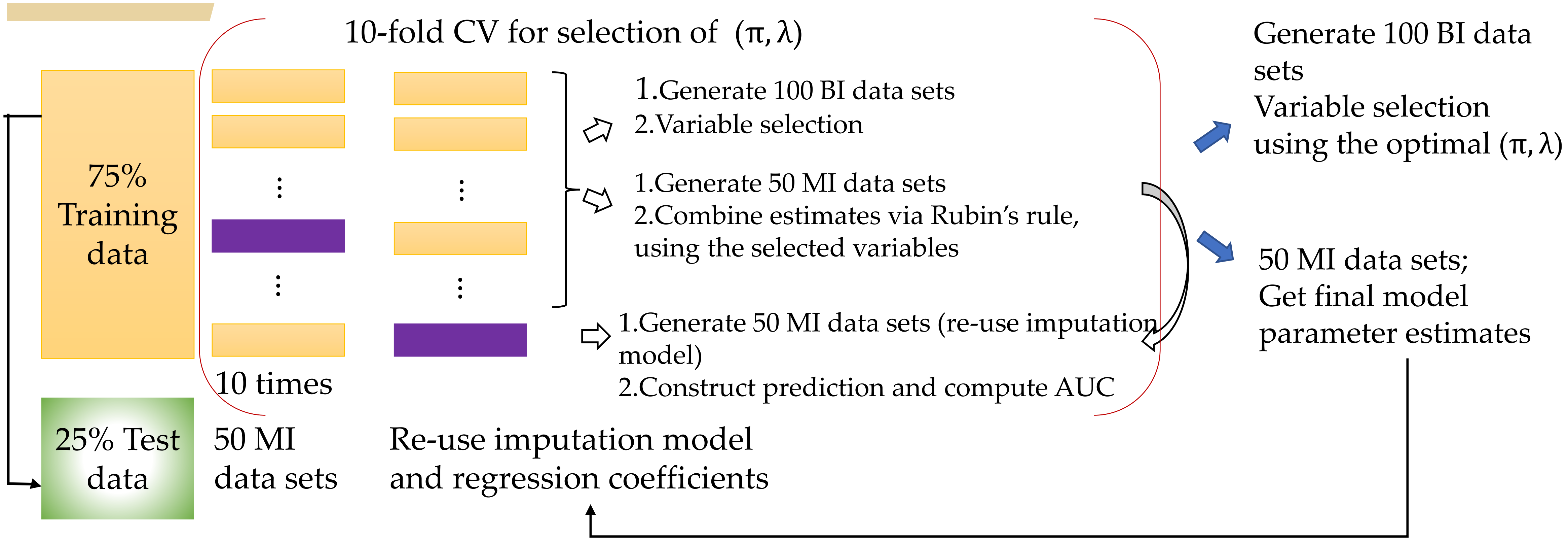
Selection of (π, λ)



Selection of (π, λ)

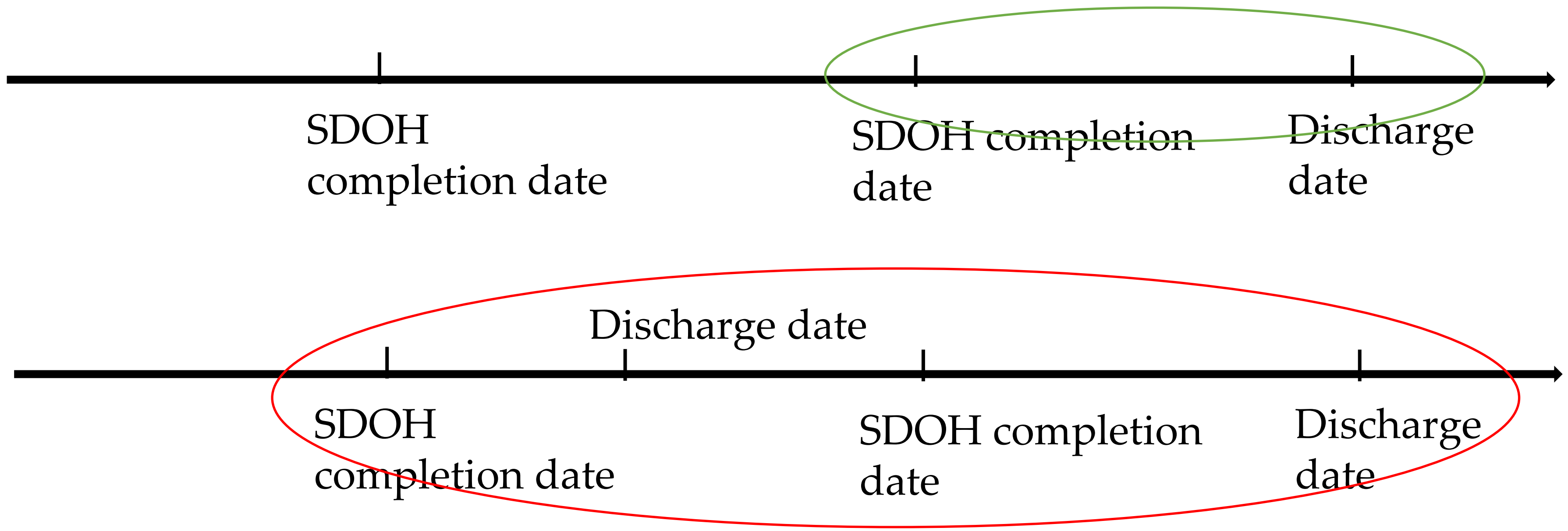


Selection of (π, λ)



Longitudinal data

- Data pattern



Male (AUC=0.716, N=15,272)

		Data	
		Home	Facility
Prediction	Home	13872	1326
	Facility	45	29

- Sensitivity=.997, Specificity=.0214
- 91.13% home discharge in the data (vs. 91.28% in the entire sample)

Female (AUC=0.803, N=18,428)

		Data	
		Home	Facility
Prediction	Home	16756	1518
	Facility	87	67

- Sensitivity=.995, Specificity=.0423
- 91.4% home discharge in the data