Variable Selection & Inference

Jeffrey D. Blume, PhD
Vanderbilt University

Topics

- Prediction, Estimation and Attribution
- Variable Selection Algorithms (overview)
- Regularization
  - Bias-variance tradeoff
  - Implication for inference tasks
- Relaxing as a general inferential strategy
- Treat 2.0 model example
  - Aligning Prediction and Inference models
  - Handling missing data

Prediction is easier than estimation

- Efron 2020
- $x_1, \ldots, x_{25} \sim N(\mu, 1)$
- $\bar{x}$ sample mean and $\hat{x}$ sample median
- $x_{new} \sim N(\mu, 1)$

- But

$$E[(\mu - \hat{x})^2]/E[(\mu - \bar{x})^2] = 1.57$$

$$E[(x_{new} - \hat{x})^2]/E[(x_{new} - \bar{x})^2] = 1.02$$

Prediction is easier than attribution

- Efron 2020; Microarray study with N genes:

$$z_j \sim N(\delta_j, 1) \text{ for } j = 1,2,\dots,N$$

- $N_0: \delta_j = 0$ $(null\ genes)$
- $N_1: \delta_j > 0$ $(non{-}null\ genes)$
- New subject's microarray: $x_j \sim N(\pm\delta_j, 1)$ $\begin{cases} + sick \\ - healthy \end{cases}$
- Prediction possible if $N_1 = O\left(N_0^{1/2}\right)$
- Attribution requires if $N_1 = O(N_0)$

From Prediction to Attribution

- Prediction models are stable in the prediction space
- Assessed on their prediction properties (and little else)

- Superior flexibility (and hence a tendency to overfit)

- Predication algorithms are fancy look-up tables ... (explain)
  - Lack smoothness, compactness important for inference

- Inference and attribution are hard because they require valid internal structure (probabilistic structure)
  - A target
  - Uniqueness (identifiability)
  - Parameters (consistent, estimable, bias, efficiency, identifiable too)
  - This structure can be correct or not

From Prediction to Attribution

- Many prediction algorithms are not internally stable as we (statisticians) think about them
  - Neural Nets, Deep learning, Support vector machines, Gradient boosted models, etc.
  - Too many parameters (not consistent or identifiable)
  - Parameters don’t represent or converge to population quantities
  - Inference is effectively impossible (beyond removing features)

- Instead, use the prediction performance as a performance benchmark for more interpretable models
  - Trade some flexibility for some interpretability
  - Assess the loss of predictive power (may or may not be ‘real’)

Moving from Prediction to Inference

- Separate variable selection and model estimation steps

- For the variable selection step, some helpful tools are
  - Shrinkage
  - Regularization (L0, L1, L2 penalties)
  - Sparsity (assumptions)
  - Bagging / model averaging (?)

- For the inference step, aim for the true model or slightly larger
  - Inference under a larger model is often more forgiving but slightly less efficient

- The act of "variable selection" assumes that the feature space is known!
  - Modern prediction algorithms construct the feature space on-the-go, but variable selection algorithms do not.
  - Enrich your feature space (splines, interactions, creativity)
  - Use a variable selection algorithm to pair it back down

Variable Selection Algorithms

- Increased bias and loss of Type I Error control arise because the algorithm settles on the ‘wrong’ model for inference
  - Instead, judge these algorithms on how often they capture the correct set of features (‘‘capture rate’’ or ‘‘support recovery’’)
  - Use them to select a set of features for fitting

- Stepwise selection has relatively poor support recovery properties
  - Backwards stepwise selection generally poor Forward stepwise selection better in some cases
  - Both need larger samples size than more modern alternative
  - Time to retire these approaches

Variable Selection (gold standard)

- Why not try every model and see which does best?
  - Best Subset Selection
  - Gold standard
  - Compare based on a cross-validated criteria (!!)
    - AIC, BIC, etc ...
  - Computationally prohibitive (although getting better)
  - Recent advancements allows up to 1000 total features (2^10)

- Careful:
  - Algorithm very sensitive to target criteria (stability can be an issue)
  - Implementations trade comprehensive search for speed
  - Pick an inference oriented loss (log-likelihood loss)
  - AIC, BIC are penalized log-likelihood losses maximizing predictive ability does not always lead to consistent model selection for the lasso
  - When it works (and is stable), it tends to work well

Notation

- Response vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$
- Design matrix $\boldsymbol{X}$ $(n \times p)$ with p features (columns)

- Model: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$ where $\epsilon_i \sim N(0, \sigma^2)$ with $\boldsymbol{\beta}_0 \in \mathbb{R}^p$

- L0-norm: number of non-zero elements in $\boldsymbol{\beta}$ or

$$||\boldsymbol{\beta}||_0 = \sum_{j=1}^{p} I(\beta_j \neq 0)$$

- L1-norm: $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|$
- L2-norm: $||\boldsymbol{\beta}||_2 = \left(\sum_{j=1}^{p} \beta_j^2\right)^{1/2}$

Best subset Selection

- Beale, Kendall, and Mann 1967
- Hocking and Leslie 1967

- Search for the subset of k predictors that produces the best fit (defined in any way)
- Equivalent to L0 optimization

- Written as

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{ \frac{1}{2} \left|\left|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2 + \lambda ||\boldsymbol{\beta}||_0 \}$$

Best subset Selection

- Wen et al. 2020: efficient R package BeSS that is scalable to identify the best sub-model in seconds or a few minutes when p is around 10,000 ($\sim$ 2^13)

- Still computationally expensive (AWS, cluster,...)
- Different fit criteria lead to different models
  - Pick an inference oriented loss

- Is useful for "small" problems with clearly defined optimality criteria

Lasso

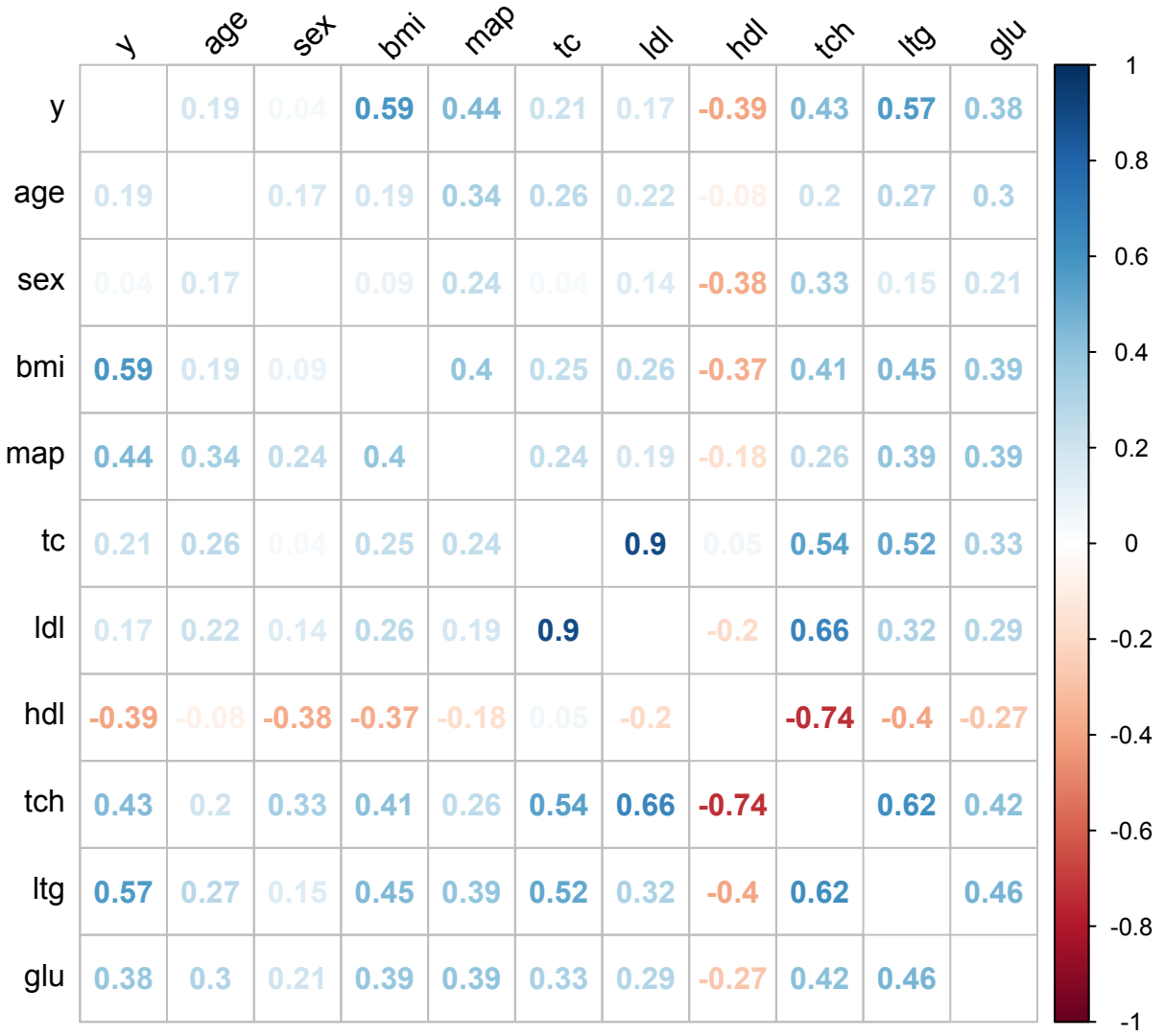
- L1 penalty encourages sparsity

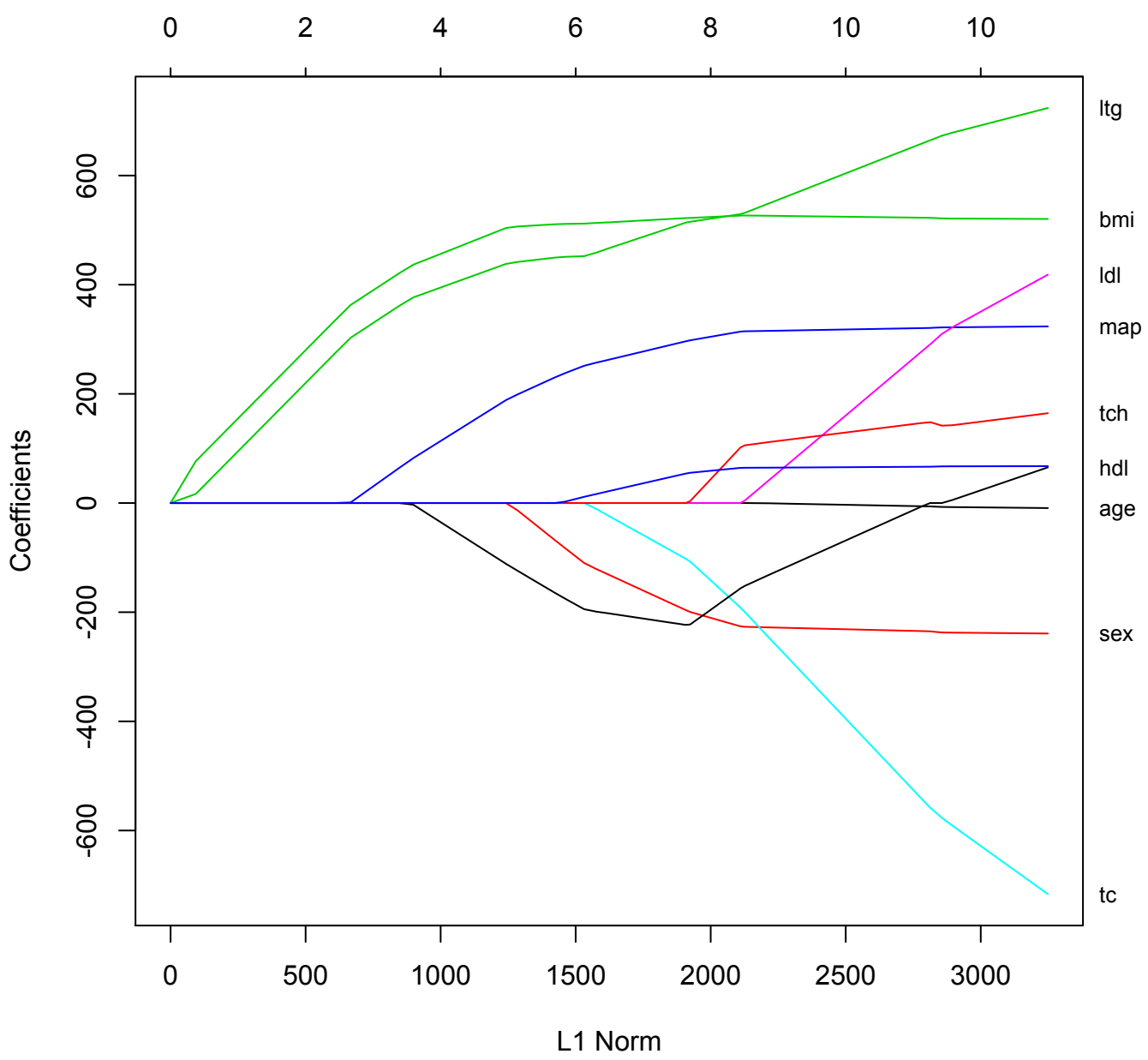
$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{ \frac{1}{2} \left|\left|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2 + \lambda ||\boldsymbol{\beta}||_1 \}$$

- Originally proposed for optimizing prediction performance
- Large literature on how to use it for model selection
- Pretty good option when sparsity assumption is realistic
- Tends to get the model size nearly correct, but not always with the correct variables
- Works well when
  - True effect sizes are large
  - Noise variables are not highly correlated with "true" features
  - Tendency to include noise features does not necessarily vanish in large N

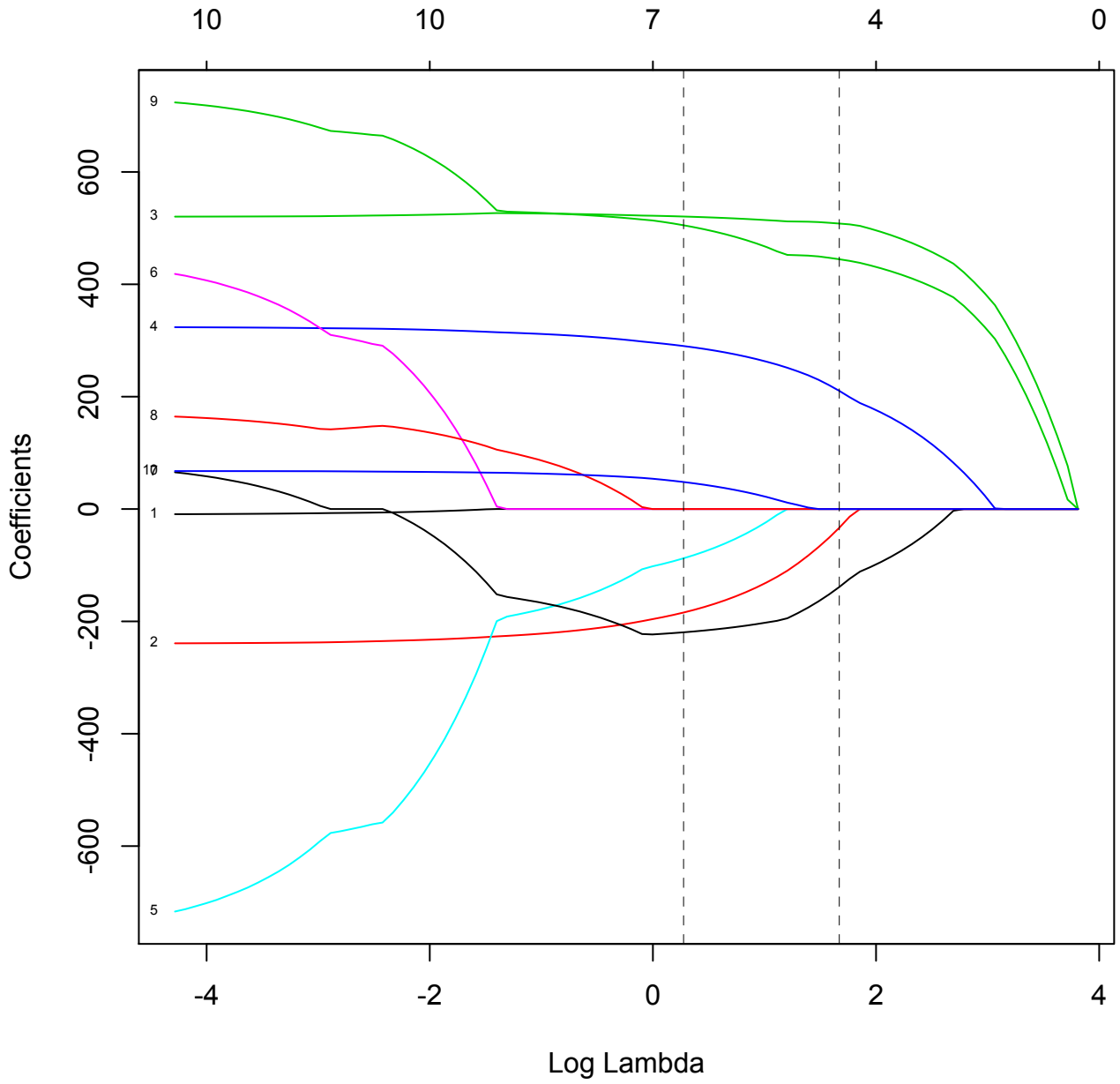
Diabetes Data

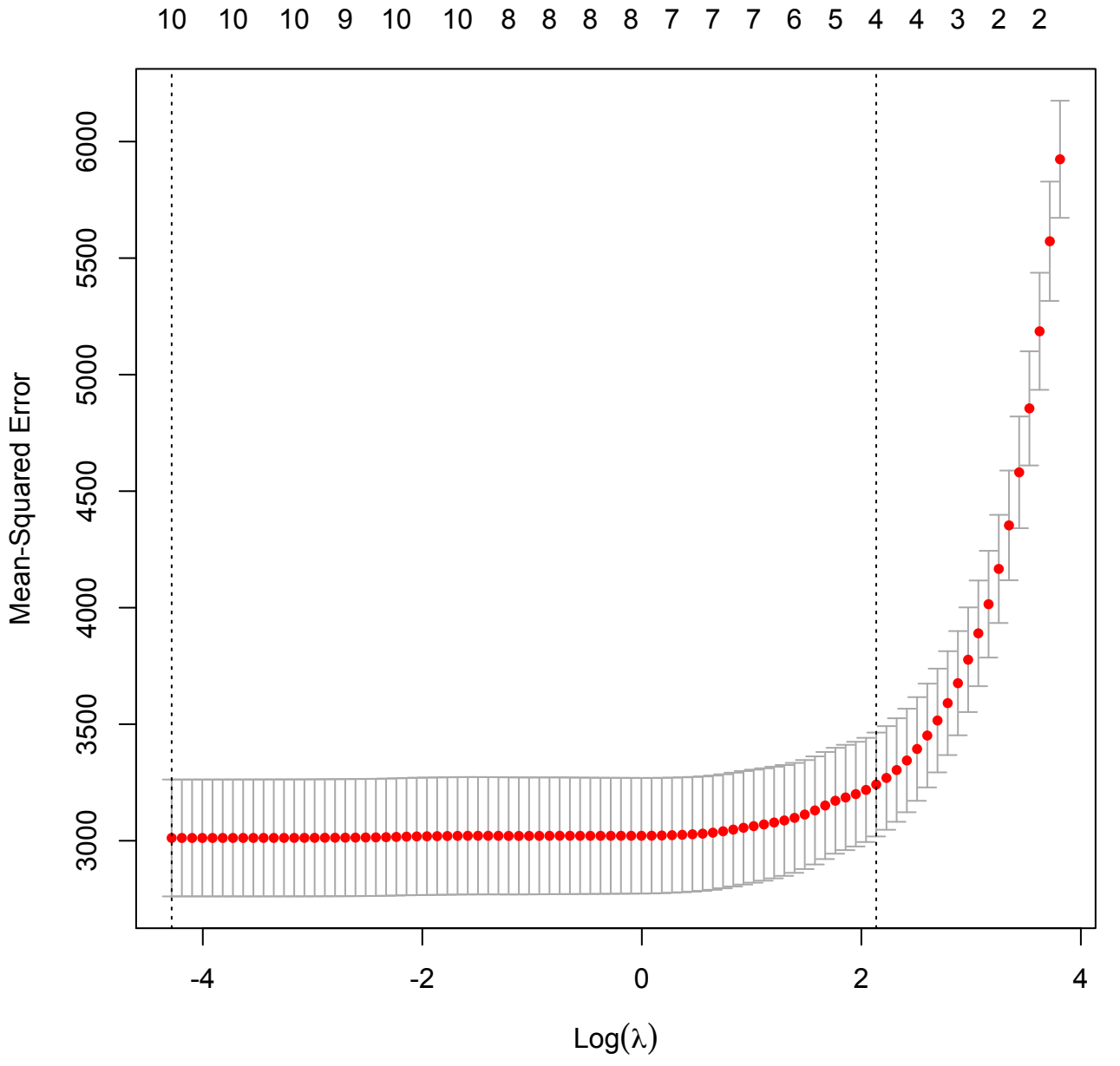
Y = Measure of Diabetes progress

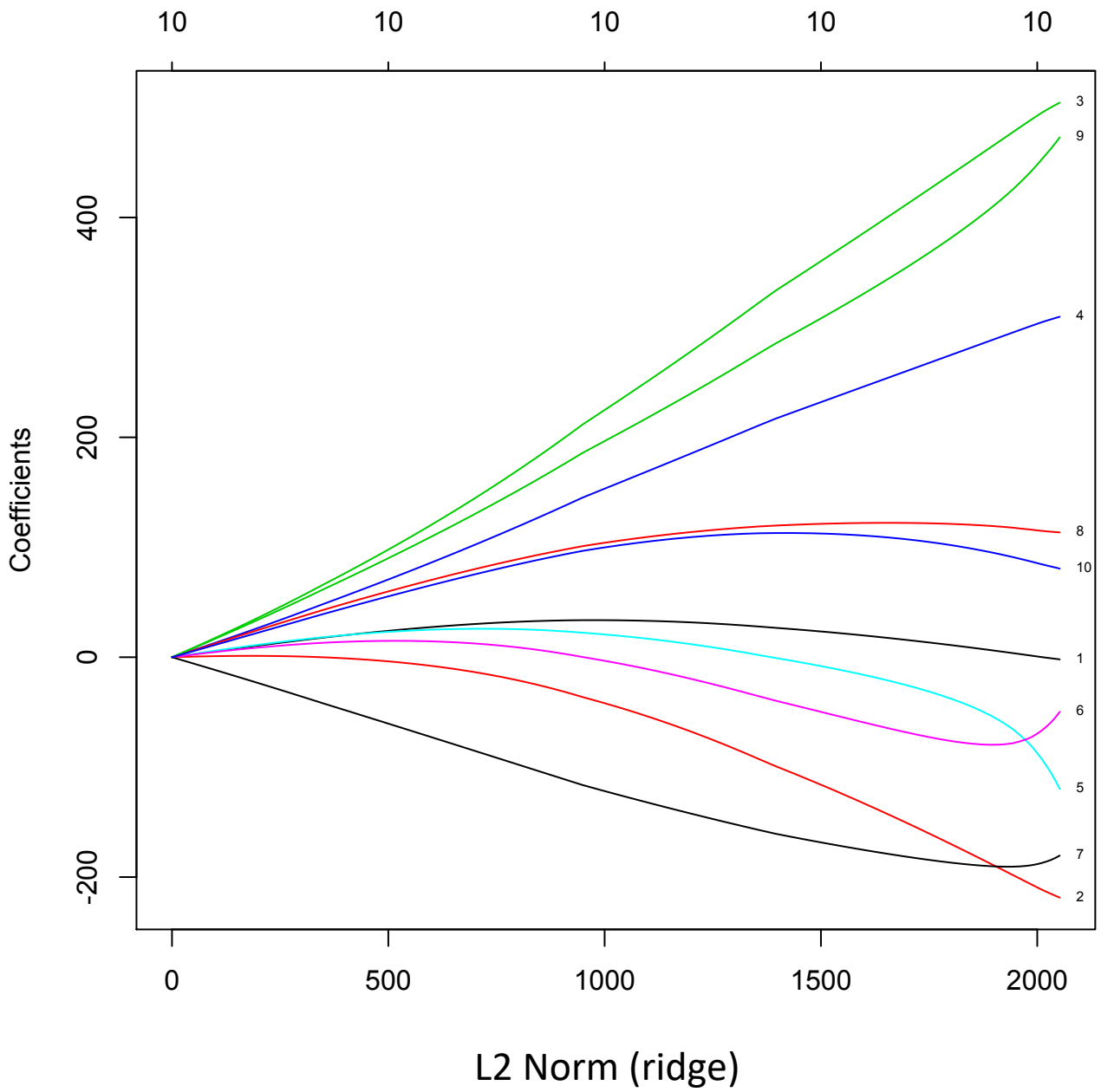


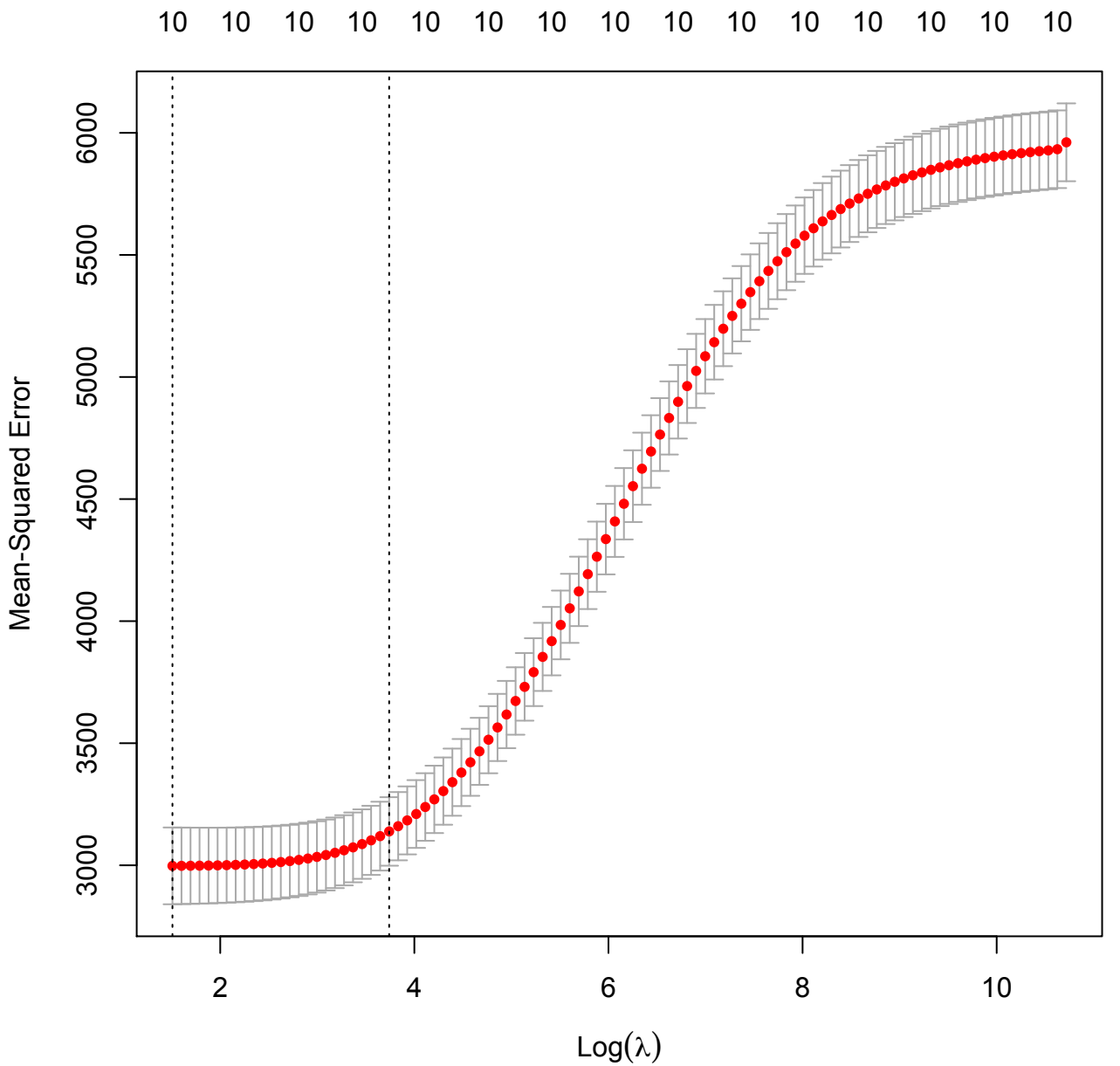


$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{ \frac{1}{2} \left|\left|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2 + \lambda ||\boldsymbol{\beta}||_1 \}$$


10
10
7
4
0
9
3
6
4
8
10
1
2
5
600
400
200
0
-200
-400
-600
Coefficients
-4
-2
0
2
4
Log Lambda


10 10 10 9 10 10 8 8 8 8 7 7 7 6 5 4 4 3 2 2
6000
5500
5000
4500
4000
3500
3000
Mean-Squared Error
-4
-2
0
2
4
Log(λ)


10 10 10 10 10
Coefficients
400
200
0
-200
3
9
4
8
10
1
6
5
7
2
0 500 1000 1500 2000
L2 Norm (ridge)


10 10 10 10 10 10 10 10 10 10 10 10 10 10 10
Mean-Squared Error
6000
5500
5000
4500
4000
3500
3000
2 4 6 8 10
Log(λ)

Adaptive Lasso

- Introduces weights in the L1 penalty term

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{ \frac{1}{2} \left|| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \right||_2^2 + \lambda_n ||\widehat{\boldsymbol{w}}\boldsymbol{\beta}||_1 \}$$

- where $\widehat{\boldsymbol{w}} = 1/|\widehat{\boldsymbol{\beta}}|^{\gamma}$
- $\gamma > 0$ is a tuning parameter and $\widehat{\boldsymbol{\beta}}$ is a root-n-consistent estimator of $\boldsymbol{\beta}^*$, for example, an OLS estimator, or a ridge estimator

- Adaptive lasso has large samples oracle properties for support recovery with proper choice of lambda
- But, hard to specify all the tuning parameters in practice

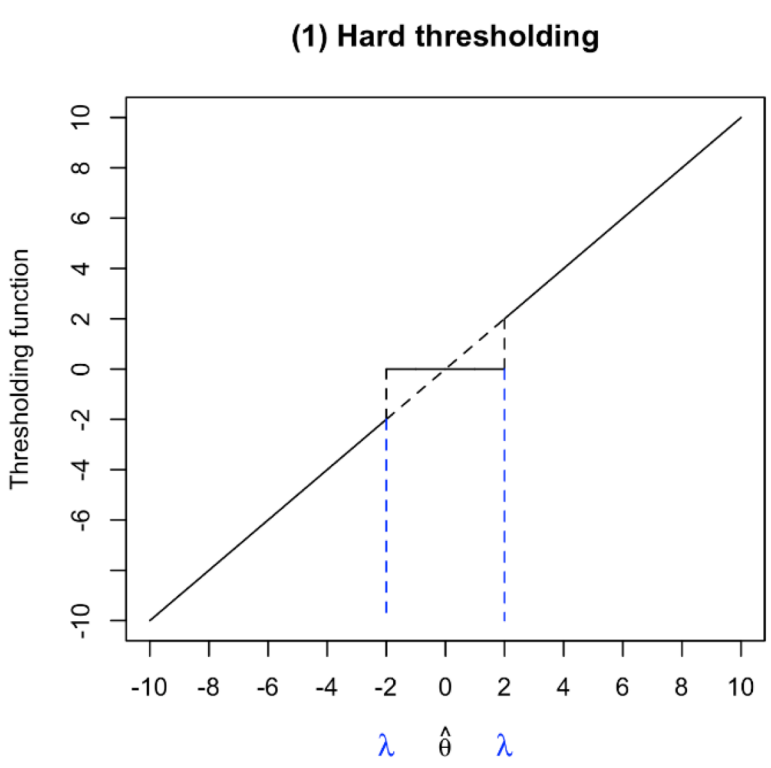
SCAD and MC+

- Smoothly clipped absolute deviation (SCAD)
- Minimax concave penalty with penalized linear unbiased selection (MC+)

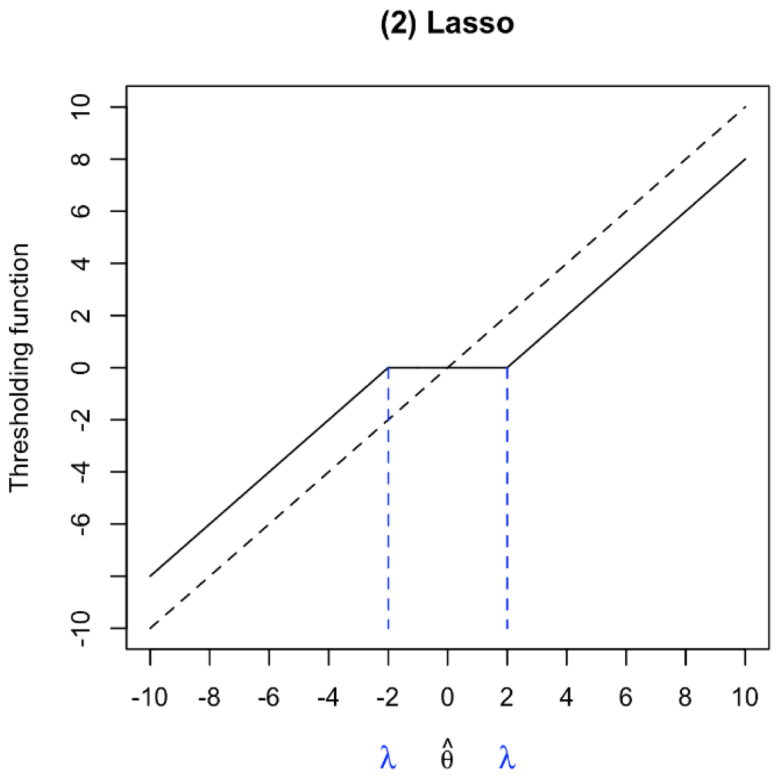
- Designed to bridge the L0 and L1 penalties
- Both use nonconvex penalties, unlike lasso

- SCAD preforms well when sparsity assumptions do not apply and p>>n and prediction is the focus
- MC+ is pretty good at support recovery and less restrictive than lasso
- MC+ is harder to implement and tune, but worth the effort

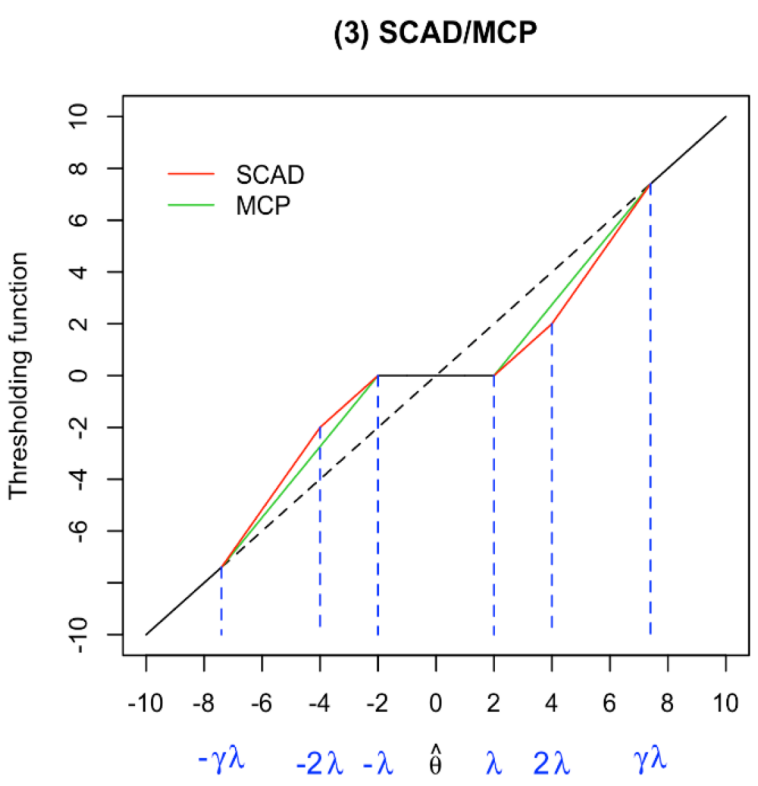
Visualization of Selection Algorithms



L0 Penalization



L1 Penalization



L0 / L1 Combination

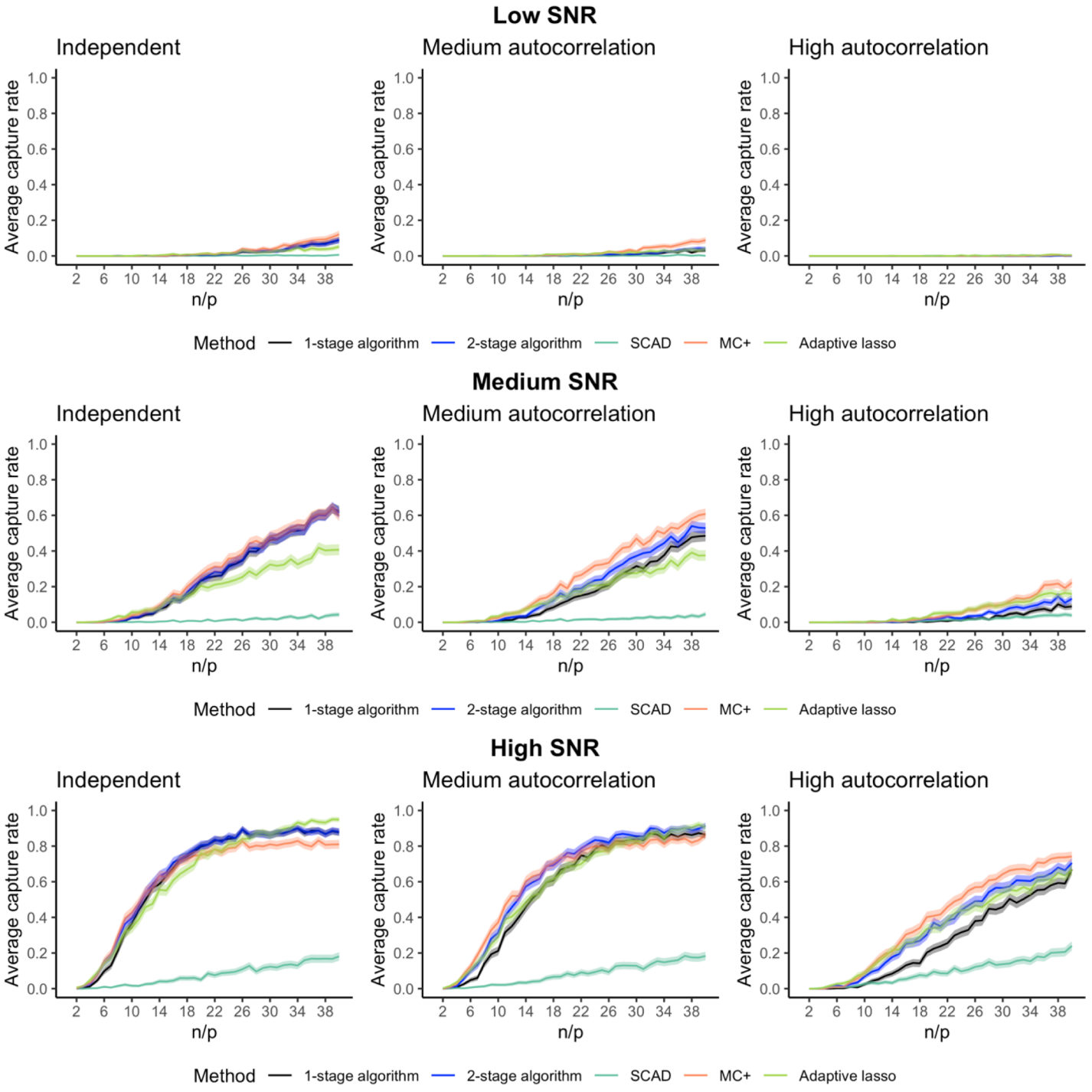
Note: none of these approaches explicity accounts for the variability in the estimate

- Simple simulations

- Linear model
- Feature correlation varies
- SNR varies
- p=50; 10 true signals

- Best subset selection not shown
  - competitive when able to run
- Lasso, stepwise do very poorly

- Zuo, Stewart, Blume(2020)
- https://arxiv.org/abs/2012.07941



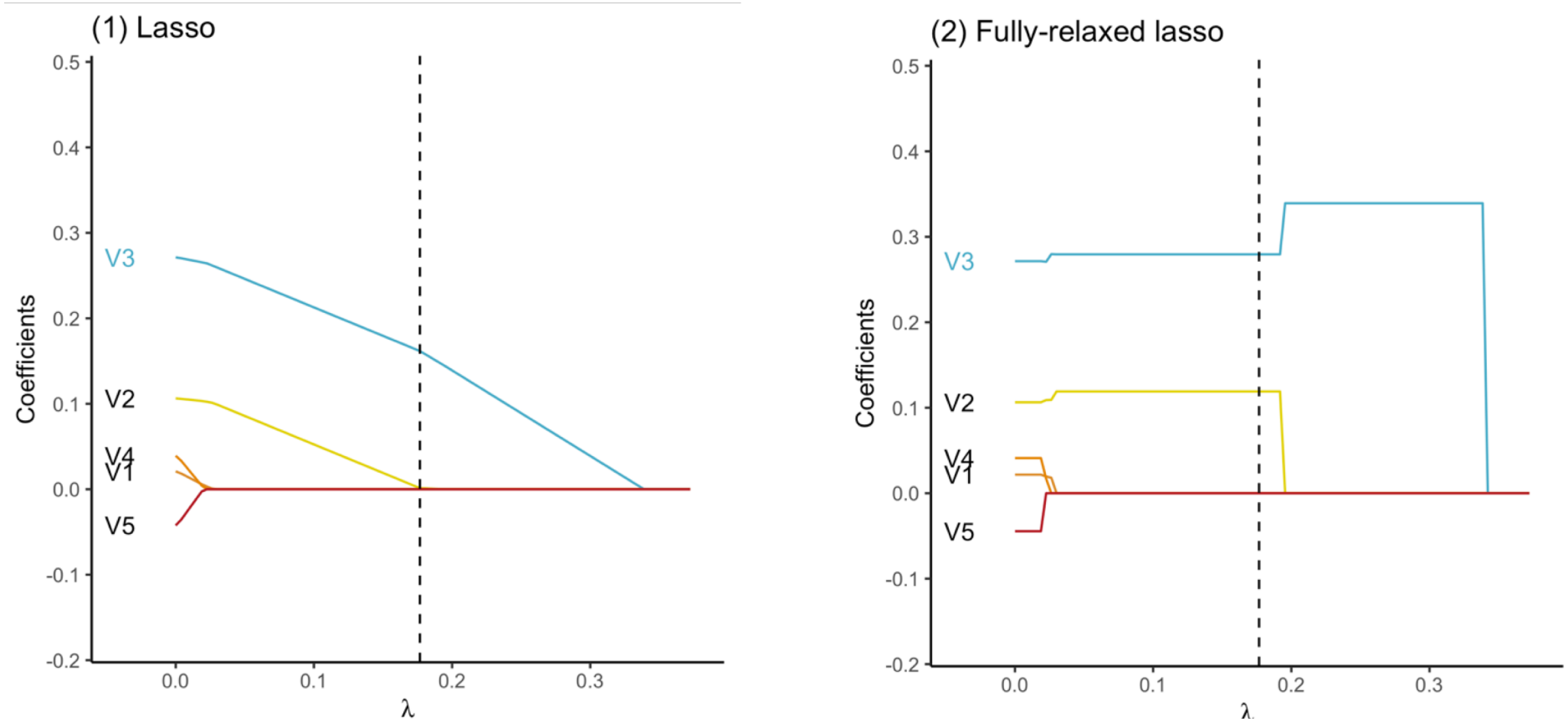
Relaxing

- Lasso, Adaptive Lasso, SCAD, MC+ all yield biased parameter estimates, even when they identify the correction subset of features

- A reliable strategy is to "fully relax" the model, by refitting the model with the variables that survived regularization step

- This has several advantages
  - Relaxing is inferentially optimal
  - Used with a lasso often yields a slightly larger model
  - Preserves the original interpretation of the coefficients

Visualization of Relaxed Lasso



Relaxed Lasso Model for Diabetes Data

	lasso.min	relaxo	ols
• (Intercept)	152.133484	152.13348	152.13348
• age	-9.220679	.	-10.01220
• sex	-239.063410	-232.98392	-239.81909
• bmi	520.458638	526.97094	519.83979
• map	323.616100	315.68769	324.39043
• tc	-716.483385	-146.49666	-792.18416
• ldl	418.448411	.	476.74584
• hdl	65.386847	-235.53890	101.04457
• tch	164.562823	.	177.06418
• ltg	723.735577	540.73661	751.27932
• glu	67.540420	72.25496	67.62539

Example: Treat Model 2.0

- Aim
  - Develop a predictive model for lung cancer in patients with an indeterminate pulmonary nodules who are referred for surgical evaluation.
  - Three existing models
    - Mayo Model (logistic model)
    - Herder Model (logistic model)
    - Treat 1.0 (logistic model)
  - Concerns
    - Poor accuracy as measured by AUC and Brier score
    - Poor calibration
- 1401 patients (415 with lung CA, 986 without)
- 14 features / covariates

Clinical practice subgroups and cohorts

Clinical Subgroup				Cohort			
Clinical Subgroup	Location	n	Cancer Prevalence	Cohort	Location	n	Cancer Prevalence
1	Pulmonary nodule clinic	374	42%	1	Tennessee	258	34%
				2	Arizona	116	60%
2	Thoracic surgery clinic	553	73%	3	Tennessee	492	72%
				4	Massachusetts	61	82%
3	Surgical resection	474	90%	5	Tennessee	216	93%
				6	Virginia	258	88%

Features (14)

- Age (age): patient age in years
- Gender (gender): 0 = female, 1 = male
- BMI (bmi_new): body mass index
- Previous cancer (prev_cancer): history of previous cancer? (0 = no, 1 = yes)
- Smoker (smoker): current smoker? (0 = no, 1 = yes)
- Pack years (packs): pack years of smoking
- FEV1 (fev1): pre-operative forced expiratory volume in first second
- Nodule size (size): size of nodule in mm
- Spiculation (spicul): nodule with spiculated edge? (0 = no, 1 = yes)
- Growth (growthcat): lesion/nodule growth (0 = no, 1 = yes, 2 = "missing")
- Upper lobe (upperlobe): lesion/nodule in upper lobe? (0 = no, 1 = yes)
- Symptoms (anysympt): presence of pre-operative symptoms, such as hemoptysis, short of breath, weight loss, fatigue, pain, pneumonia? (0 = no, 1 = yes)
- Pet avidity (petavid): FDG-PET avidity (<2.5 or >= 2.5 SUV).
- Clinical group (group): "referral type" (Pulm = pulmonary nodule clinic, Thoracic = thoracic surgery clinic, Surgery = surgical resection)

The Data

	Population (n=1401)	Missing data (n)
Age mean (SD)	64 (11.7)	0
Male gender	58% (813)	0
BMI mean (SD), kg/m^2	28 (6.1)	9% (129)
Smokers	82% (1149)	0.6% (8)
Pack years mean (SD)	41 (34)	3% (42)
Symptomatic	45% (182)	11% (149)
Previous cancer	31% (438)	0.2% (3)
FEV1 mean (SD), %	77 (20)	20% (277)
Nodule size (mm)	25 (17)	1% (18)
Spiculation	41% (578)	4% (60)
Growth	35% (490)	45% (635)
Upper lobe	53% (743)	2% (23)
PET avid	65% (904)	22% (311)
Cancer	42% (986)	0

Plan of Attack

- Expand the feature space and saturate it
  - Restricted cubic splines for continuous variables
  - Interactions (3,4,5-way etc.)
  - Pair down by regularization, then fill out model so interpretable

- Develop multiple candidate models for prediction
  - Logistic regression (regularized)
    - Main effects + interactions (prespecified)
    - Shrink interactions
    - Shrink all
  - Random forest
  - Gradient booted model

- Cross-validate models (10-fold, repeated 50 times)
- Relax and "correct" best preforming models with a focus on interpretability
- Assess the reduction in predictive performance

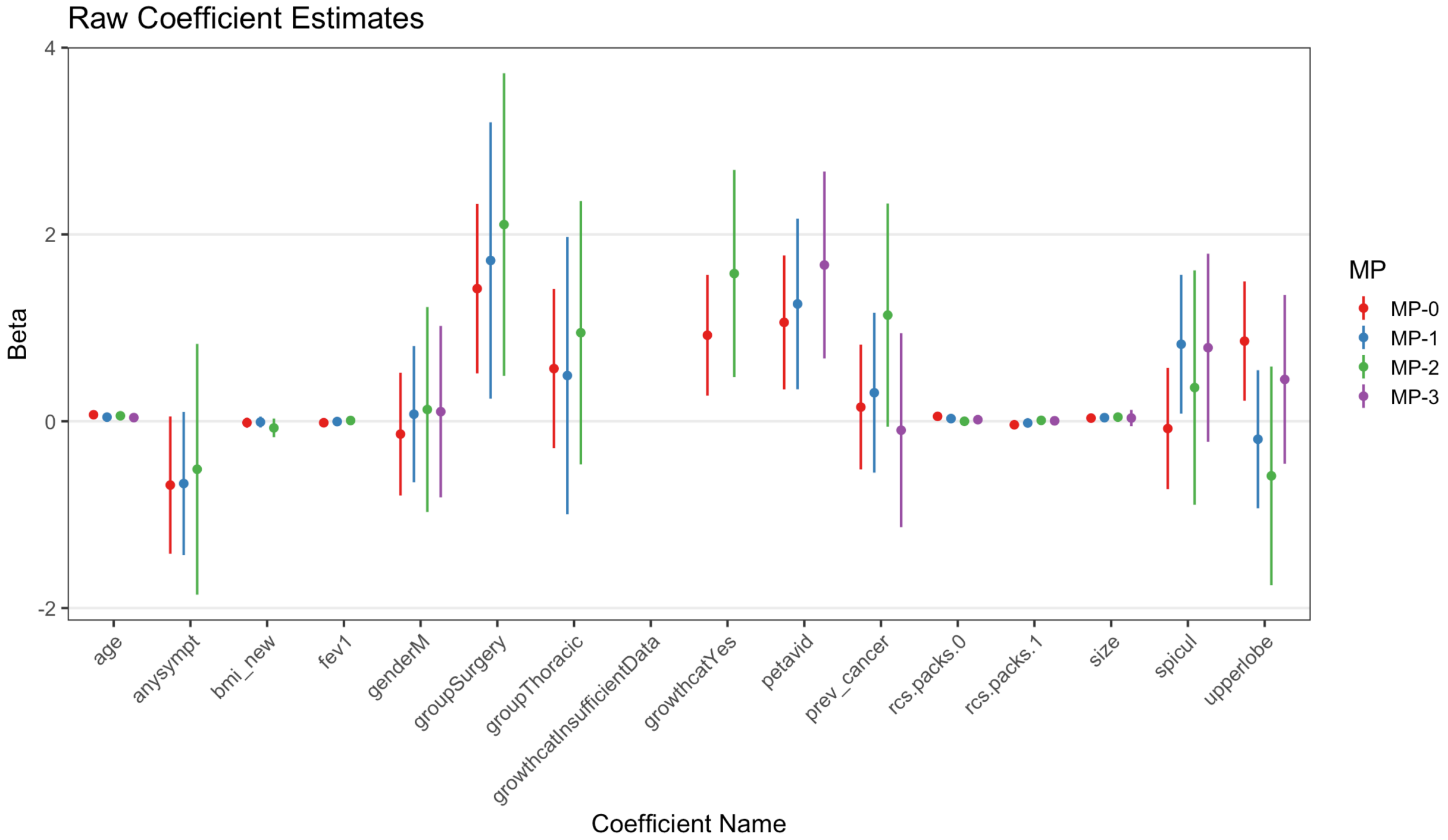
Handling missing data

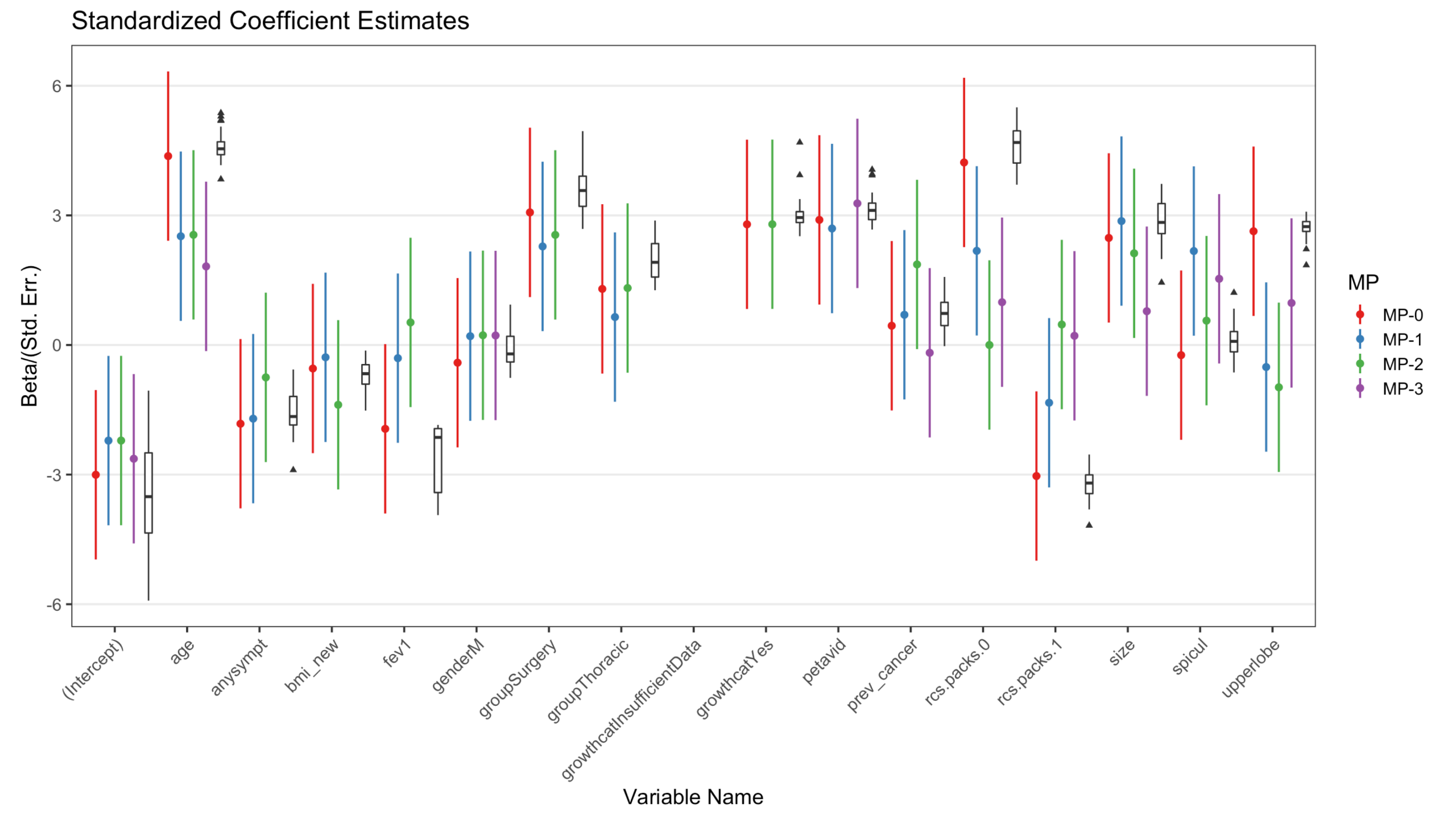
- Use the pattern submodel approach by Mercaldo & Blume (2018)
- Implement the plan of attach in subset with complete data
- Repeated in each subset of missing data pattern

- Relaxes the missing-at-random assumptions
  - Coefficient interpretation now depends on missingness pattern
  - Pro: improve predictive ability (sometimes a lot)
  - Con: loses simple coefficient interpretation

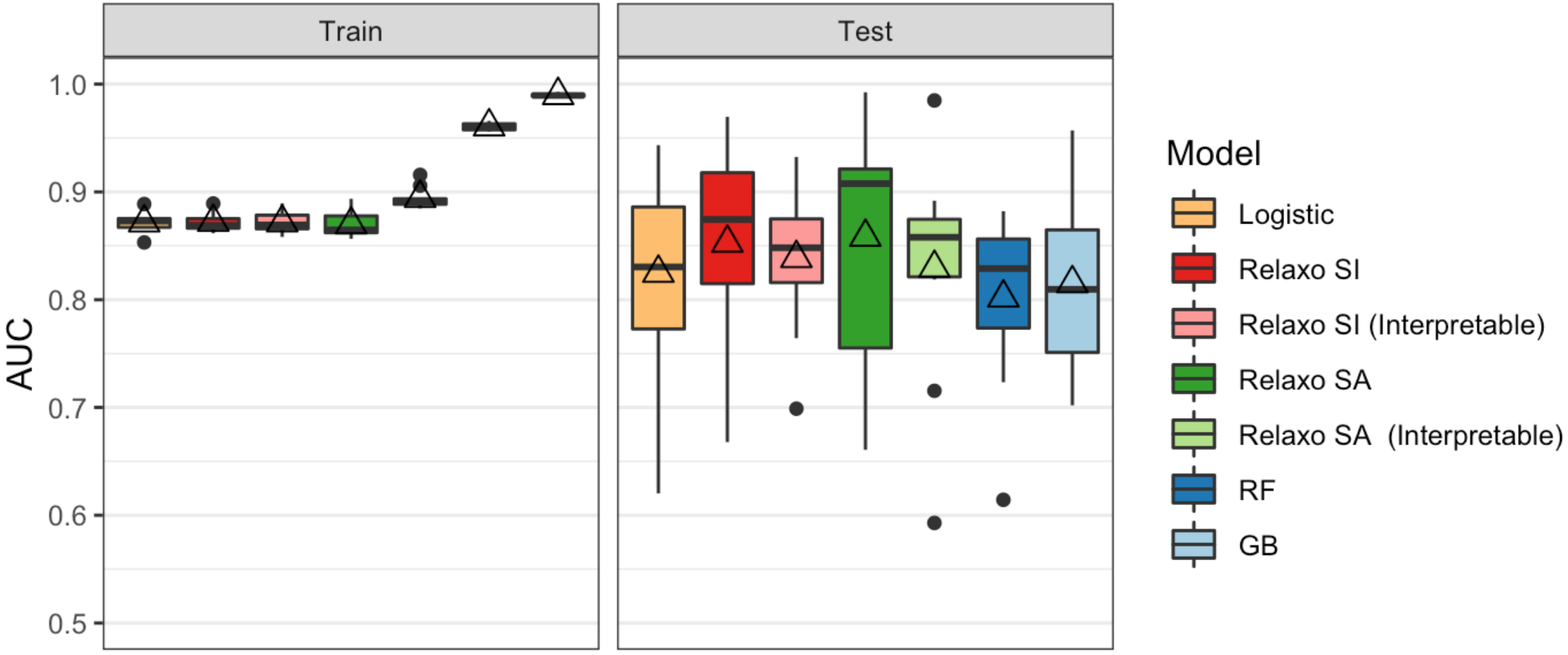
Missing Data Patterns

Pattern ID	frequency	age	anysympt	bmi_new	fev1	gender	group	growthcat	packs	petavid	prev_cancer	size	spicul	upper lobe
MP-0	471	.	.	.	.	.	.	.	.	.	.	.	.	.
MP-1	356	.	.	.	.	.	.	X	.	.	.	.	.	.
MP-2	120	.	.	.	.	.	.	.	.	X	.	.	.	.
MP-3	116	.	X	X	X	.	.	X	.	.	.	.	.	.
MP-4	50	.	.	.	X	.	.	.	.	X	.	.	.	.
MP-5	41	.	.	.	.	.	.	X	.	X	.	.	.	.
MP-6	35	.	.	.	X	.	.	.	.	.	.	.	.	.
MP-7	23	.	.	.	X	.	.	X	.	.	.	.	.	.
MP-8	18	.	.	.	.	.	.	X	.	X	.	.	X	.

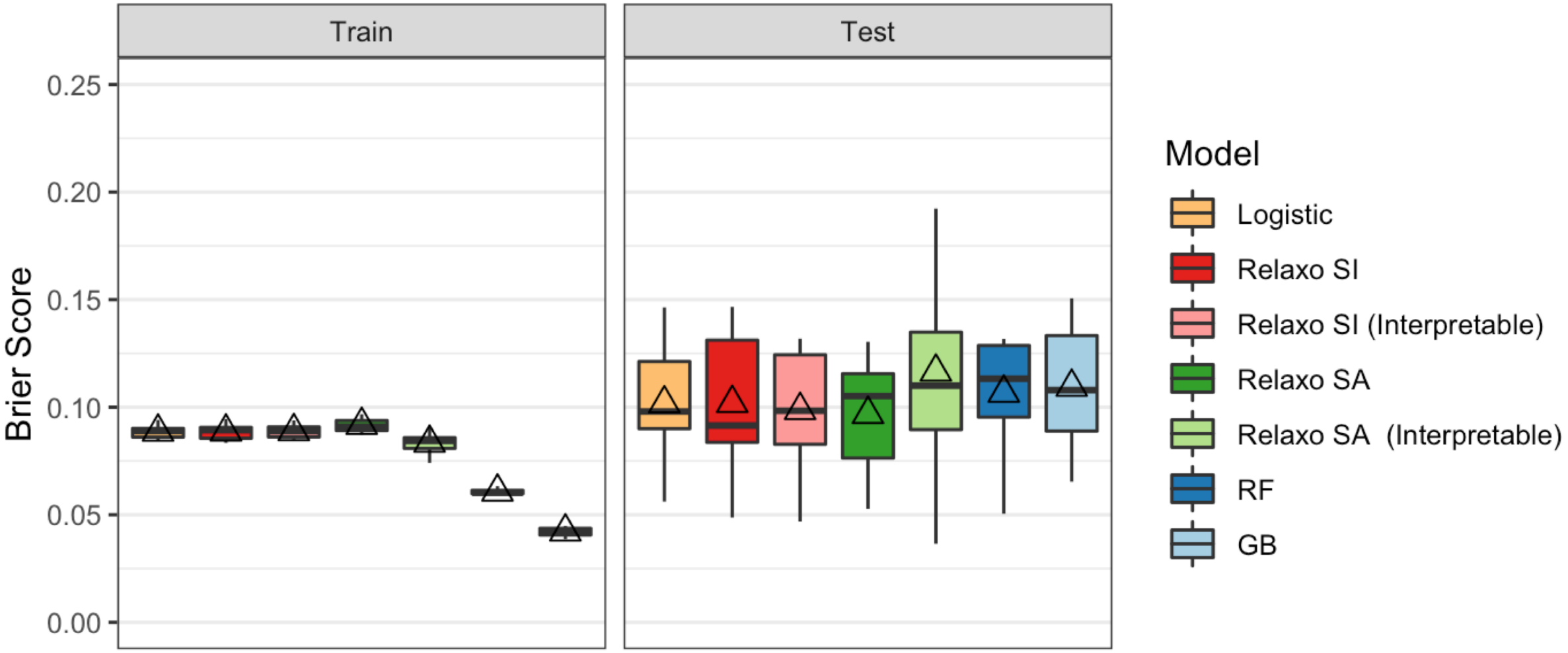

Raw Coefficient Estimates
Beta
4
2
0
-2
age
anysympt
bmi_new
fev1
genderM
groupSurgery
groupThoracic
growthcatInsufficientData
growthcatYes
petavid
prev_cancer
rcs.packs.0
rcs.packs.1
size
spicul
upperlobe
Coefficient Name
MP
MP-0
MP-1
MP-2
MP-3


Standardized Coefficient Estimates
Beta/(Std. Err.)
6
3
0
-3
-6
(Intercept)
age
anysympt
bmi_new
fev1
genderM
groupSurgery
groupThoracic
growthcatInsufficientData
growthcatYes
petavid
prev_cancer
rcs.packs.0
rcs.packs.1
size
spicul
upperlobe
Variable Name
MP
MP-0
MP-1
MP-2
MP-3

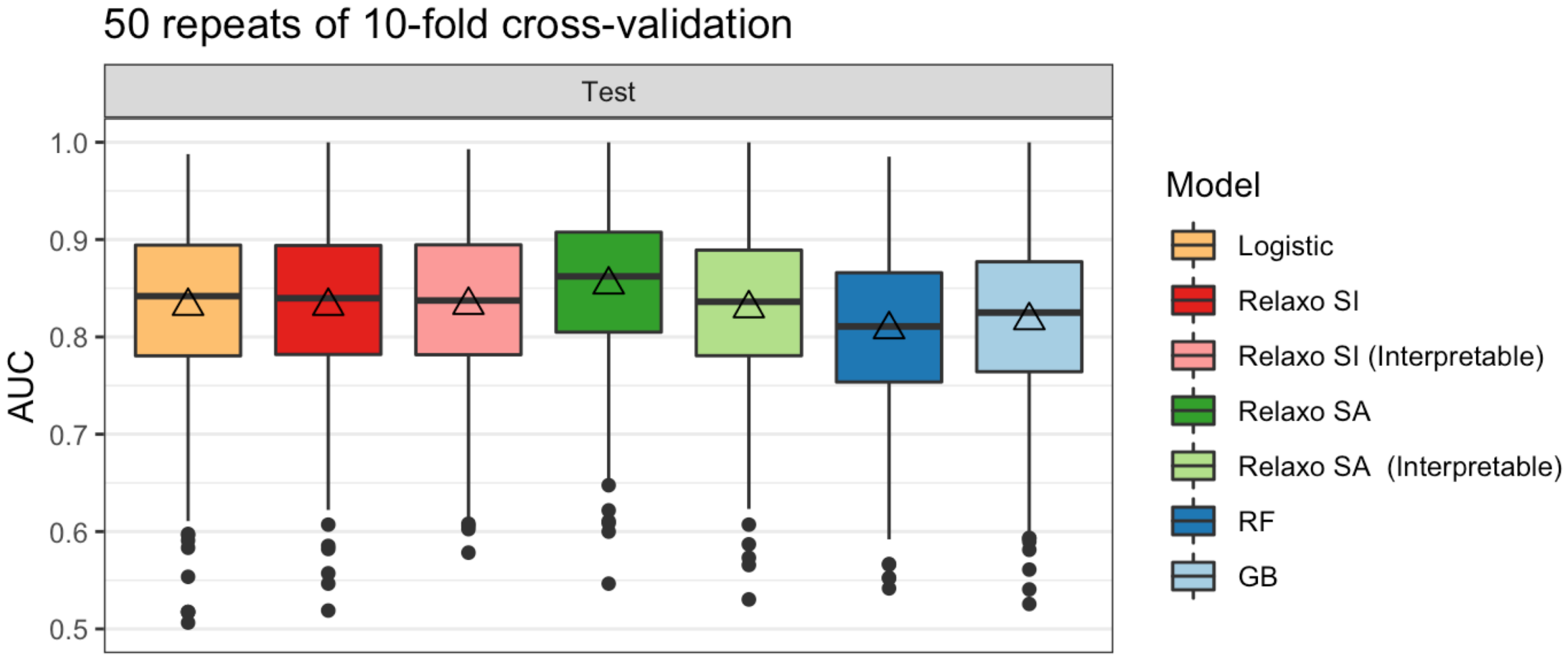
Cross-Validated Performance (AUC)

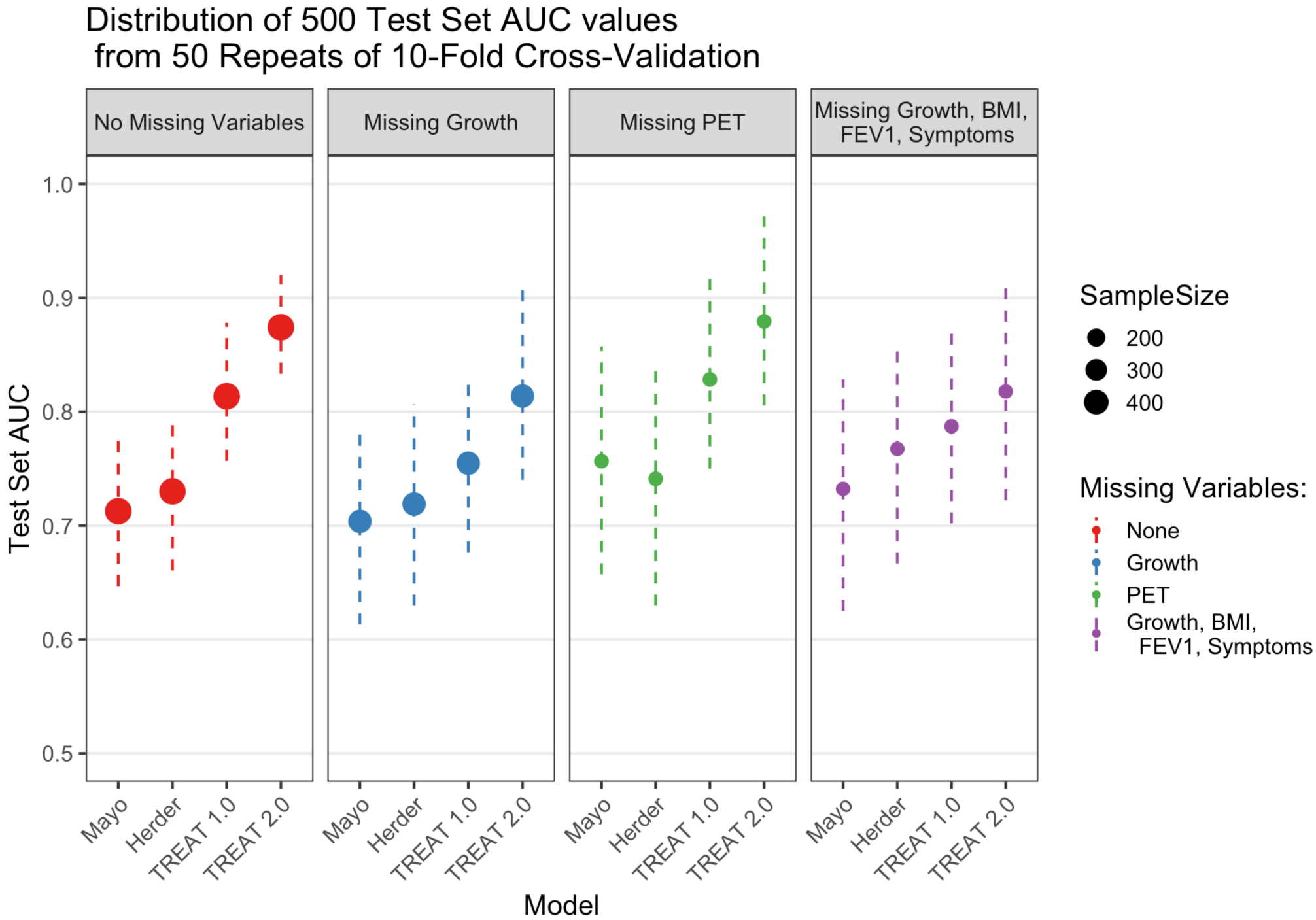


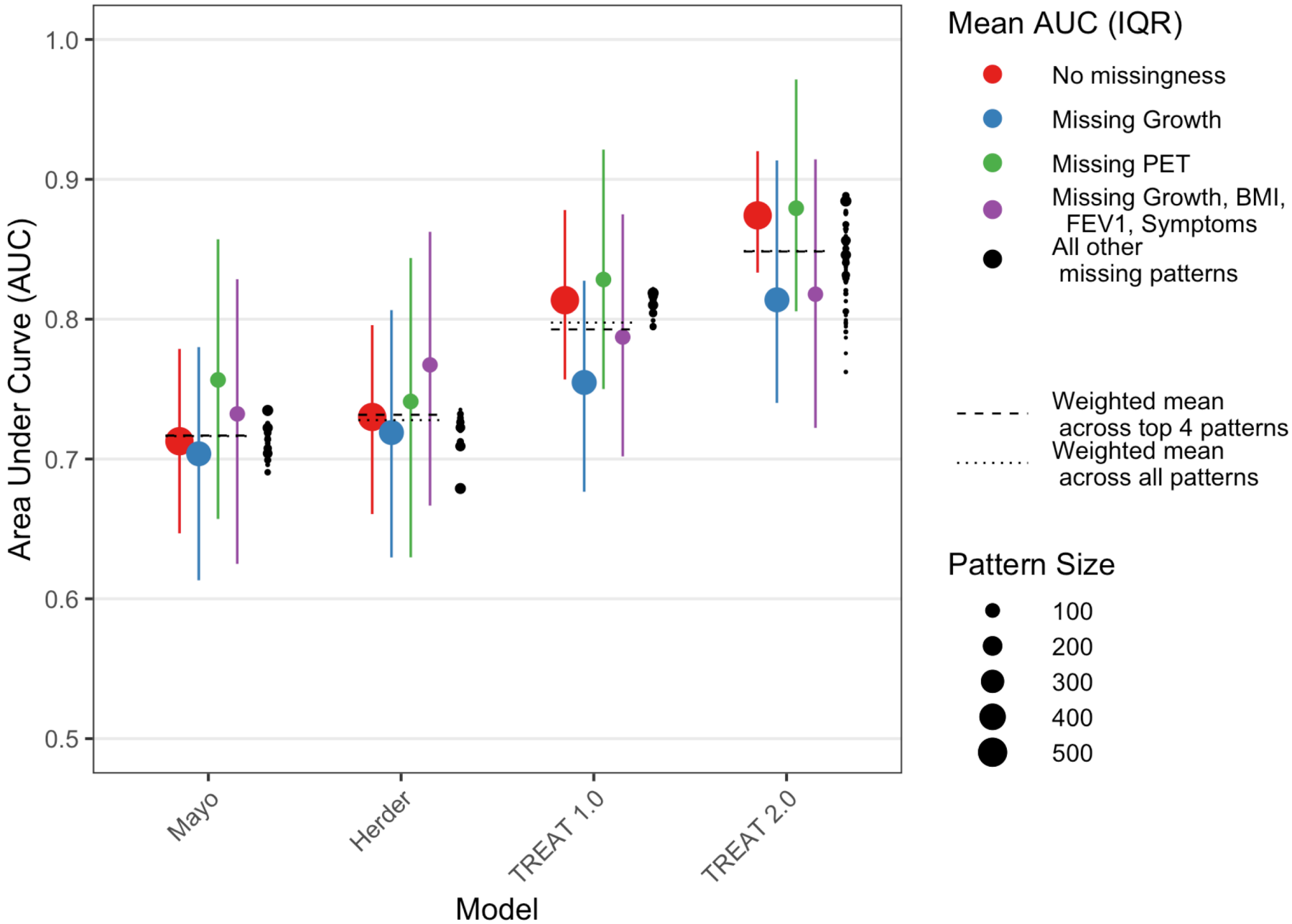
Cross-Validated Performance (Brier)



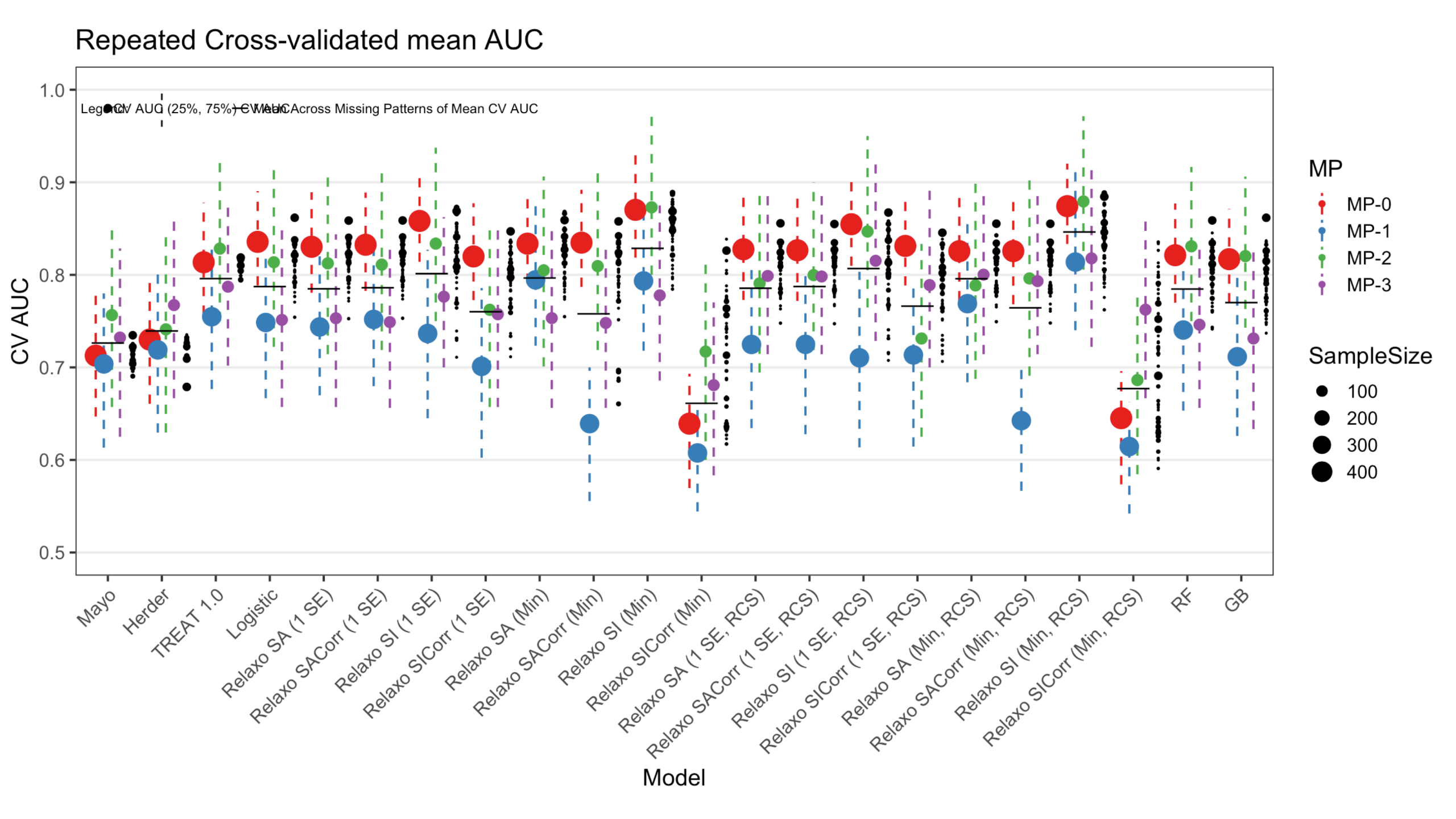
Repeated Cross-Validation (50 x 10-fold)

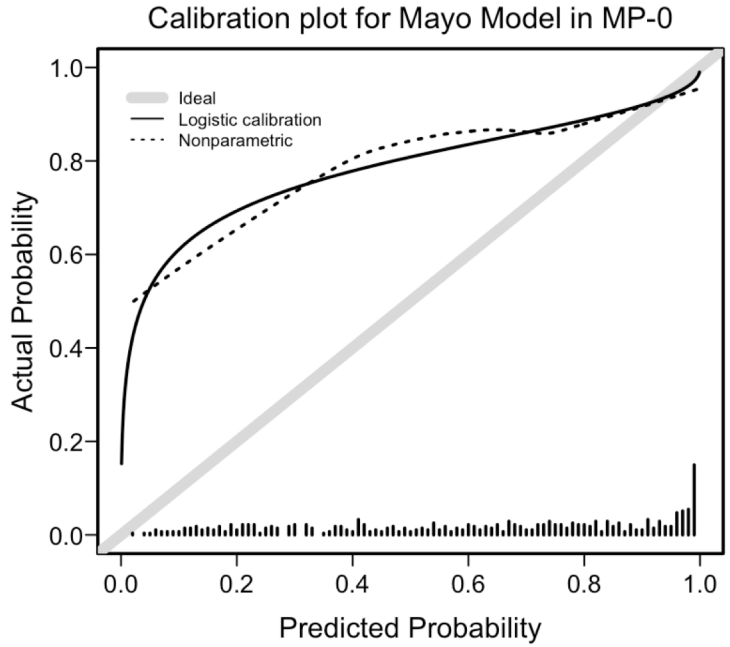


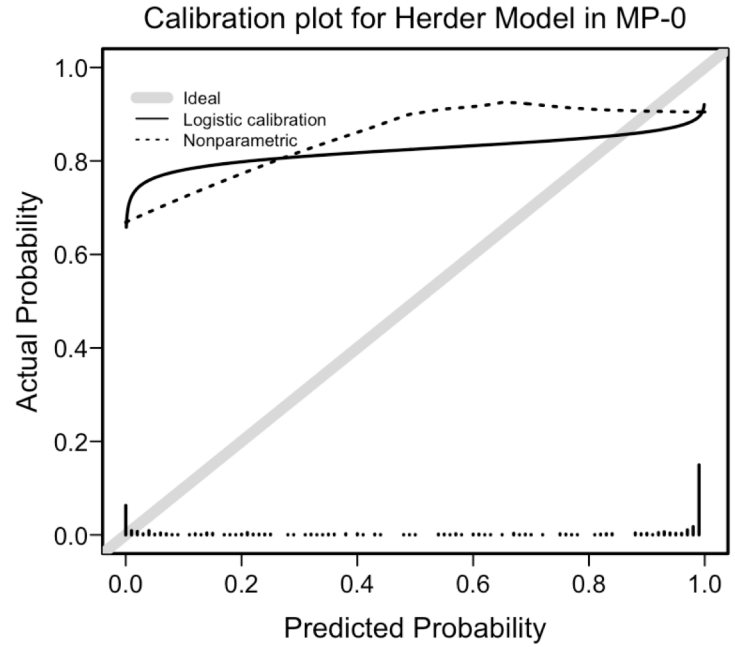

Distribution of 500 Test Set AUC values
from 50 Repeats of 10-Fold Cross-Validation
No Missing Variables
Missing Growth
Missing PET
Missing Growth, BMI, FEV1, Symptoms
Test Set AUC
1.0
0.9
0.8
0.7
0.6
0.5
Mayo
Herder
TREAT 1.0
TREAT 2.0
Model
SampleSize
200
300
400
Missing Variables:
None
Growth
PET
Growth, BMI, FEV1, Symptoms

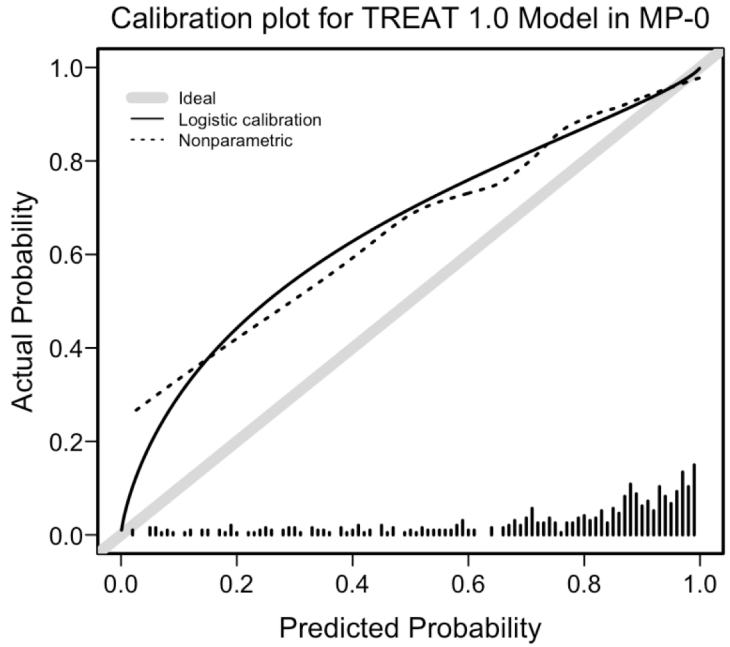


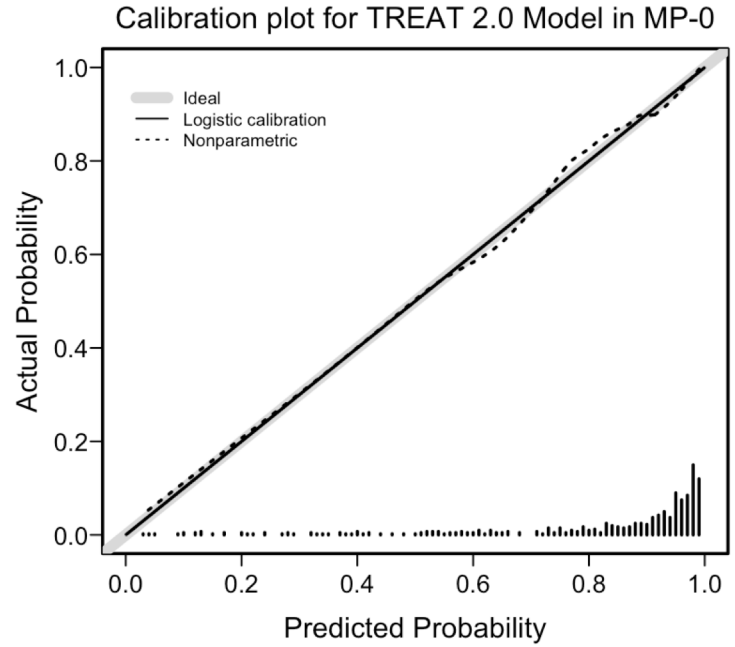
Distribution of Test Set AUC values from 50 repeats of 10-fold Cross-validation


Repeated Cross-validated mean AUC
Legend: CV AUC (25%, 75%) CV Mean Across Missing Patterns of Mean CV AUC
CV AUC
1.0
0.9
0.8
0.7
0.6
0.5
Mayo
Herder
TREAT 1.0
Logistic
Relaxo SA (1 SE)
Relaxo SACorr (1 SE)
Relaxo SI (1 SE)
Relaxo SICorr (1 SE)
Relaxo SA (Min)
Relaxo SACorr (Min)
Relaxo SI (Min)
Relaxo SICorr (Min)
Relaxo SA (1 SE, RCS)
Relaxo SACorr (1 SE, RCS)
Relaxo SI (1 SE, RCS)
Relaxo SICorr (1 SE, RCS)
Relaxo SA (Min, RCS)
Relaxo SACorr (Min, RCS)
Relaxo SI (Min, RCS)
Relaxo SICorr (Min, RCS)
RF
GB
Model
MP
MP-0
MP-1
MP-2
MP-3
SampleSize
100
200
300
400


Calibration plot for Mayo Model in MP-0
Ideal
Logistic calibration
Nonparametric
Actual Probability
Predicted Probability
0.0
0.2
0.4
0.6
0.8
1.0


Calibration plot for Herder Model in MP-0
Ideal
Logistic calibration
Nonparametric
Actual Probability
Predicted Probability
0.0
0.2
0.4
0.6
0.8
1.0


Calibration plot for TREAT 1.0 Model in MP-0
Ideal
Logistic calibration
Nonparametric
Actual Probability
Predicted Probability
0.0
0.2
0.4
0.6
0.8
1.0


Calibration plot for TREAT 2.0 Model in MP-0
Ideal
Logistic calibration
Nonparametric
Actual Probability
Predicted Probability
0.0
0.2
0.4
0.6
0.8
1.0

Questions

- Thank you for your attention

- Special thanks to
  - Valerie Welty (Treat 2.0 analysis)
  - Yi Zuo (Simulations of Variable selection properties)