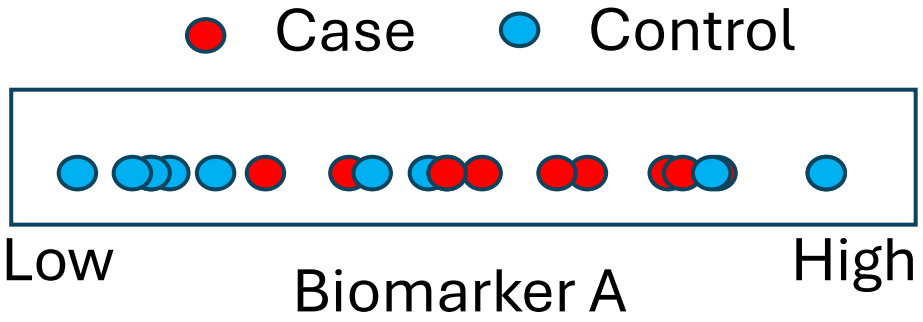
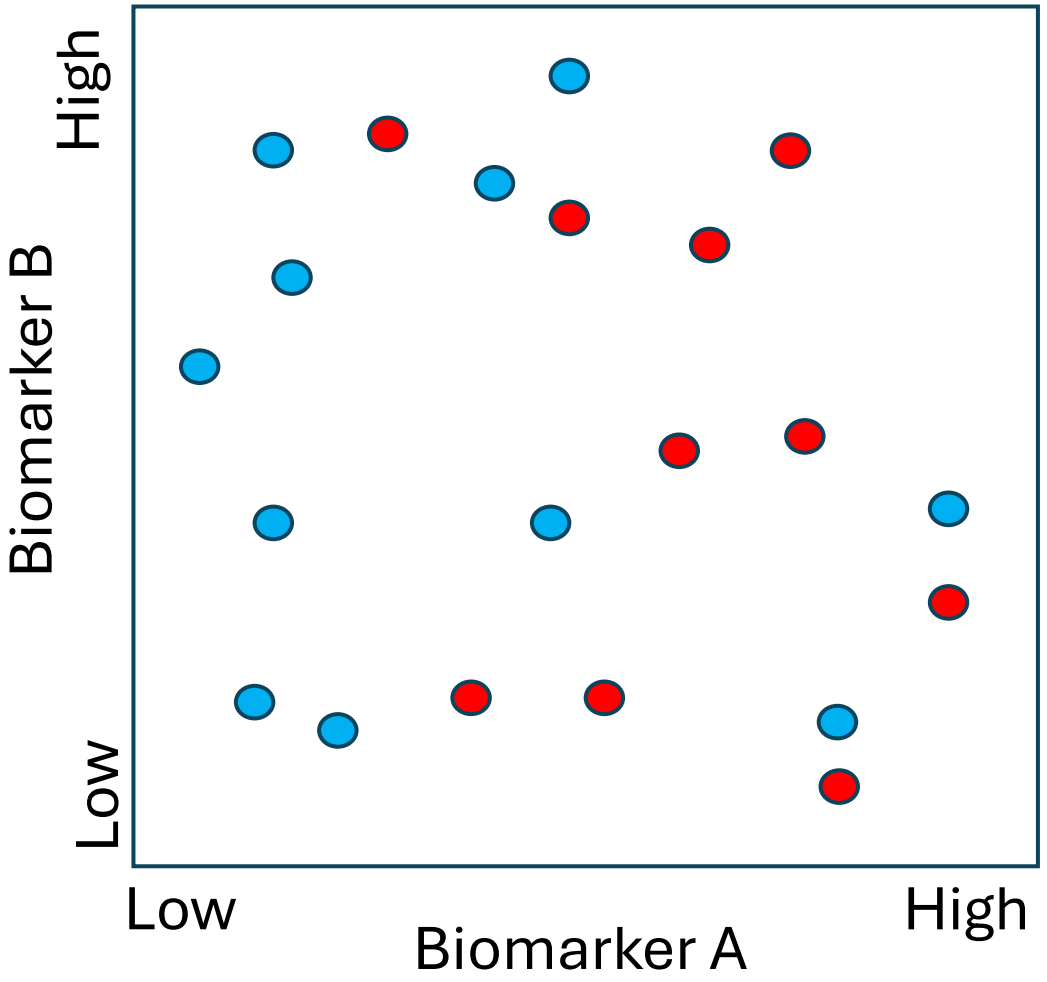
Statistical analysis for NAPPA

Yunro Chung, Ph.D.
Assistant Professor
Arizona State University

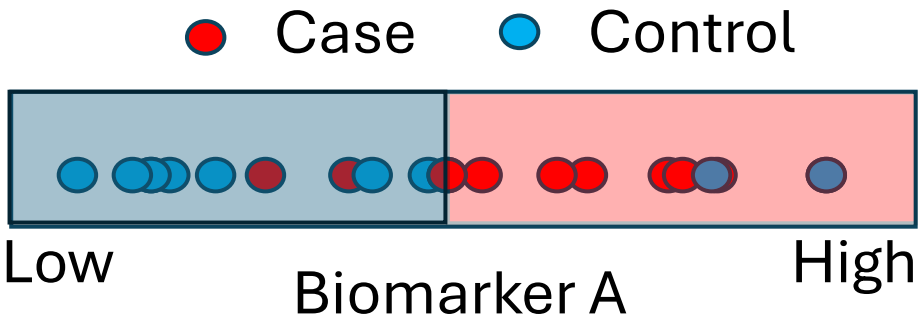
Overview

Overview: combination of biomarkers

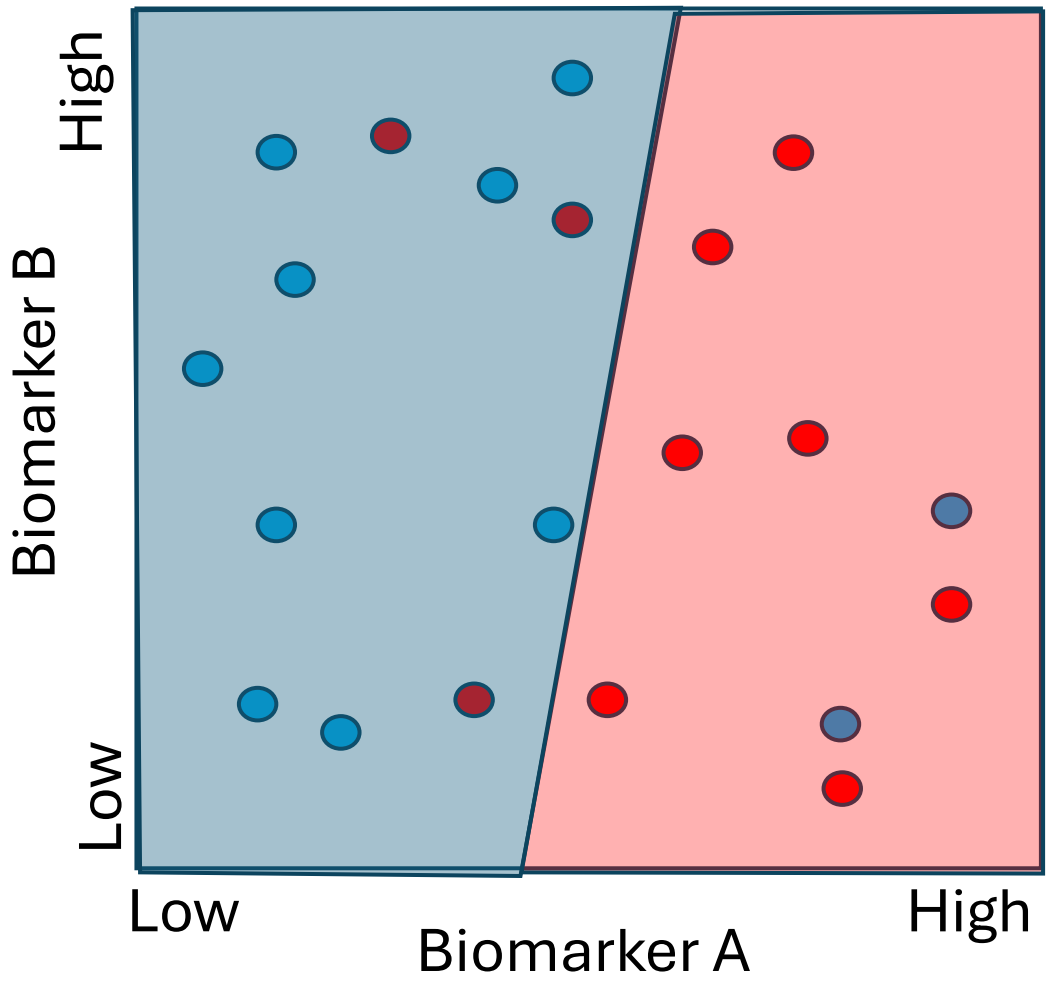




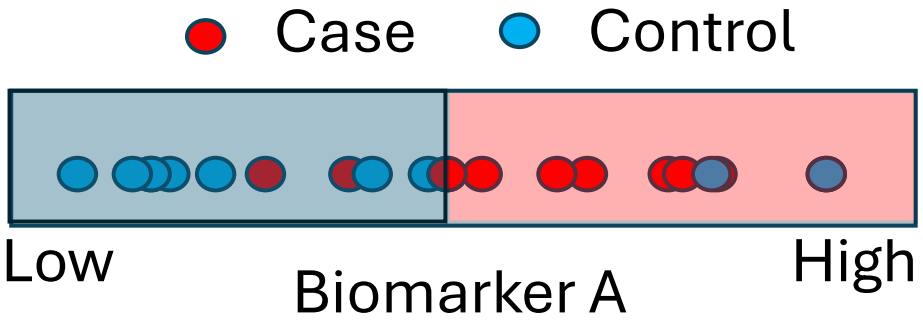
Overview: combination of biomarkers

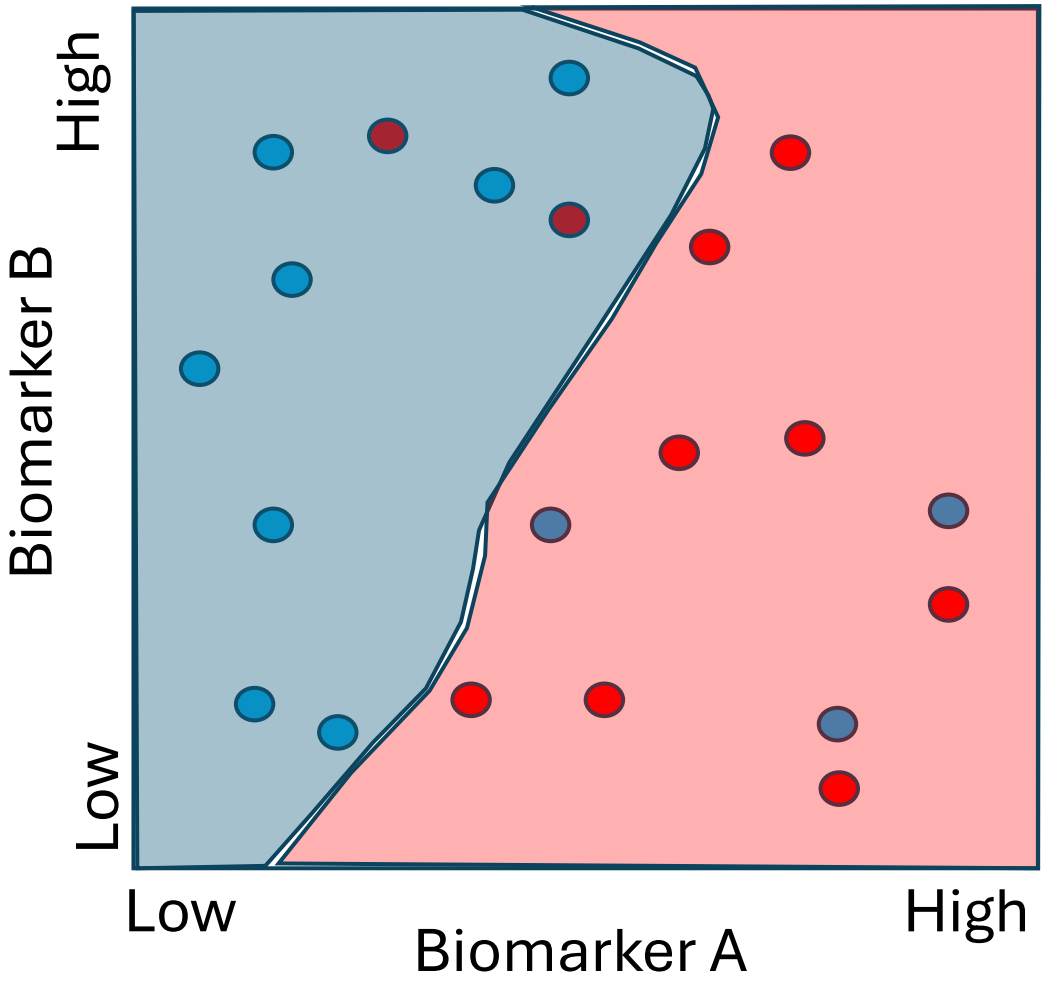


- Linear combination
  - Easy to interpret
  - Best classifier may not be linear



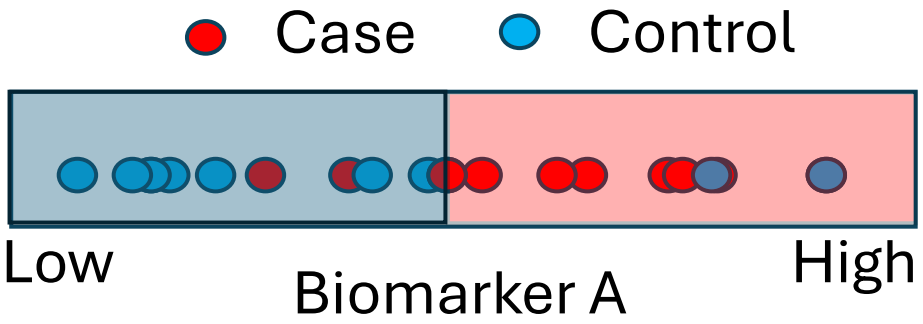
Overview: combination of biomarkers



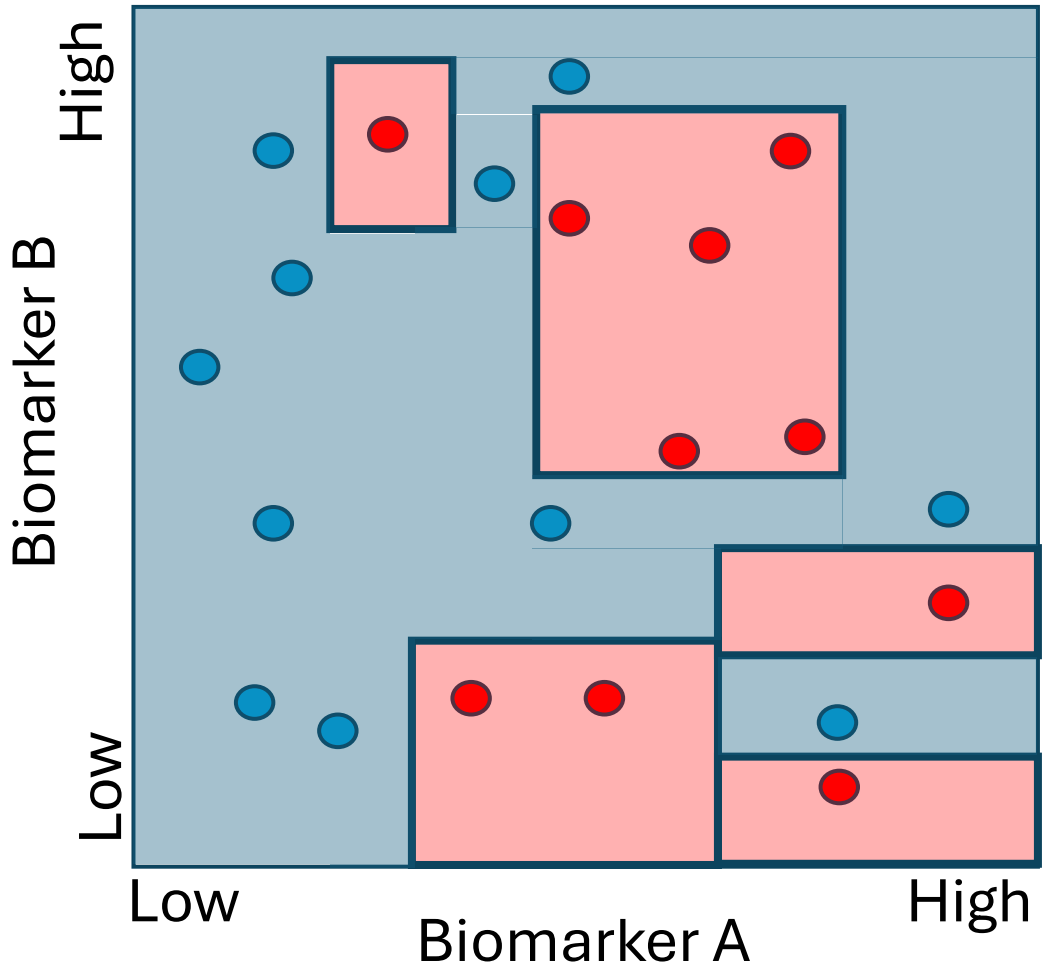


- Nonlinear combination
  - It may improve accuracy
  - Interpretation is often a challenge

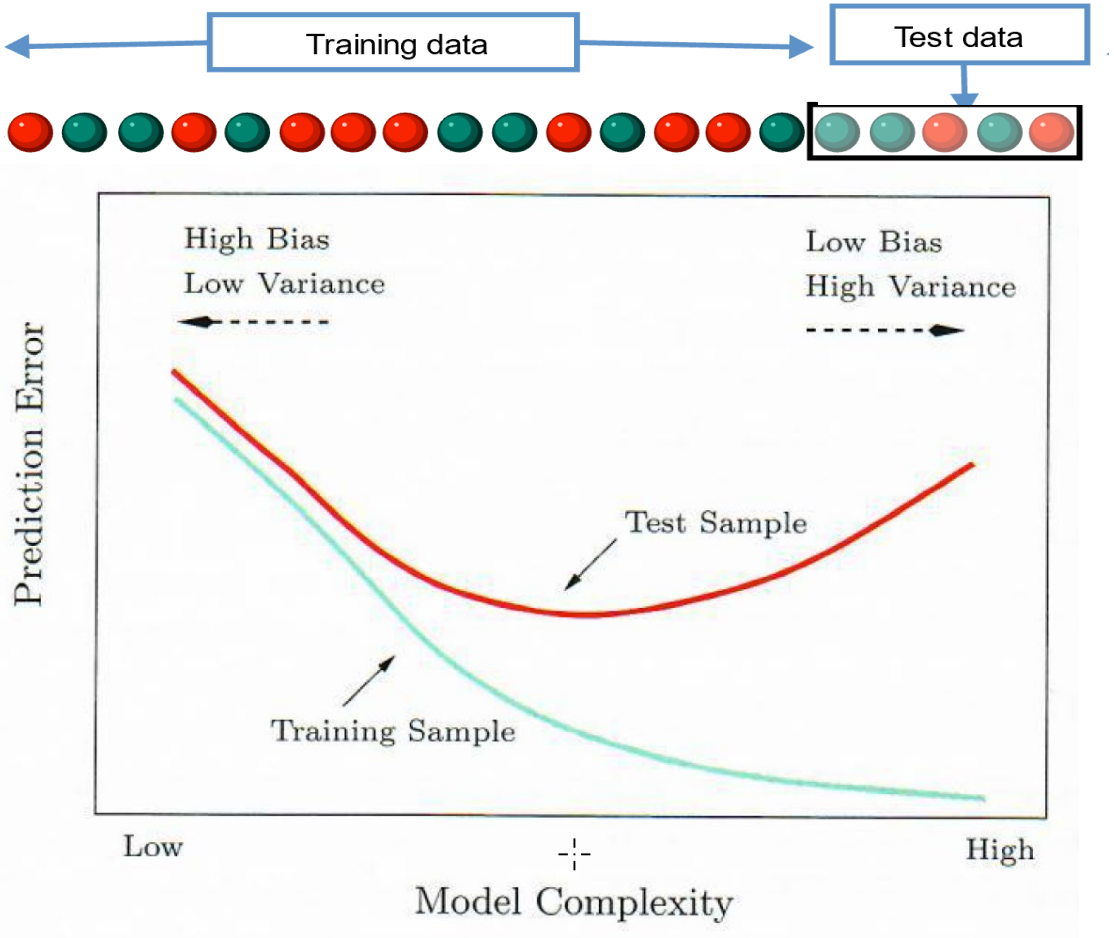
Overview: combination of biomarkers

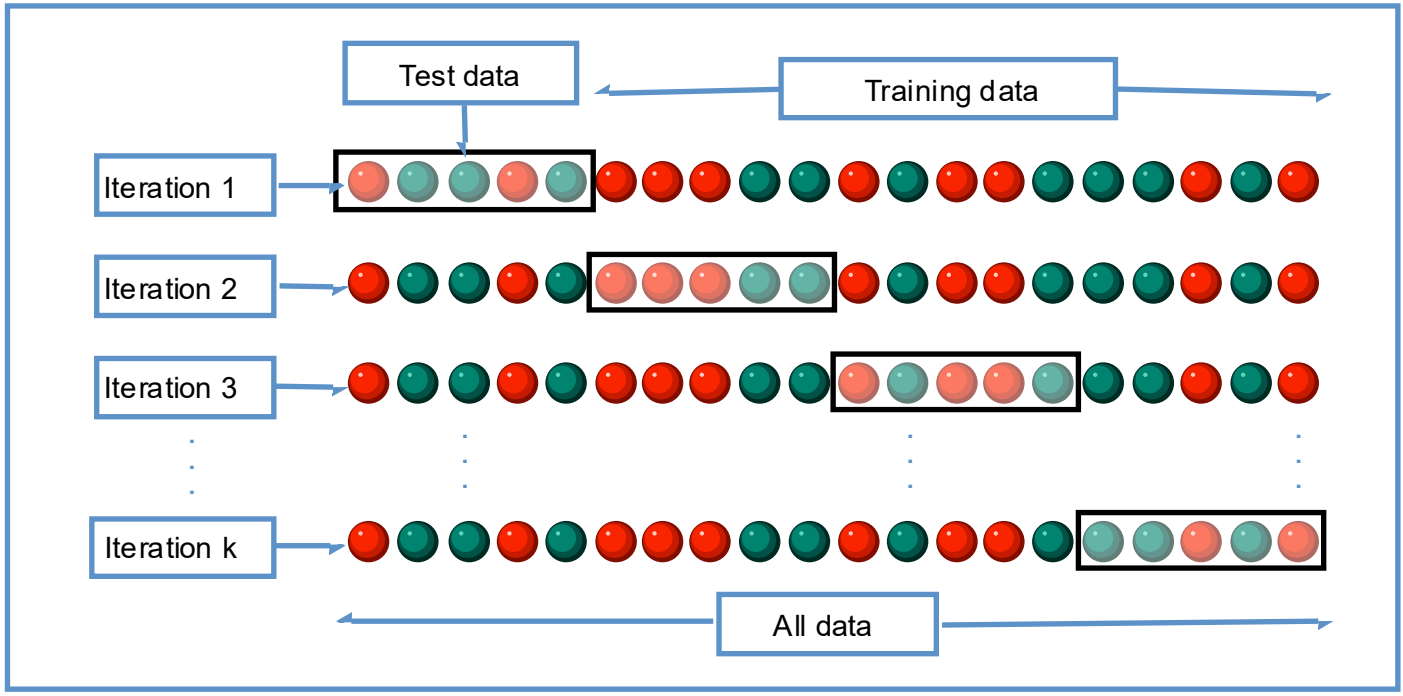


- Increase complexity
  - May not be interpretable.
  - 100% accuracy due to overfitting.



Training and test set / cross-validation





Hastie, T., 2009. The elements of statistical learning: data mining, inference, and prediction.

https://en.wikipedia.org/wiki/Cross-validation_(statistics)

Logistic regression

Preliminary

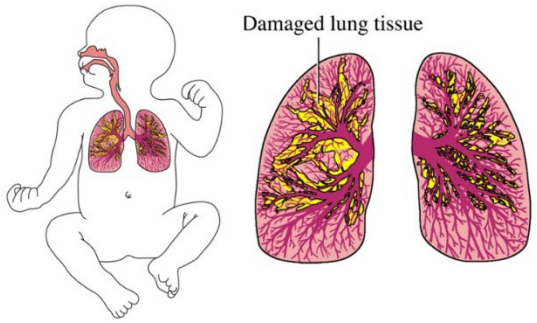
- Y: outcome, target variable
- X: independent variable, explanator, feature variable (or biomarker)

- Y is a binary outcome variable, Y=1 (Event) or 0 (No event)
  - Death or alive
  - Diseased or non-diseased
  - Exposed or unexposed
- X is an independent variable.
- Pr(Event)=P(Y=1) is probability of Y=1, e.g. probability of disease
- Odds of event=Pr(Event)/Pr(No event)
  - Probability is between (0,1).
  - Odds is between (0,∞)

Probability	Odds
0.1	0.11
0.2	0.25
0.3	0.43
0.4	0.67
0.5	1.00
0.6	1.50
0.7	2.33
0.8	4.00
0.9	9.00

Bronchopulmonary dysplasia (BPD)

- BPD is a condition that affects babies born prematurely who have underdeveloped lungs and need to use breathing equipment to help them breathe. Pressure from the oxygen they receive causes lung tissue damage.
- Most newborns who develop BPD are born more than 10 weeks before their due dates, weigh less than 2 pounds at birth, and have breathing problems.



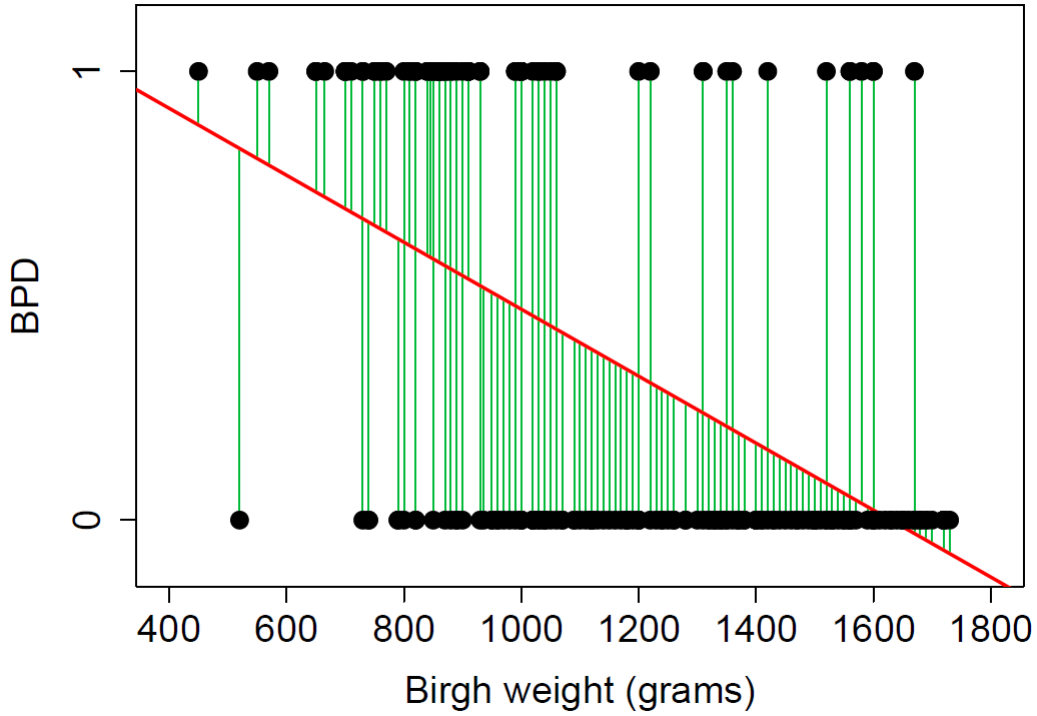
https://www.nhlbi.nih.gov/health/bronchopulmonary-dysplasia

BPD Data

- Population: low birth weight infants whose weights are less than 1,750 grams.
- Sample: 223 infants drawn from the population.
- 76 were diagnosed with BPD (Y=1), and the remaining 147 were not (Y=0).
- Can we use infant's birth weight (X) to estimate the likelihood that he or she will develop BPD (Y)?

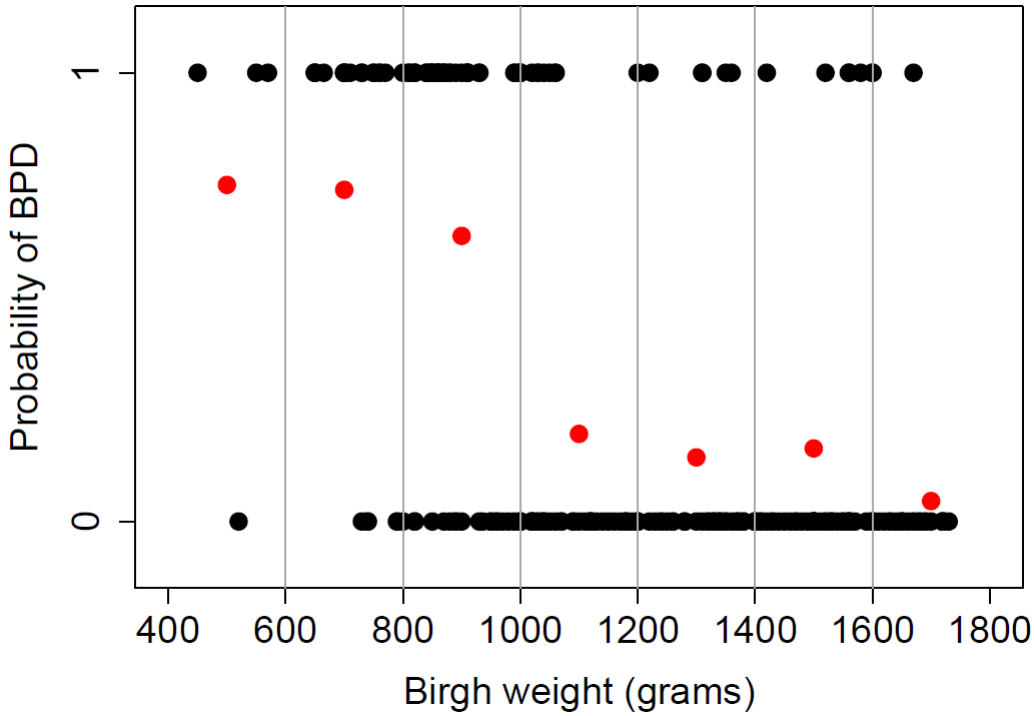
Pagano, M., Gauvreau, K. and Mattie, H., 2022. *Principles of biostatistics*. Chapman and Hall/CRC.

Why not use linear regression for binary data?



- $\hat{Y}$=1.21621-0.00075X
- Is it reasonable?
- How to interpret $\hat{Y}$ =0.9 or -0.1?

Modeling with probability of event



Birth Weight (grams)	Sample size	Number with BPD	Fraction
<600	4	3	0.75
600-800	23	17	0.74
800-1000	55	35	0.64
1000-1200	41	8	0.20
1200-1400	35	5	0.14
1400-1600	43	7	0.16
>1600	22	1	0.05
Overall	223	76	0.34

Modeling with probability of event

- Let p be the probability that Y equals 1, i.e. Pr(Y=1).
- Assume p depend on the value of X=x, denoted as p(x)=Pr(Y=1|X=x).
- A linear model:
  - $p(x) = \beta_0 + \beta_1 x$ is not appropriate because p(x) is bounded by 0 and 1, but $\beta_0 + \beta_1 x$ is unbounded.
- We might consider a nonlinear model.

Logistic Regression

- The logistic regression model:

  $$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x,$$

  where:

  - **p(x)/(1-p(x))**, called the odds, removes the upper bound.
  - **Log**, the natural logarithm, removes the lower bound.
  - $\log\left(\frac{p(x)}{1-p(x)}\right)$ is called a logit.
  - β_0 is an intercept, i.e. the logit value when X=0
  - β_1 is a slope, i.e. the change of logit for a one unit increase in X

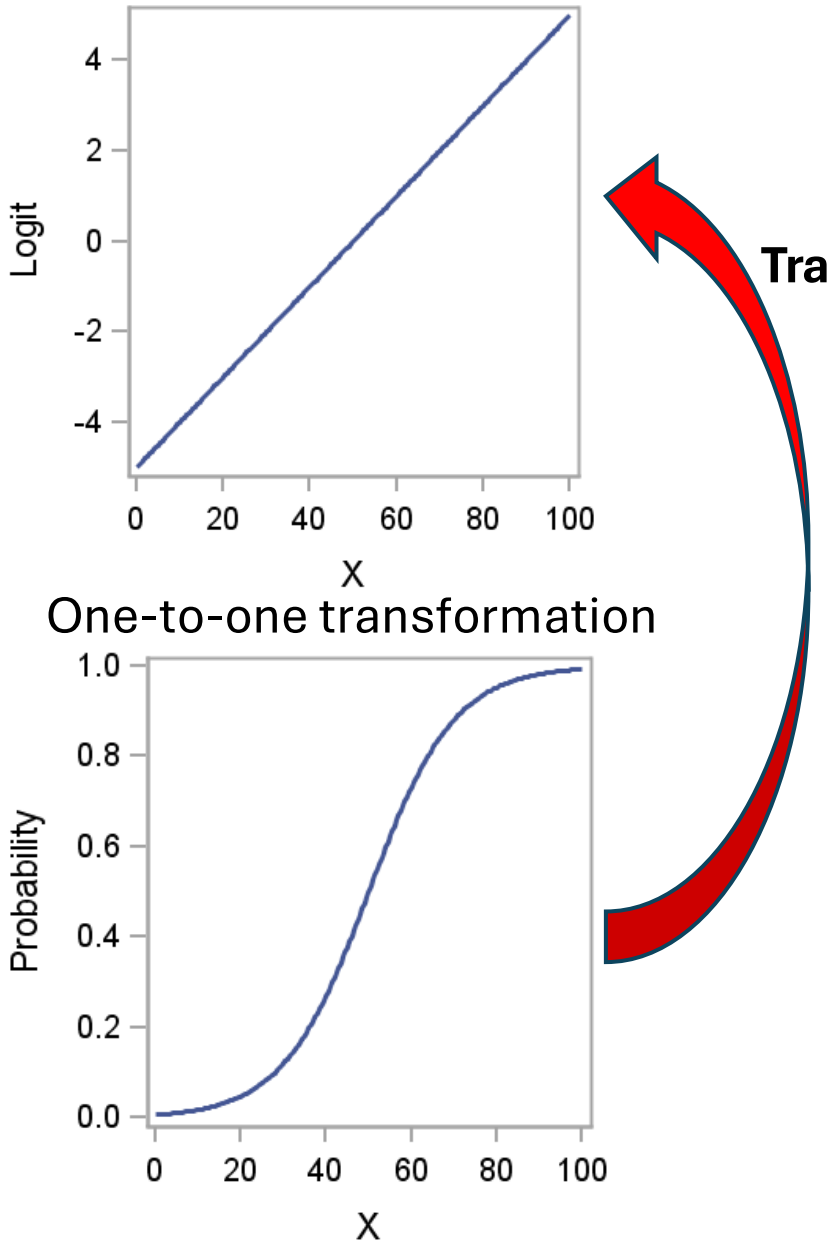
Logistic Regression

- The logistic regression model is equivalent to

$$p(x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}}$$

- Why?
  - $\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 X}$
  - $p(x) = e^{\beta_0+\beta_1 X}(1 - p(x))$
  - $\left(1 + e^{\beta_0+\beta_1 X}\right)p(x) = e^{\beta_0+\beta_1 X}$
  - $p(x) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$
  - $p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$

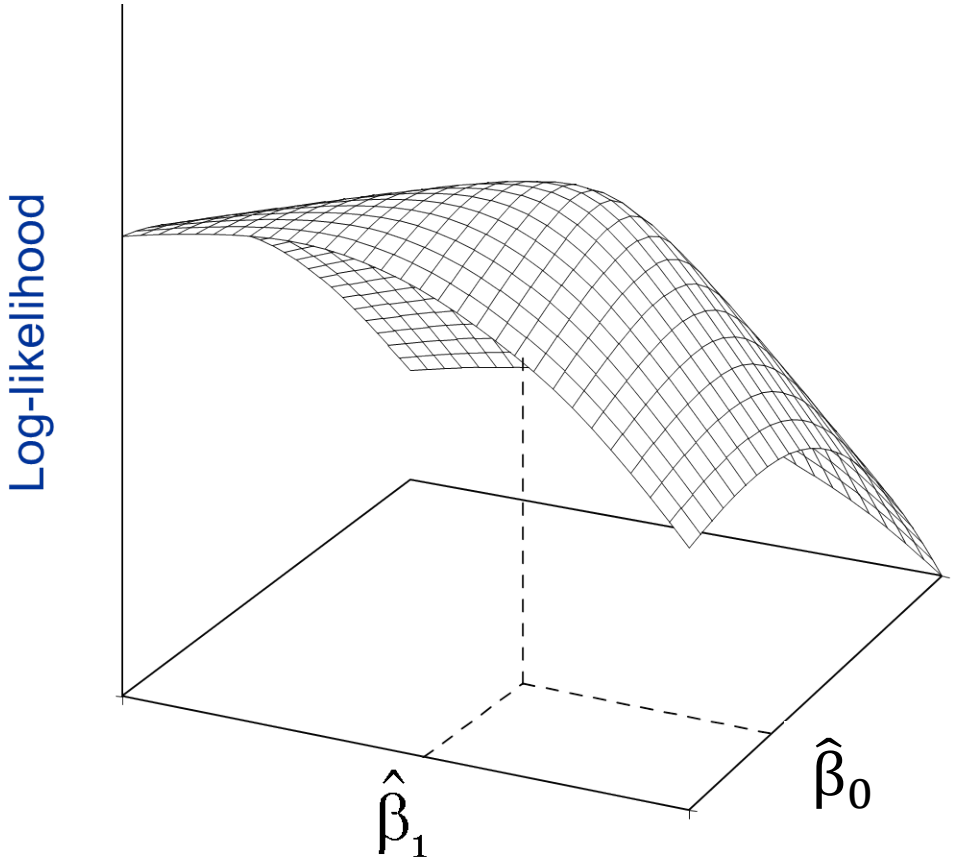
Back to probability (or inverse logit transformation)



Logit Transformation

How to estimate β_0 and β_1?

- Define an objective function, such as loss function, likelihood, AUC.
- Find $\hat{\beta}_0$ and $\hat{\beta}_1$ that maximize (or minimize) the objective function.

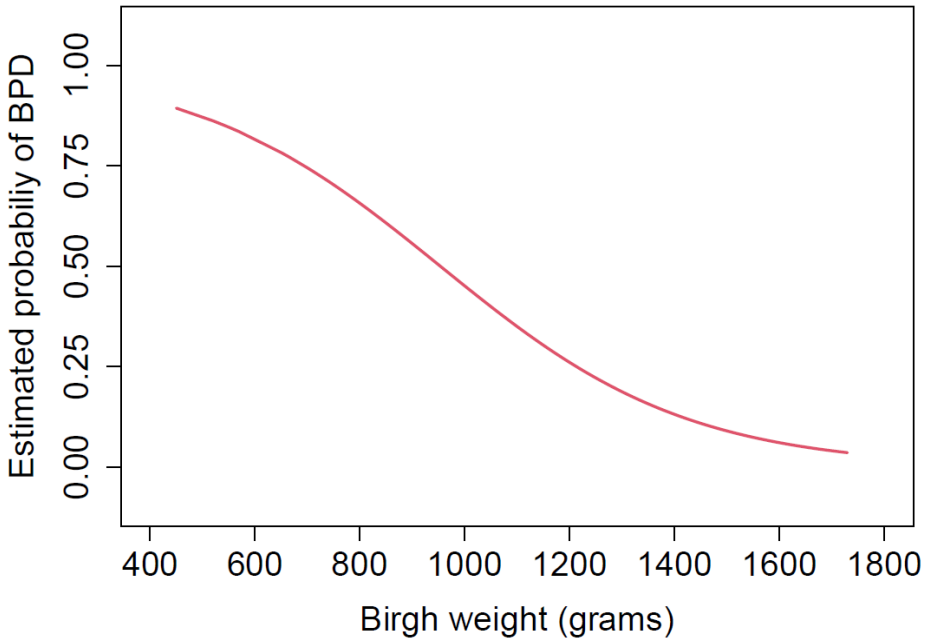


Evaluation: BPD data

- Based on the logistic regression with an independent variable of birth weights:

$$\log\left(\frac{\hat{p}(x)}{1-\hat{p}(x)}\right) = 0.40342 - 0.0042x$$

- $\hat{p}(x) = \frac{1}{1+e^{-(4.0343-0.0042x)}}$



- Fit a logistic regression based on the training set (y_training,x_training)
- Produce a test ROC curve based on (y_test, $\hat{p}(x\_test)$).

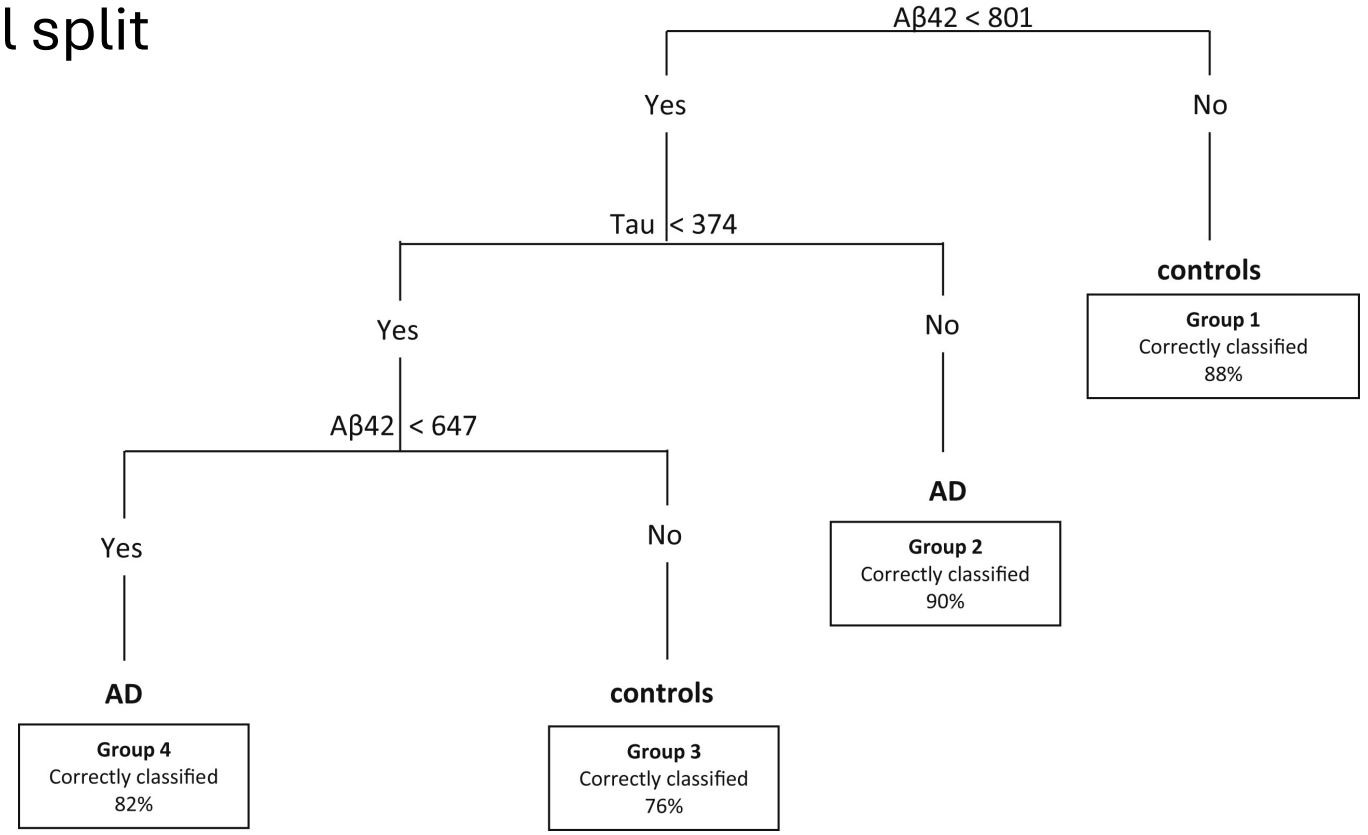
Multiple Logistic Regression

- $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q$
  - $X_1, X_2, \ldots, X_q$ are q biomarkers.
- Regularization methods:
  - RIDGE: L2 penalty → selected biomarkers are relatively large.
  - LASSO (least absolute shrinkage and selection operator): L1 penalty → selected biomarkers are relatively small.
  - Elastic net: L1 & L2 linearly combined penalty → selected biomarkers are usually between the two models.
- Some of βs (often many βs) are estimated to zeros. Easy to interpret.

Random Forest

Classification and regression tree (CART)

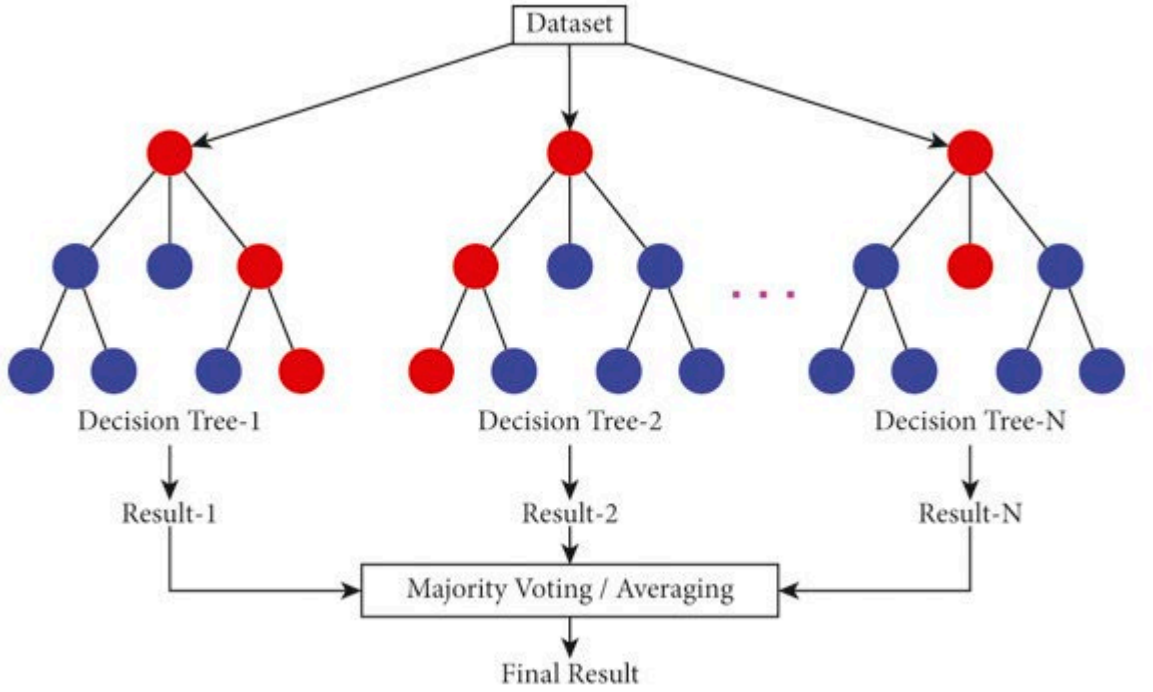
- CSF biomarkers in predicting Alzheimer's disease
- Recursively search the optimal split for each terminal node.
- Prune some of the nodes.
- Nonlinear but interpretable.
- May not work well for high-dimensional data.



Mofrad, R.B., Schoonenboom, N.S., Tijms, B.M., Scheltens, P., Visser, P.J., van der Flier, W.M. and Teunissen, C.E., 2019. Decision tree supports the interpretation of CSF biomarkers in Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *11*, pp.1-9.

Random Forest

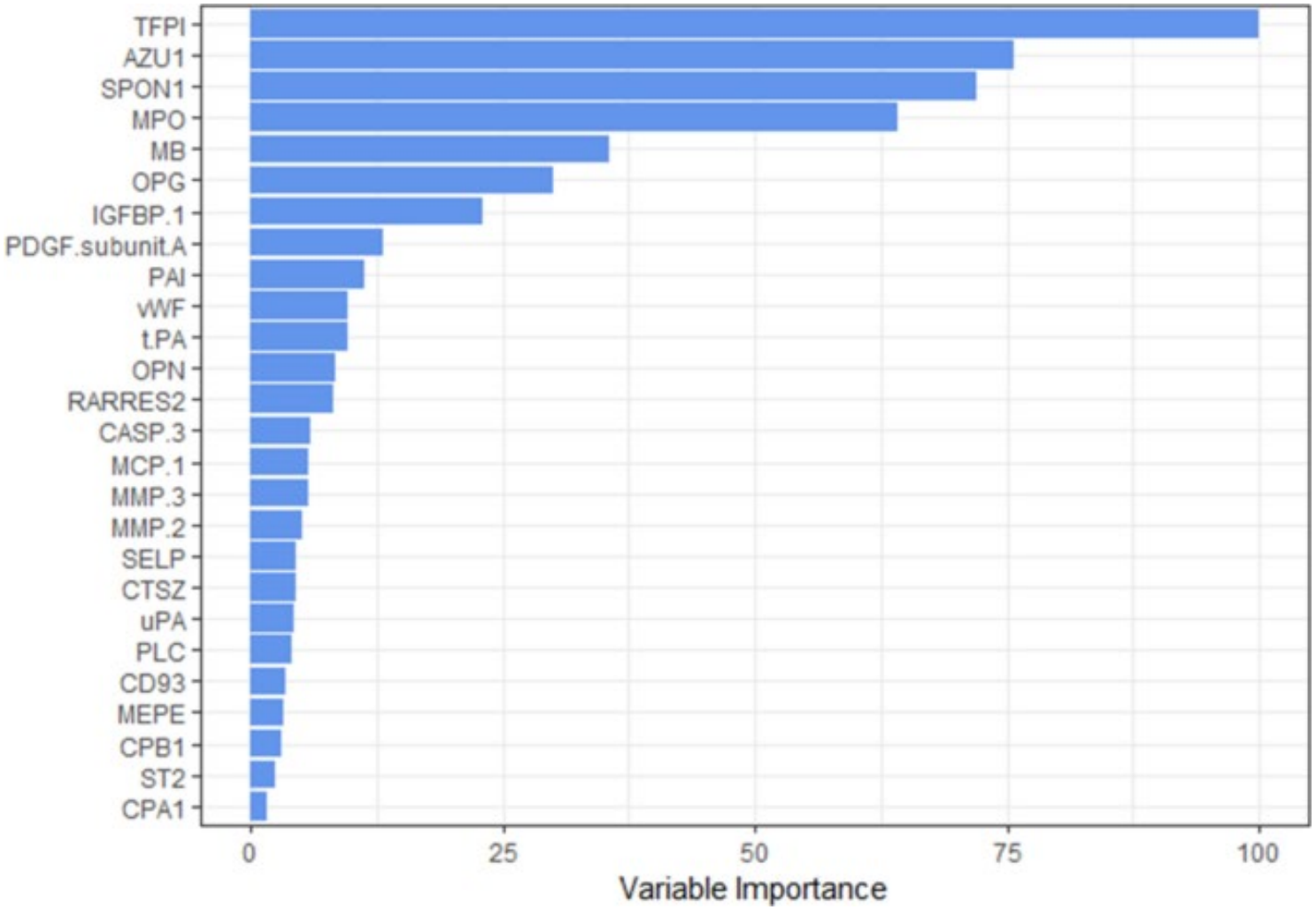
- A collection of decision trees, based on bootstrap resamples.



Khan, M.Y., Qayoom, A., Nizami, M.S., Siddiqui, M.S., Wasi, S. and Raazi, S.M.K.U.R., 2021. Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques. *Complexity*, *2021*(1), p.2553199.

- RF is usually better than CART but less interpretable.

Random Forest



Maag, E., Kulasingam, A., Grove, E.L., Pedersen, K.S., Kristensen, S.D. and Hvas, A.M., 2021. Statistical and machine learning methods for analysis of multiplex protein data from a novel proximity extension assay in patients with ST-elevation myocardial infarction. *Scientific Reports*, *11*(1), p.13787.

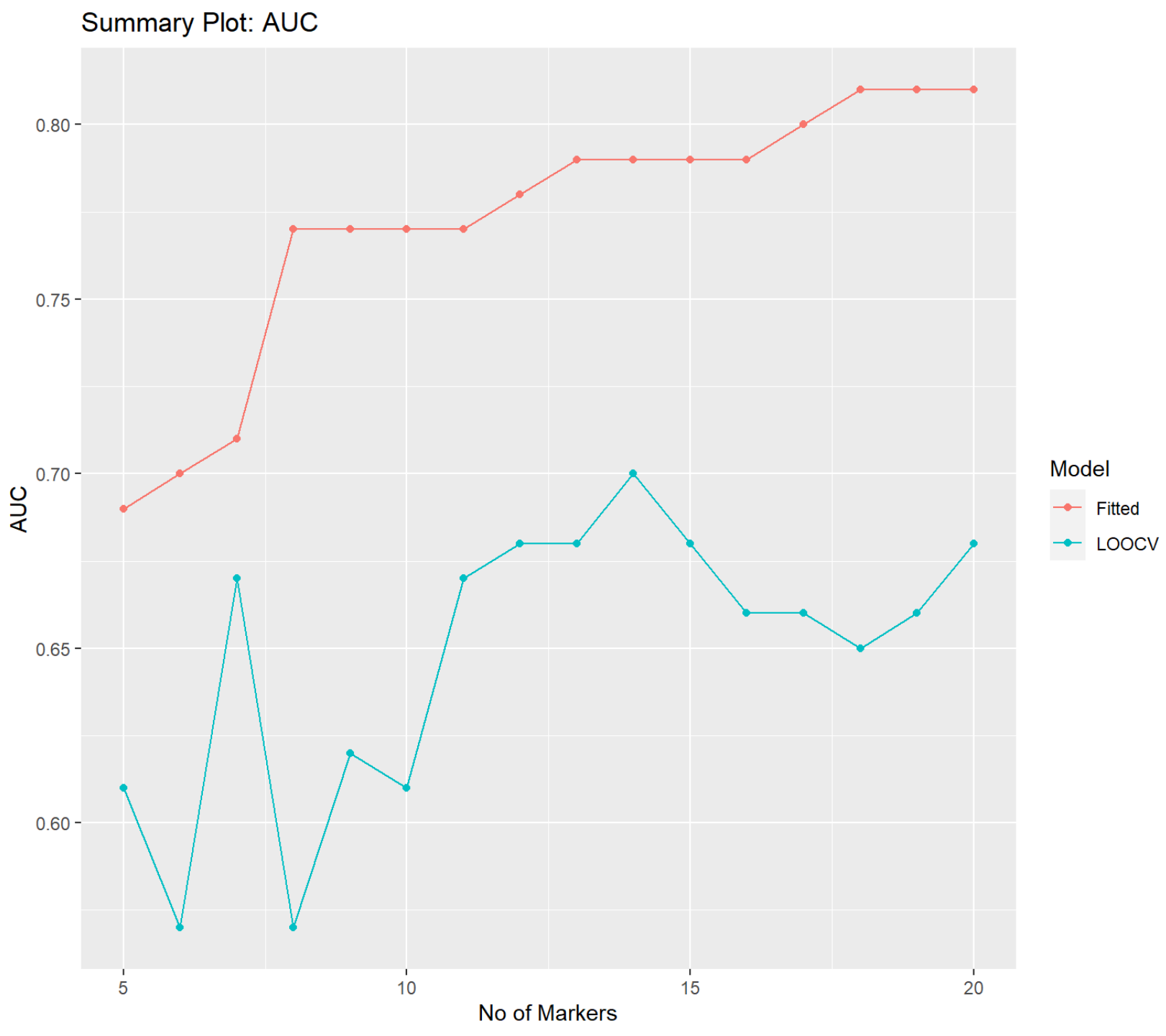
- There are many other ML models:
  - SVM (support vector machine),
  - Neural network,
  - Naïve Bayes,
  - Bayesian CART,
  - XgBoost (eXtreme Gradient Boosting),
  - Etc.
- Some of these models are not interpretable / not appropriate for low-dimensional data.

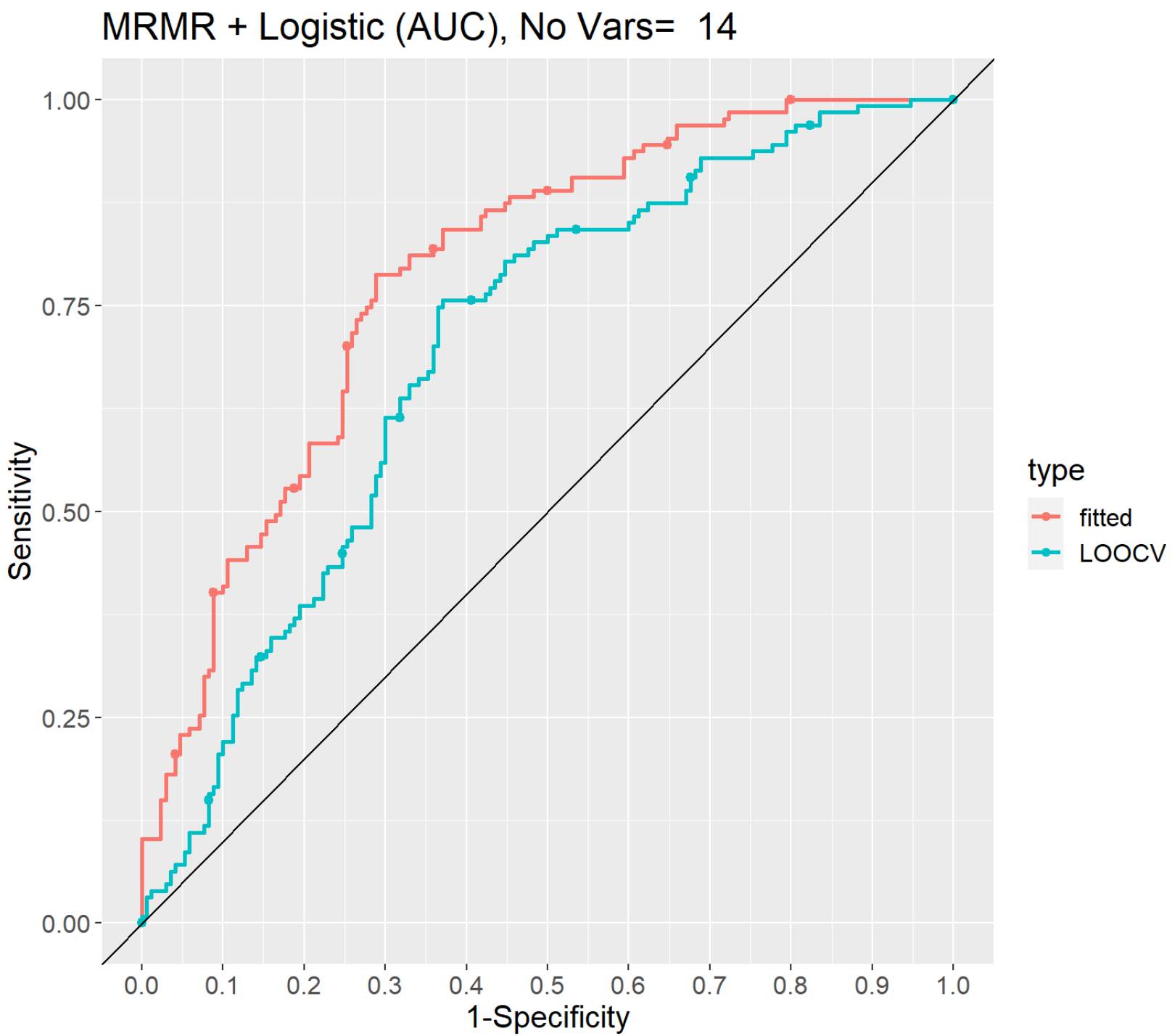
Two-stage ML models

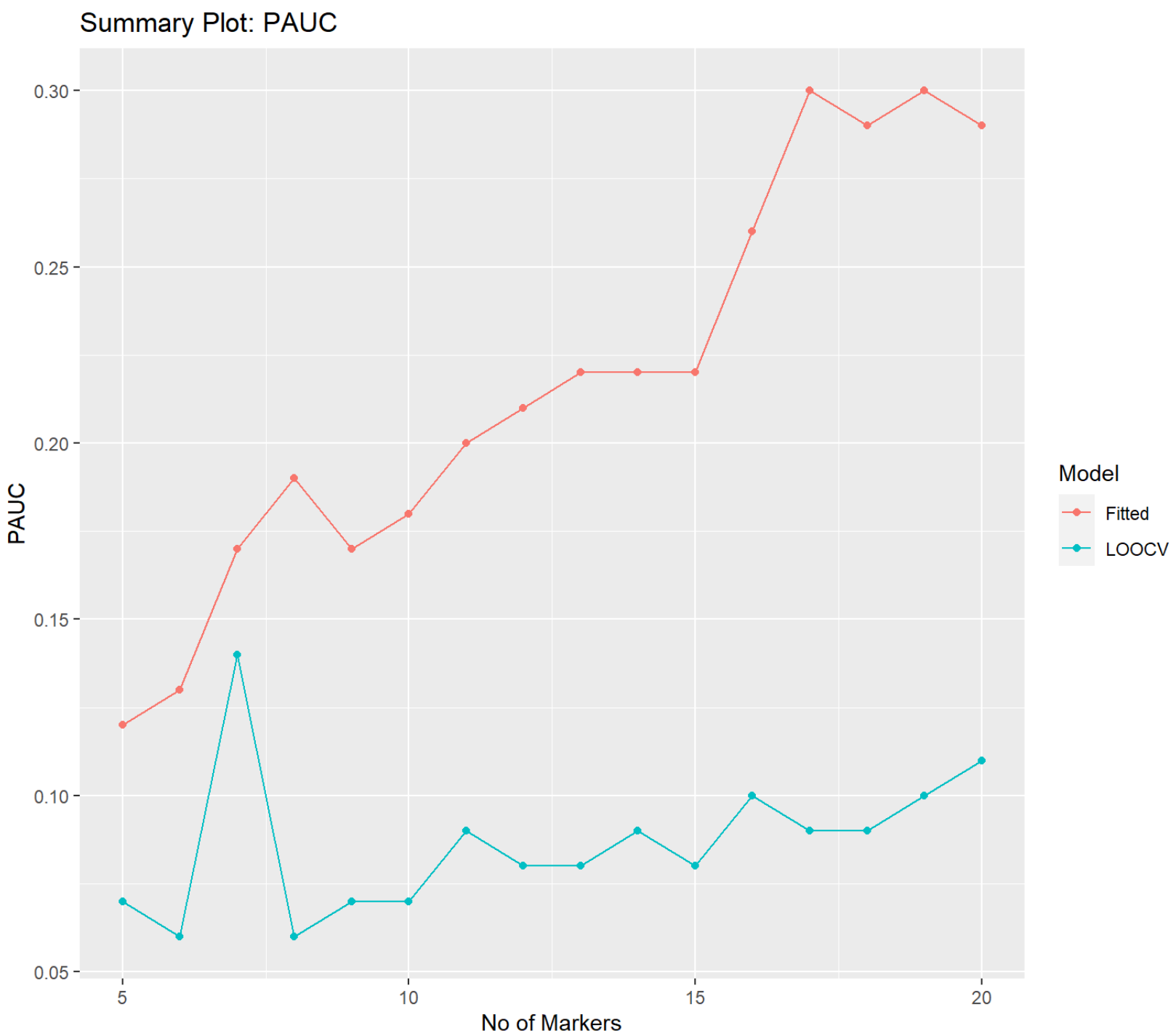
Two-stage ML models

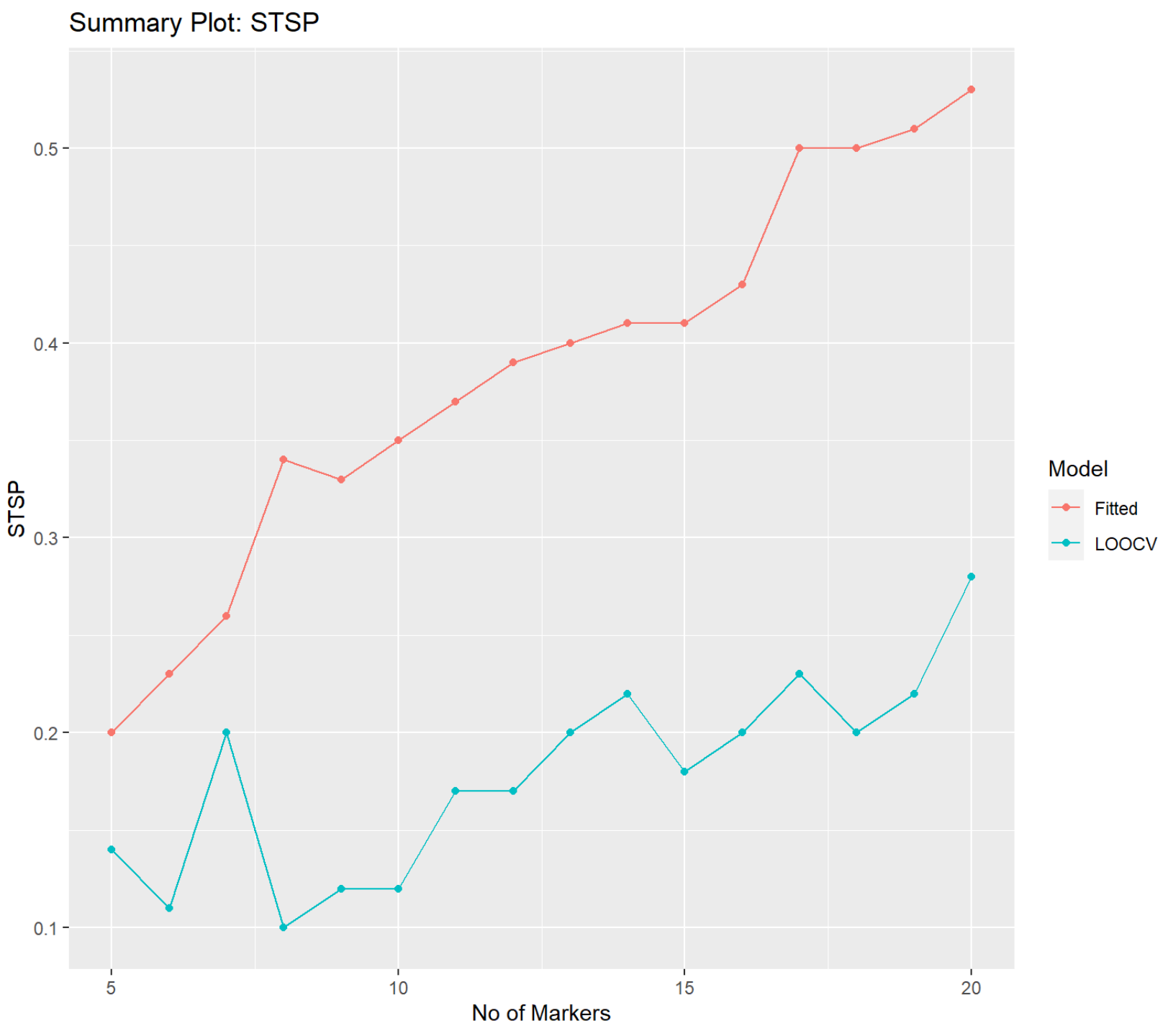
- We often prefer parsimonious and interpretable models.

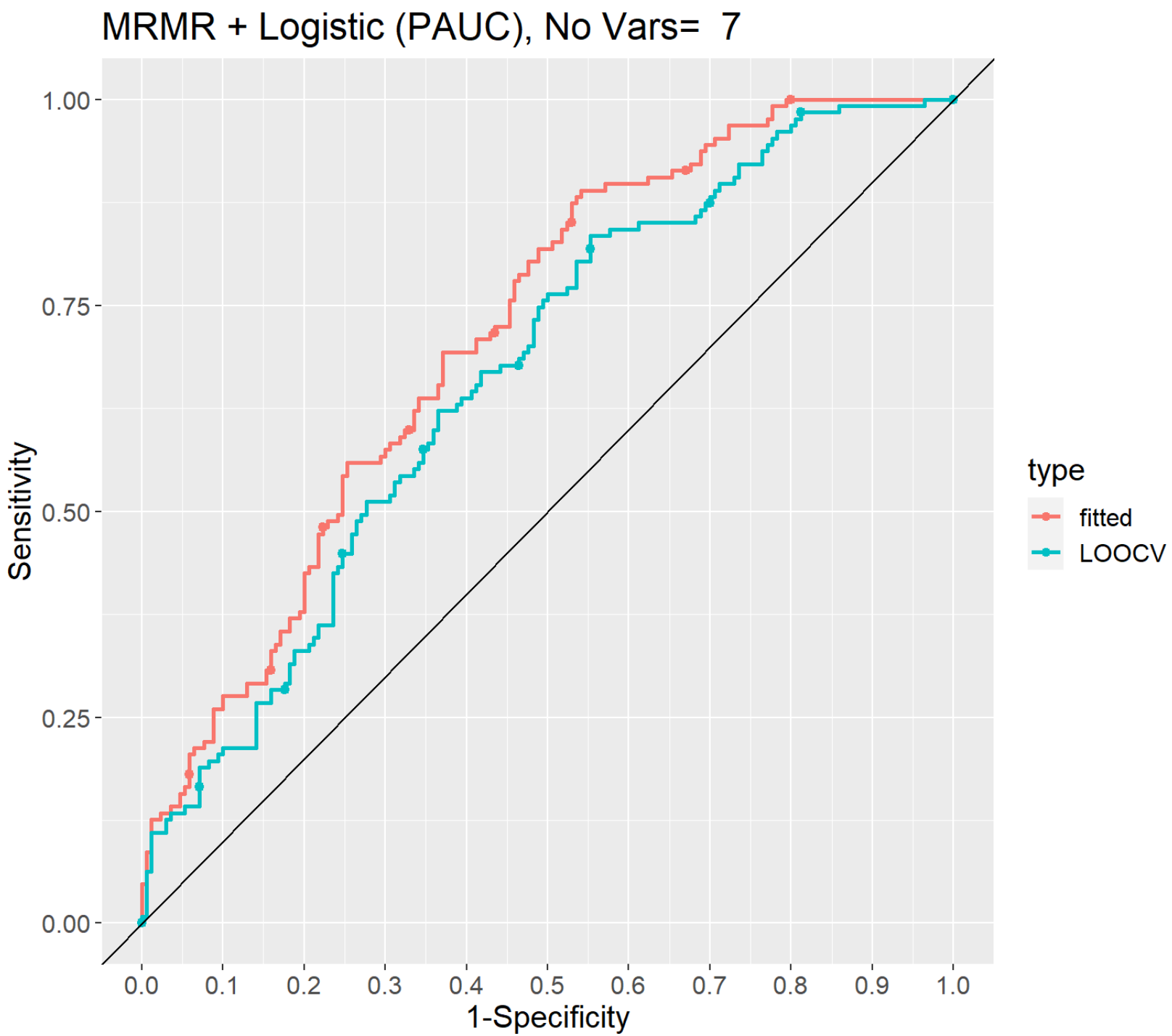
- Stage 1: Select potentially useful biomarkers.
  - The Minimum Redundancy Maximum Relevance (MRMR) algorithm can be used to select K biomarkers that are effective and less correlated.
  - K is a hyperparameter.
- Stage 2: Use ML models with the selected biomarkers.

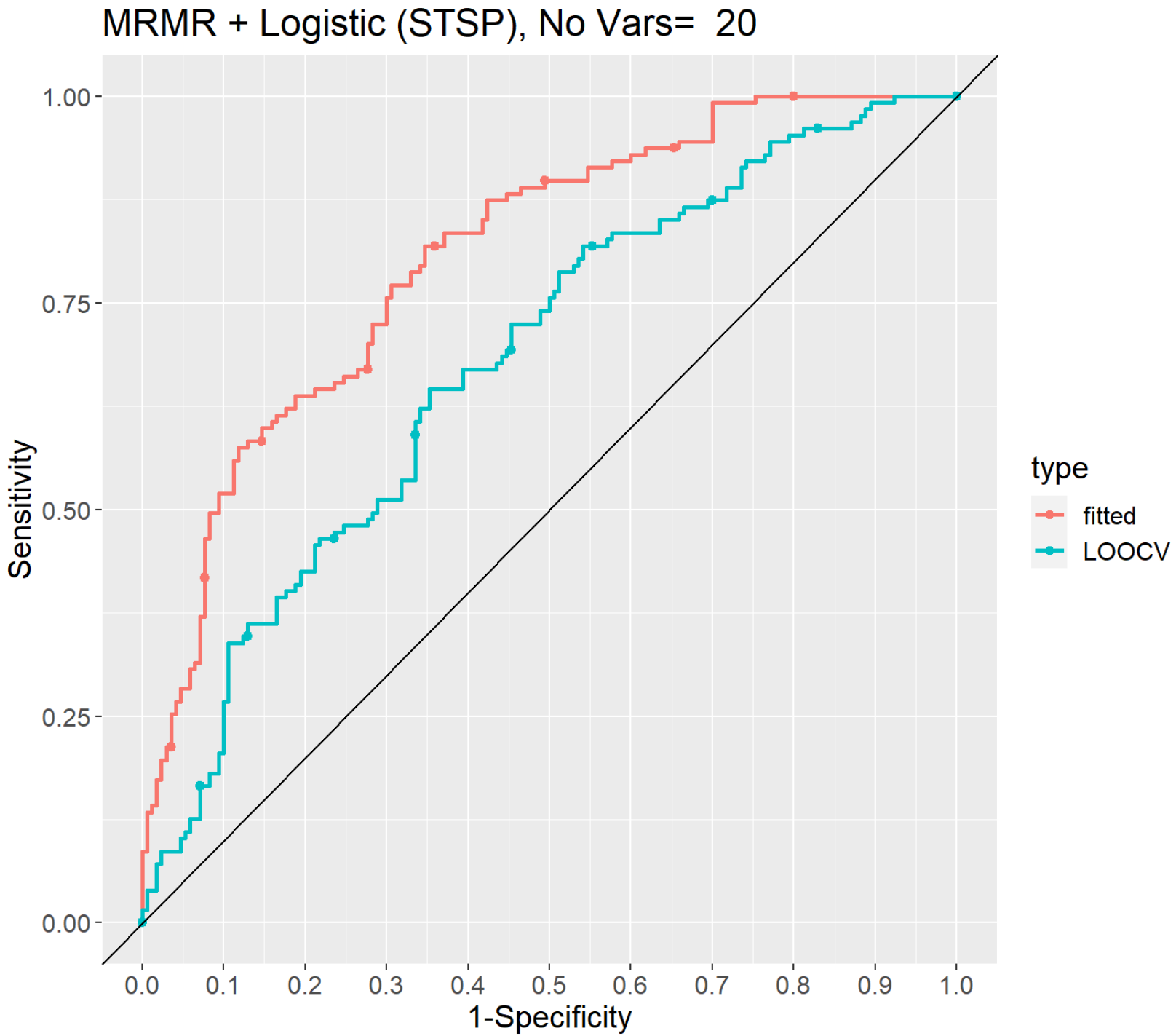

Summary Plot: AUC
AUC
0.80
0.75
0.70
0.65
0.60
5
10
15
20
No of Markers
Model
Fitted
LOOCV


MRMR + Logistic (AUC), No Vars= 14
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1-Specificity
type
fitted
LOOCV


Summary Plot: PAUC
PAUC
0.30
0.25
0.20
0.15
0.10
0.05
5
10
15
20
No of Markers
Model
Fitted
LOOCV


Summary Plot: STSP
STSP
0.5
0.4
0.3
0.2
0.1
5
10
15
20
No of Markers
Model
Fitted
LOOCV


MRMR + Logistic (PAUC), No Vars= 7
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1-Specificity
type
fitted
LOOCV


MRMR + Logistic (STSP), No Vars= 20
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1-Specificity
type
fitted
LOOCV

Evaluation of two-stage ML models.

- The below approach causes overfitting problems.
  - Stage 1: select K biomarkers using the entire dataset.
  - Stage 2: train ML models on the training set & evaluate their performances using the test set.

- Instead, we recommend the below approach.
  - Select K biomarkers & train ML models on the training set.
  - Evaluate these two-stage ML models using the test set.

Summary

- Any other ML models can be used to find the best combination of biomarkers.
  - Are black-box models acceptable? E.g., 1000 of biomarker combinations.
- Evaluation of two-stage ML models.
  - We do not recommend the following approach. Stage 1: select K biomarkers using the entire dataset. Stage 2: train ML models on the training set & evaluate their performances using the test set.
  - Instead, we recommend the below. Selecting K biomarkers & train ML models on the training set. Evaluate these two-stage ML models using the test set.
- Unless the dataset is extremely large, leave-one-out cross-validation (loocv) can be used instead of 5-fold cross-validation.