

# Stability of clinical prediction models developed using statistical or machine learning methods

**Richard D. Riley**<sup>1</sup>  | **Gary S. Collins**<sup>2</sup> 

<sup>1</sup>Institute of Applied Health Research,  
College of Medical and Dental Sciences,  
University of Birmingham, Birmingham,  
UK

<sup>2</sup>Centre for Statistics in Medicine,  
Nuffield Department of Orthopaedics,  
Rheumatology and Musculoskeletal  
Sciences, University of Oxford, Oxford,  
UK

## Correspondence

Richard D. Riley, Institute of Applied  
Health Research, College of Medical and  
Dental Sciences, University of  
Birmingham, Birmingham B15 2TT, UK.  
Email: [r.d.riley@bham.ac.uk](mailto:r.d.riley@bham.ac.uk)

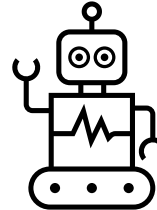
**Nat Sirirutbunkajorn**

Student, Data Science for Healthcare and Clinical Informatics

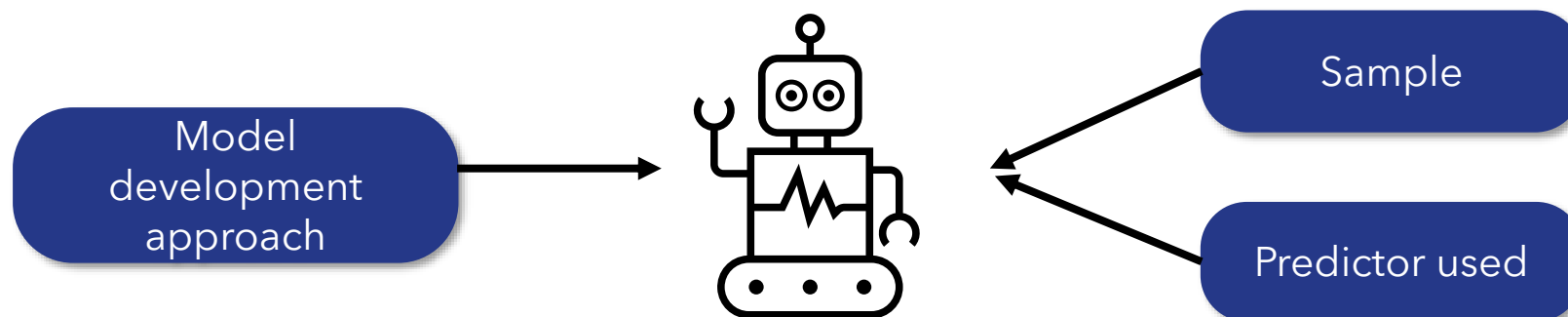
# Introduction

## Introduction

- Clinical prediction models are used in healthcare to:
  - predict an outcome value
  - estimate risk of an outcome (now or in the future)
- The major methods/approach used to develop prediction models are:
  - Statistical method
  - Machine learning method
- Predictors use for such models might includes:
  - Basic characteristic (age, sex, etc...)
  - Measurements (blood pressure, biomarkers, test results)
  - Imaging



## Introduction



## Introduction

Reliability ↑

- Relevant predictors
- Large sample size
- No adjustment to reduce overfitting

Reliability ↓

- Small sample size relative to number of predictors
- Complex model relative to sample size
- No adjustment to reduce overfitting

## Introduction

Reliability ↓

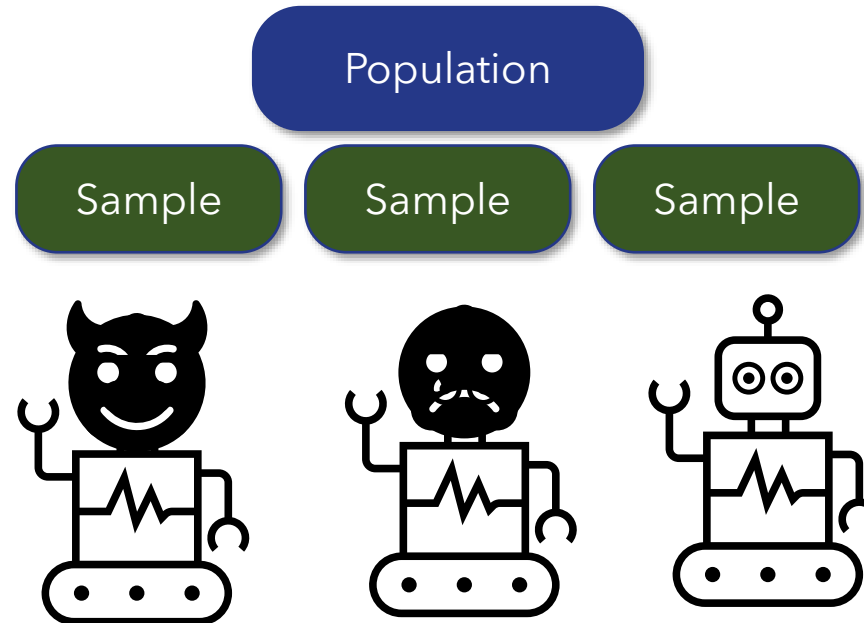
- Small sample size relative to number of predictors
- Complex model relative to sample size
- No adjustment to reduce overfitting



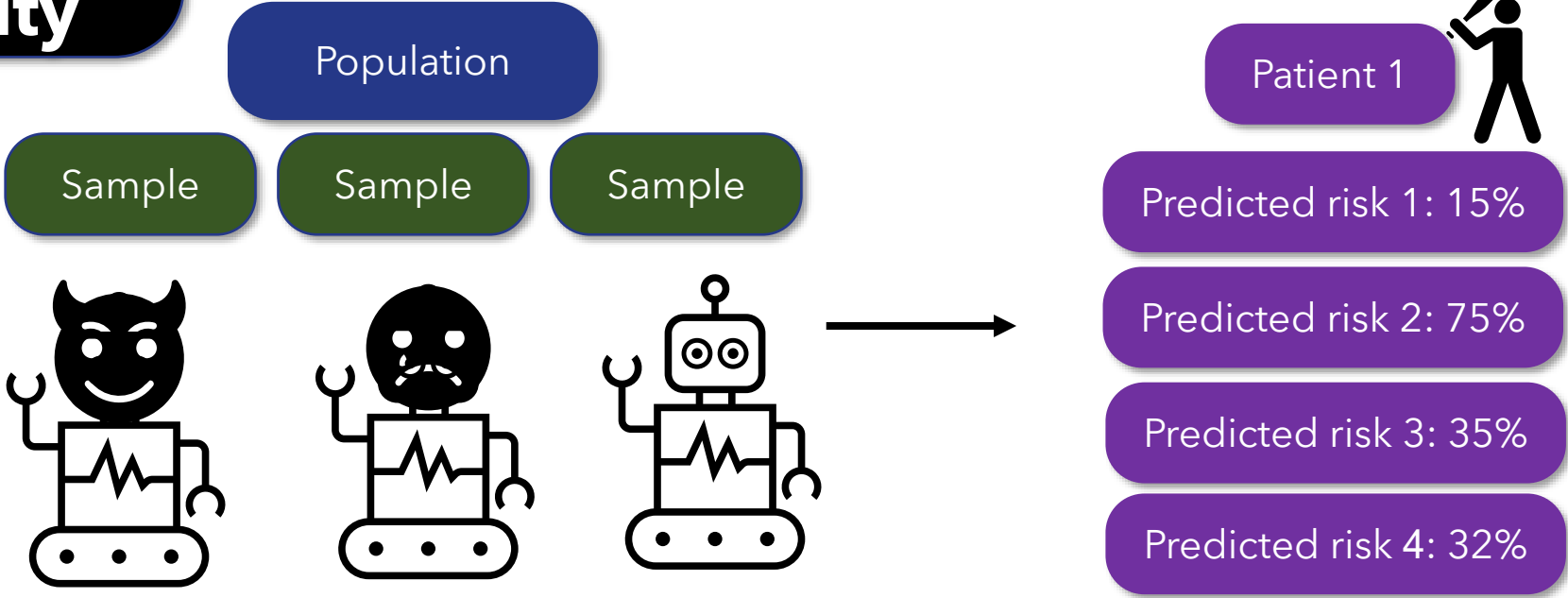
# Instability Volatility

Different:

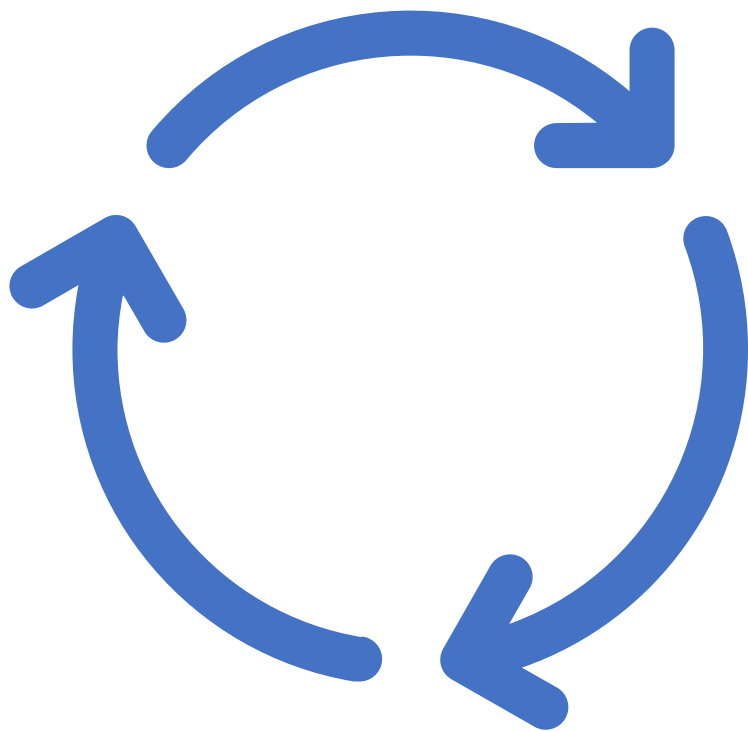
- Weights
- Predictors selected
- Parameters
- etc...

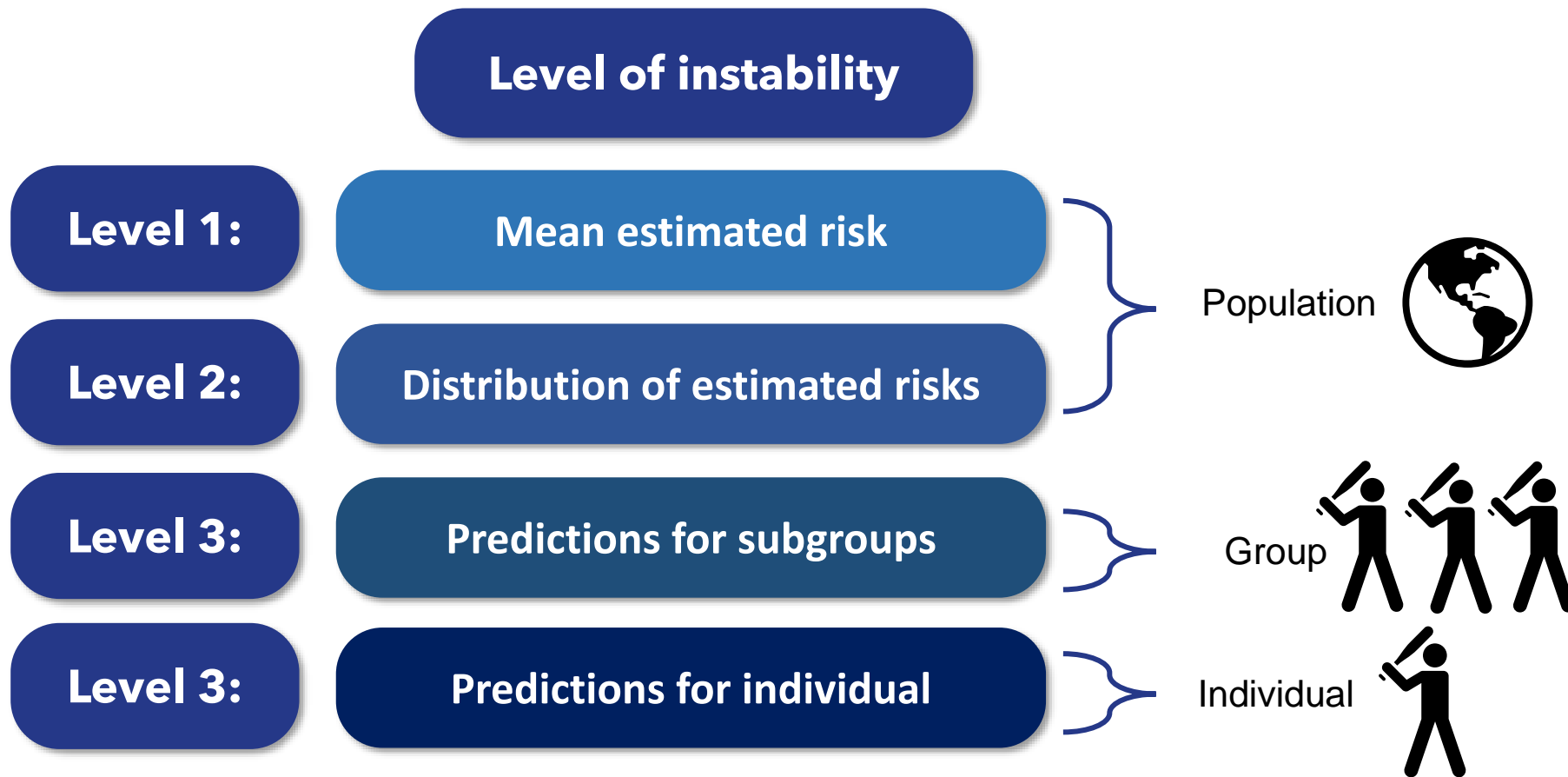


# Instability Volatility









# Quantifying Stability

# QUANTIFYING INSTABILITY IN MODEL DEVELOPMENT STUDIES

## Bootstrap assessment of instability

- In practice when developing a model researchers will not know the “true” risk of each individual, and so need to examine instability using **bootstrapping**.
- Some parts of the model-building process may not be easily implemented (automatically) in each bootstrap and in these instances, some compromise may be required to make the approach practical.

### BOX 1 The bootstrap process to examine instability of predictions from a clinical prediction model after model development

Context: a prediction model has just been developed using a particular model-building strategy, and the model developers want to examine the potential instability of predictions from this model. To do this using the model development dataset of  $N$  participants, we recommend the following bootstrap process is applied:

Step 1: Use the developed model to make predictions ( $\hat{p}_i$ ) for each individual participant ( $i = 1$  to  $N$ ) in the development dataset.

Step 2: Generate a bootstrap sample with replacement, ensuring the same size ( $N$ ) as the model development dataset.

Step 3: Develop a bootstrap prediction model in the bootstrap sample, replicating exactly (or as far as practically possible) the same model-building strategy as used originally.

Step 4: Use the bootstrap model developed in step 3 to make predictions for each individual ( $j$ ) in the original dataset. We refer to these predictions as  $\hat{p}_{bi}$ , where  $b$  indicates which bootstrap sample the model was generated in ( $b = 1$  to  $B$ ).

Step 5: Repeat steps 2–4 a total of  $(B - 1)$  times, and we suggest  $B$  is at least 200.

Step 6: Store all the predictions from the  $B$  iterations of steps 2–5 in a single dataset, containing for each individual a prediction ( $\hat{p}_i$ ) from the original model and  $B$  predictions ( $\hat{p}_{1i}, \hat{p}_{2i}, \dots, \hat{p}_{Bi}$ ) from the bootstrap models.

Step 7: Summarize the instability in model predictions, including prediction instability plots, calibration instability plots, and the mean absolute predictor error (MAPE), see main text.

## QUANTIFYING INSTABILITY IN MODEL DEVELOPMENT STUDIES

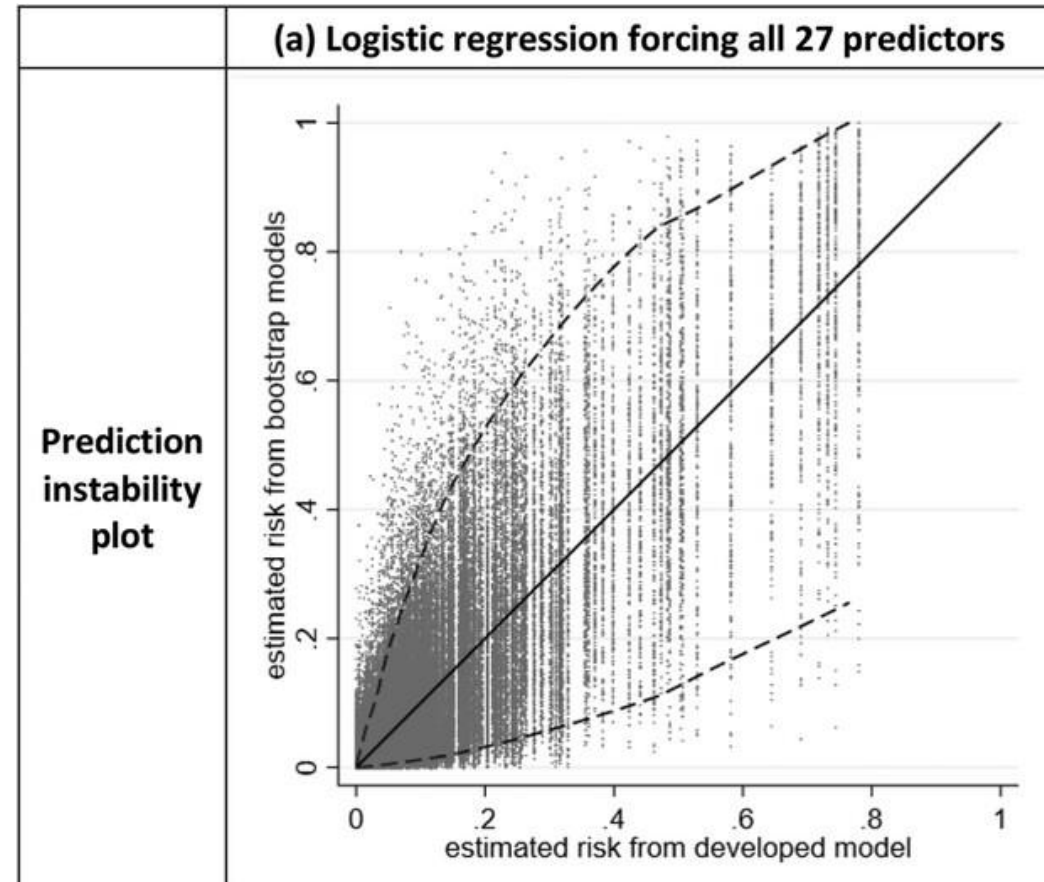
### Numerical summaries and graphical presentations of instability

- Prediction from multiple boot strap sample can be used for instability plots and measures.
  - **Prediction instability plot**
  - **Calibration instability plot**
  - **Mean absolute predictor error (MAPE)**
  - **MAPE instability plot**

# QUANTIFYING INSTABILITY IN MODEL DEVELOPMENT STUDIES

## Prediction instability plot

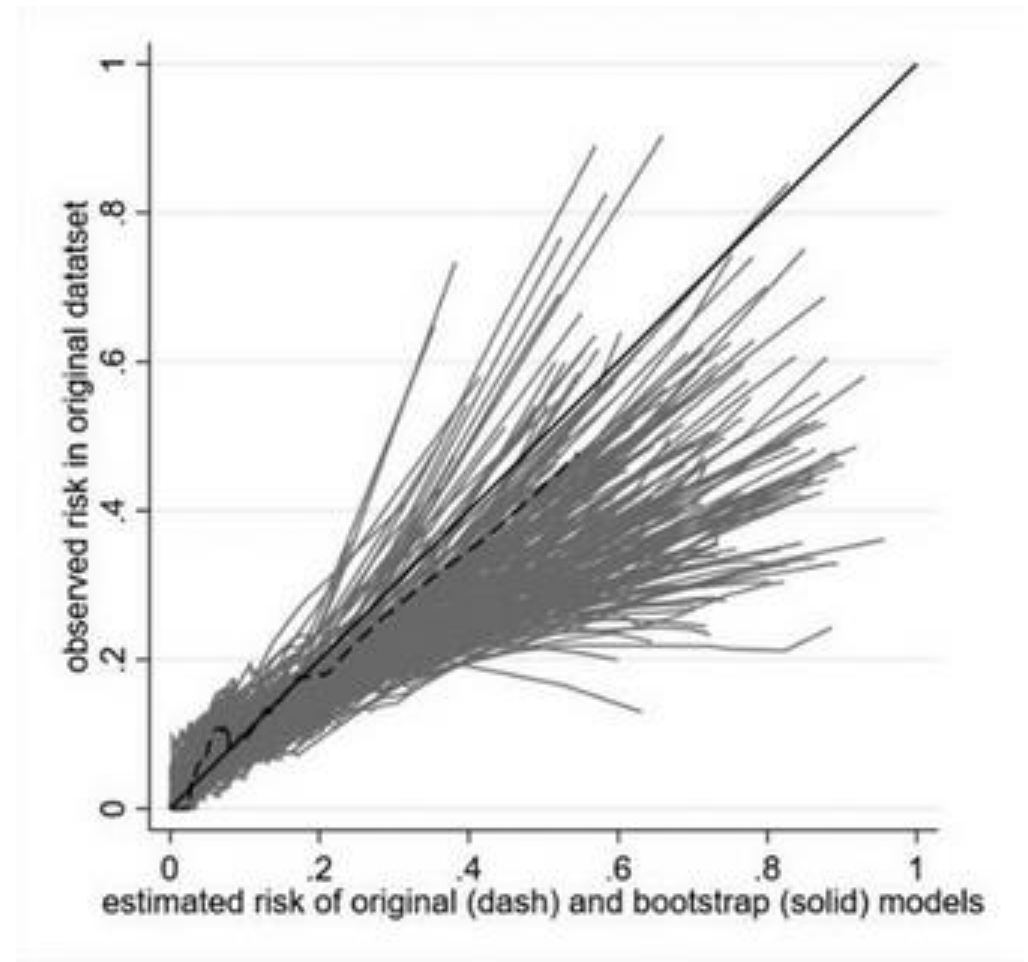
- Instability of a prediction model is reflected by variability of individual-level predictions from the B bootstrap models.
- Plot is B predicted values (y-axis) for each individual against their original predicted value (x-axis).
- A 95% range could be presented for each individual, defined by the 2.5th and 97.5th percentile of prediction values.



## QUANTIFYING INSTABILITY IN MODEL DEVELOPMENT STUDIES

### Calibration instability plot

- Just overlap all calibration curve for all B model onto the same curve.
- The wider the spread of the B calibration curves, the greater the instability
- With many curves the plot may be dense and unclear, and so displaying a random sample of 100 or 200 will often be fine.



## Mean absolute predictor error (MAPE)

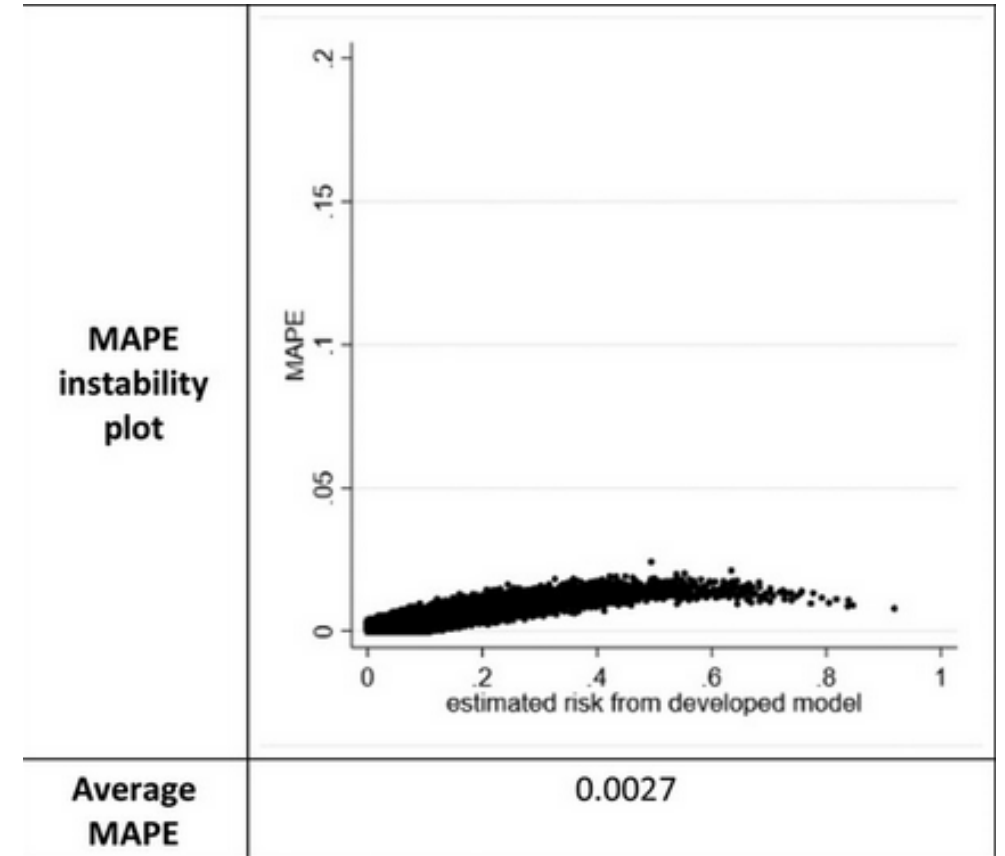
- *Mean Absolute Prediction Error* = the mean absolute difference between the bootstrap model predictions and the original model prediction

$$\text{MAPE for individual } i = \frac{\sum_{b=1}^B |\hat{p}_{bi} - \hat{p}_i|}{B}.$$

- Then we can calculate the average MAPE from all individual

$$\text{average MAPE} = \frac{\sum_{b=1}^B \sum_{i=1}^N |\hat{p}_{bi} - \hat{p}_i|}{BN}.$$

- As average MAPE may obfuscate large individual variability in some subgroup, MAPE should always be accompanied by an instability plot of individual-level MAPE values.





# Experiment

## Experiment setup

### Simulated data

- Generated with **logistic regression**.
- An individual's true logit event risk is determined by a single predictor,  $X$ .
  - $X$  is drawn from normal distribution  $N(0,4)$  with mean 0 and standard deviation of 2.
  - c-statistic of about 0.86 in the population based on  $X$ .
- The outcome,  $Y$  is generated from Bernoulli( $\pi$ ).
  - $\text{Logit}(\pi) = \text{Linear predictor } (\beta_0) + (\beta_1) X$ 
    - where  $\beta_0 = 0$  and  $\beta_1 = 1$
  - $\text{Logit}(\pi) = X$
  - $\pi = \frac{1}{1+e^{-X}}$
- For each sample, also draw 10 noise variable  $Z_1$  to from  $N(0, 1)$ .
- Sample size ( $n$ ) for evaluation = 100 (though also consider 50, 385, 500, 1000, and 5000 in further experiment).
- True overall risk of an outcome = 0.5

### Prediction model

- **Logistic regression** model with  $n$  sample
  - with LASSO regularization
  - with 10-fold cross-validation
  - Using 11 predictors ( $X$  and  $Z_1$  to  $Z_{10}$ ).
- Use the developed model to predict 100,000 other samples generated with the same process.
- **Random forest** is also examined.

Repeat 1,000 times

Total 1,000 models

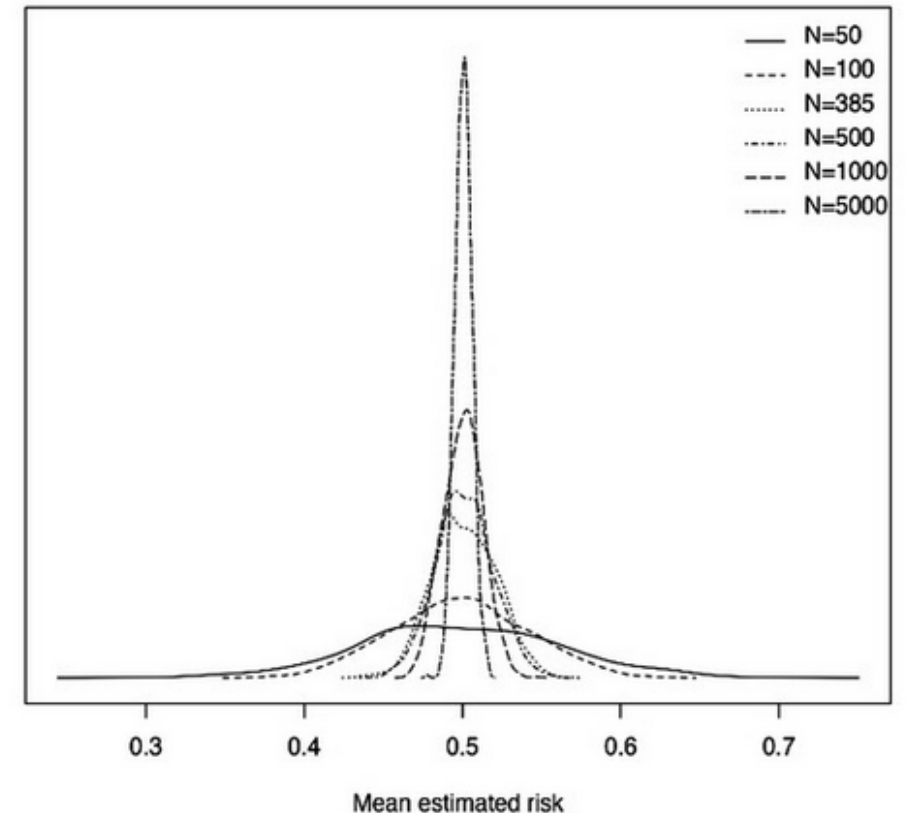
## Level 1:

## Mean estimated risk

- **The bare minimum requirement**
- $n = 50$ 
  - 95% of mean estimated risks = 0.36 - 0.64
- $n = 100$ 
  - 95% of mean estimated risks = 0.42 - 0.58
- $n = 5,000$ 
  - 95% of mean estimated risks = 0.49 - 0.51
- The smaller the development dataset, the greater the instability in a model's mean estimated risk,
- Downstream consequence is miscalibration between the mean estimated and mean observed risk in the population (also known as miscalibration-in-the-large).
- Method to estimate minimum sample size to estimate the mean risk within 0.05 of the true value has been proposed (Riley et al., 2019a, [2020](#)).

### LEVEL 1: INSTABILITY IN MEAN ESTIMATED RISK

(a) Distribution of the mean estimated risk from 1000 example models, for model development sample sizes of 50, 100, 385, 500, 1000 and 5000



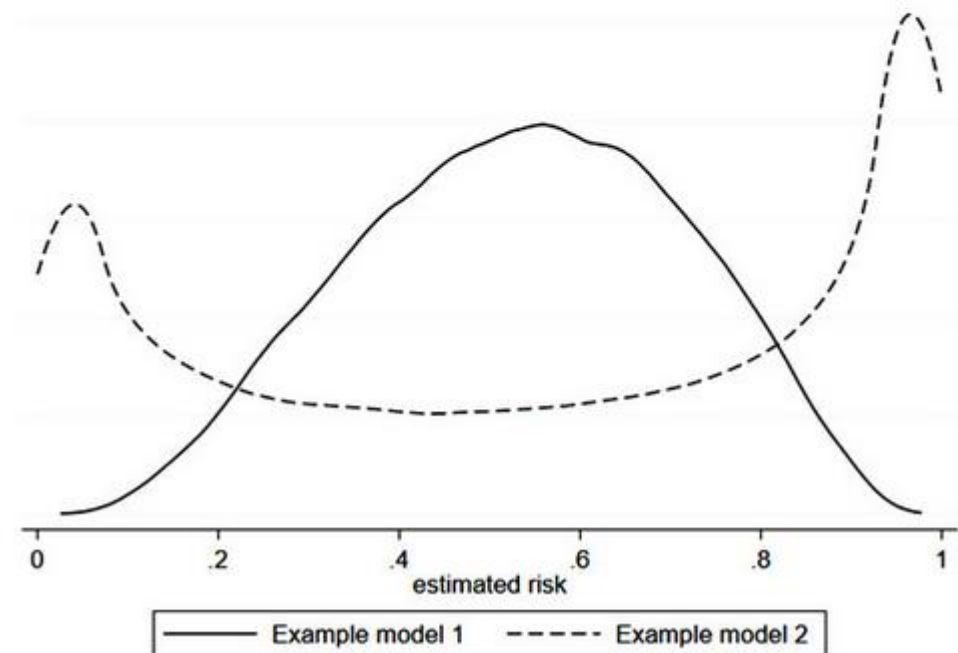
## Level 2:

## Distribution of estimated risks

- With  $n = 100$ , distribution of predicted risk (from 100,000 samples) can be very different.
- Downstream consequence is also miscalibration.

### LEVEL 2: INSTABILITY IN DISTRIBUTION OF ESTIMATED RISKS

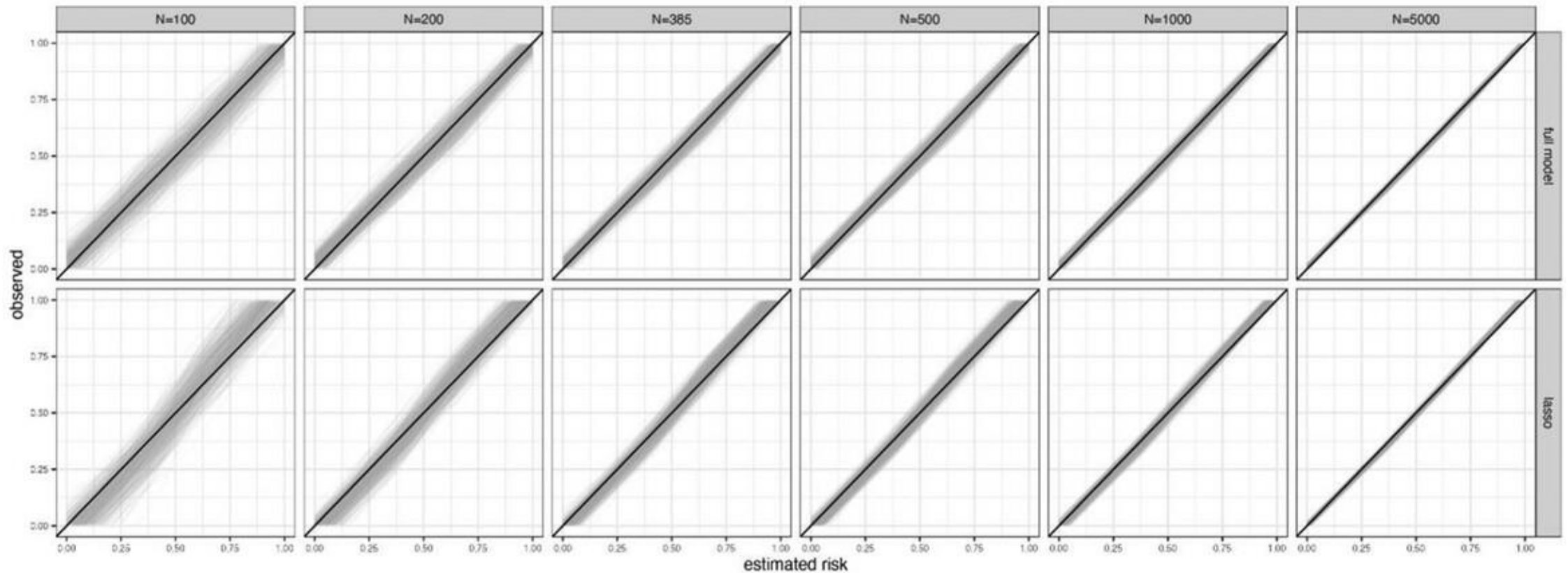
- (b) Distribution of estimated risks from two example models developed using a sample size of 100, which gave very different distributions



## Level 2:

## Distribution of estimated risks

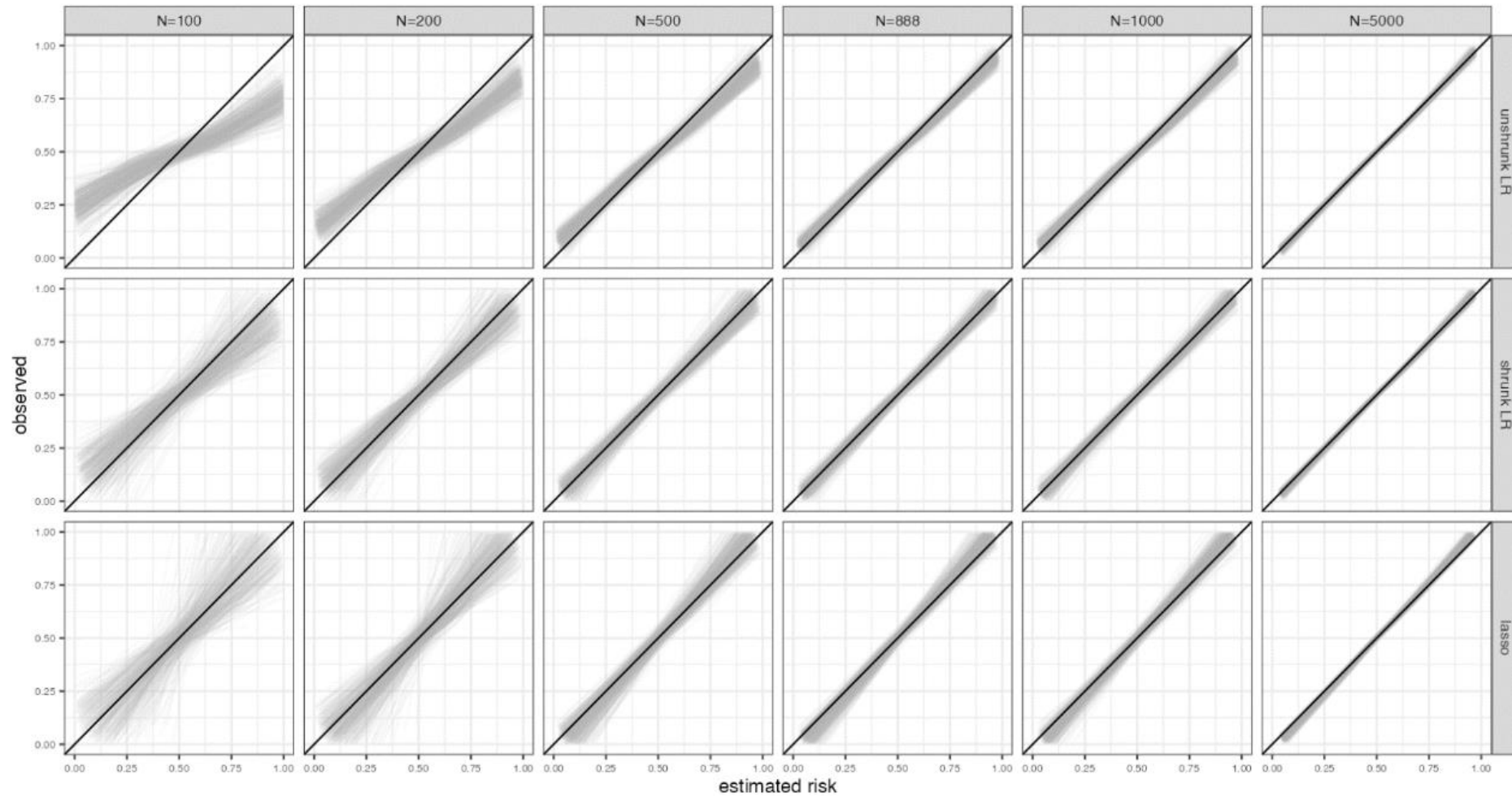
- Calibration curve for each model developed with different N.



## Level 2:

## Distribution of estimated risks

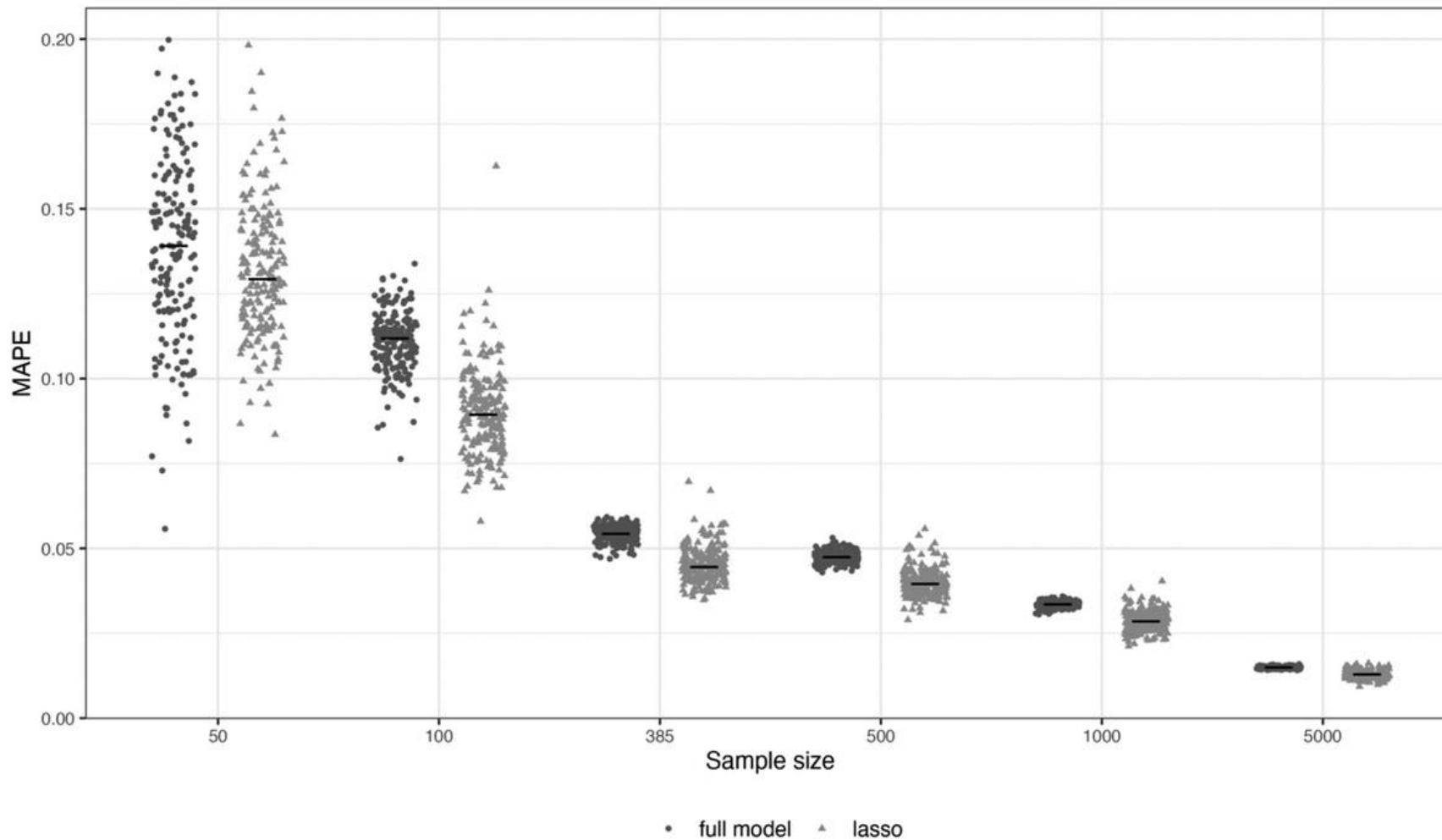
- **More predictors** (5 true predictors and 10 noise in this case) **require larger samples.**



## Level 2:

## Distribution of estimated risks

- Instability demonstrated by mean absolute prediction error (MAPE)



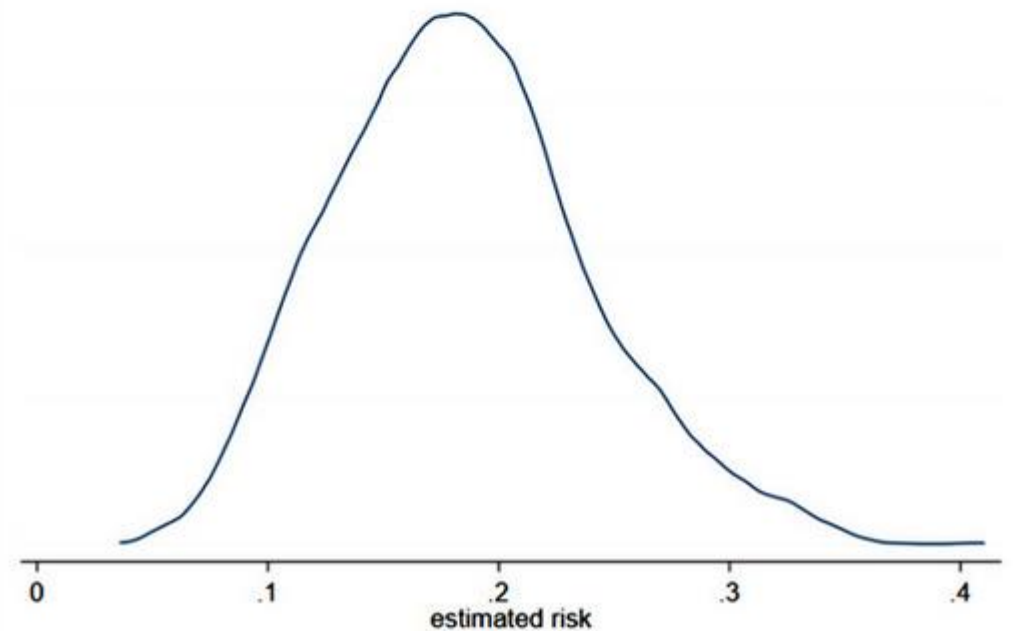
## Level 3:

## Predictions for subgroups

- Even if level 1 and 2 is stable, there may be instability in subgroups.
- Consider the subgroup with  $X$  value  $< -1$  for models developed with  $n=100$ , variability is between 0.1-0.3
- Especially relevant for algorithmic fairness.

### LEVEL 3: INSTABILITY IN ESTIMATED RISK FOR A SUBGROUP

(c) Distribution of the mean estimated risk for the subgroup of individuals with  $X < -1$ , as derived from 1000 example models each developed using a sample size of 100.





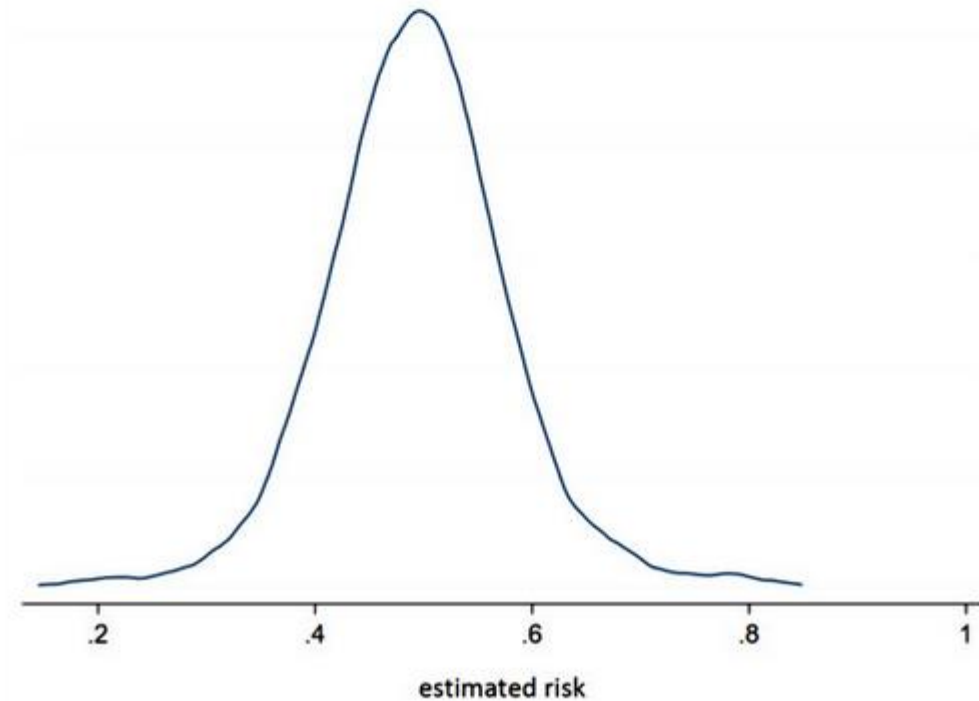
## Level 4:

## Predictions for individual

- Often more severe than other level.
- Is a huge concern in actual clinical usage.
- For the same sample (selected randomly), predicted risk from 1,000 models developed with  $n=100$  range from 0.2-0.8
- This is seriously unreliable.

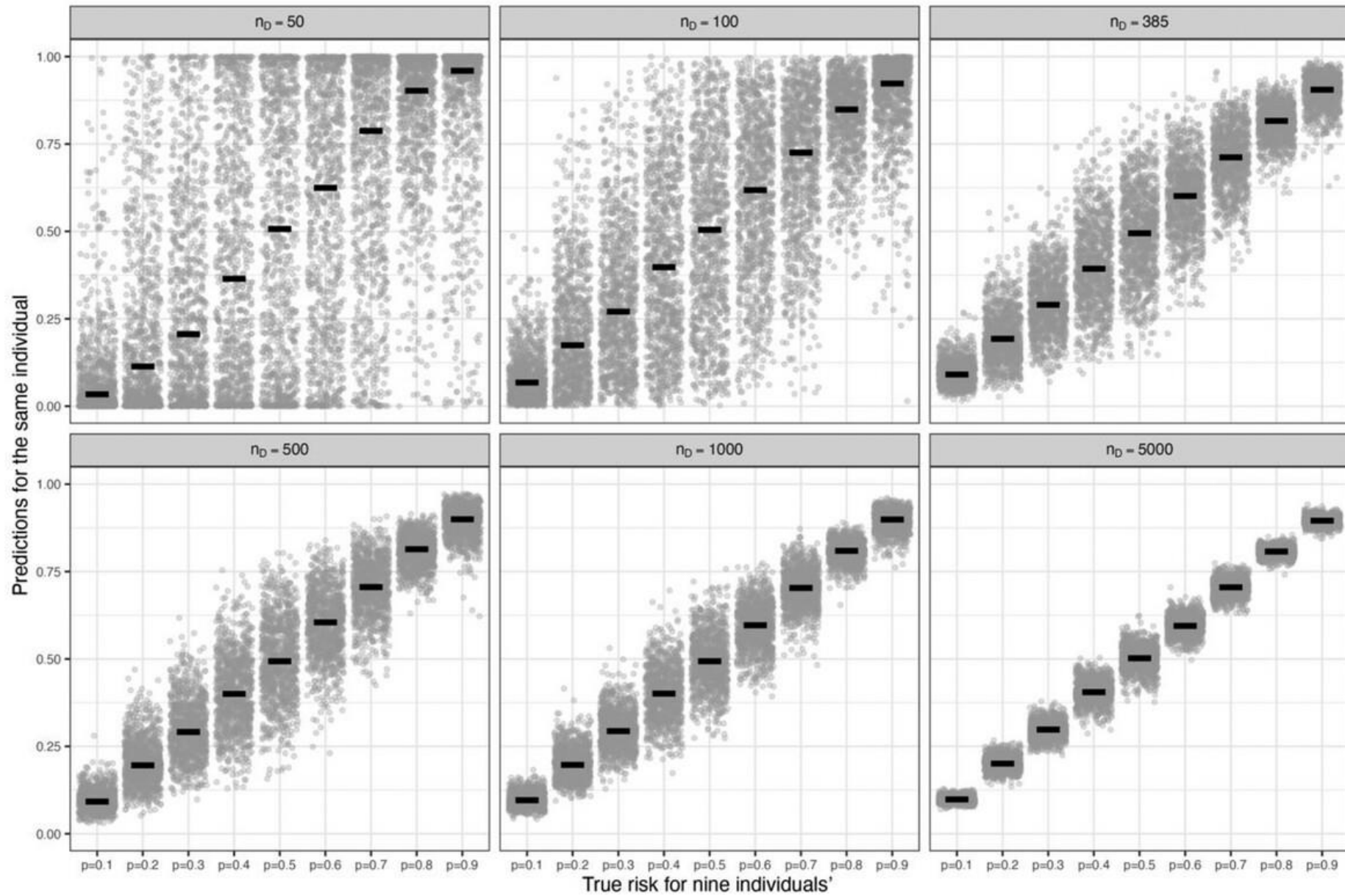
### LEVEL 4: INSTABILITY IN ESTIMATED RISK FOR INDIVIDUAL

(d) Distribution of estimated risk from 1000 example models (each developed using a sample size of 100 participants) for one particular individual whose true risk is 0.5



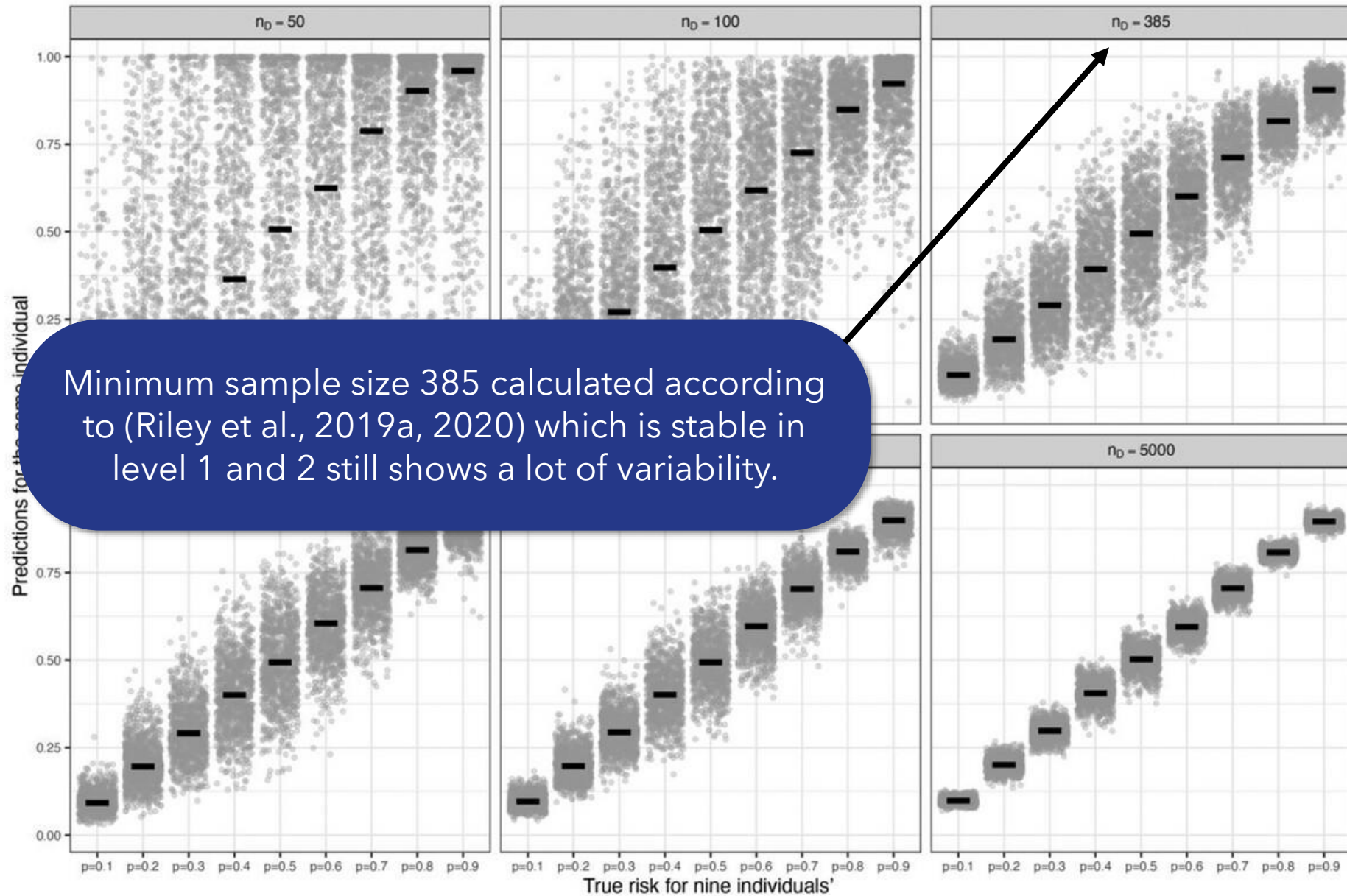
# Level 4:

# Predictions for individual



## Level 4:

## Predictions for individual



## Experiment 2: Other cases for instability

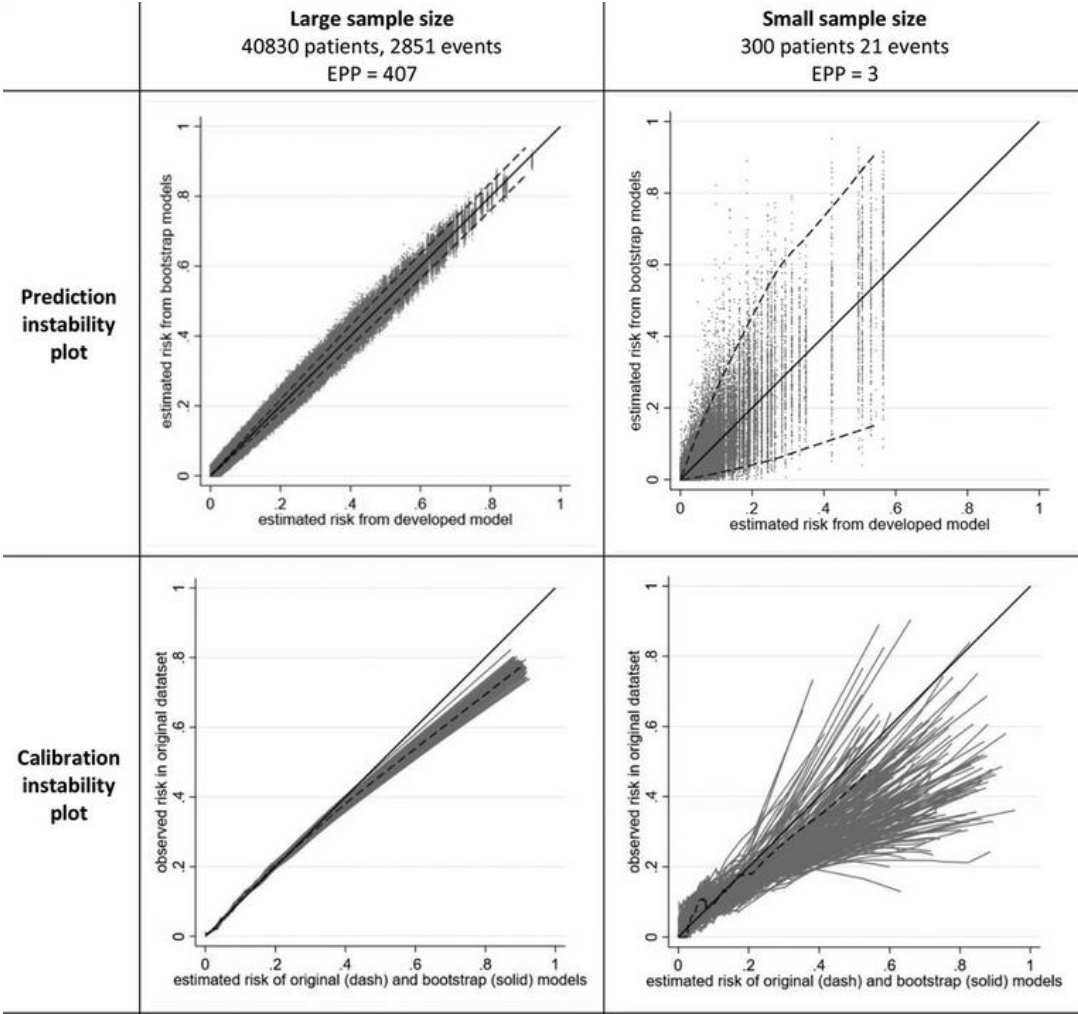
### Data

- GUSTO-I dataset
  - Contains individual participant level information on 30-day mortality following an acute myocardial infarction
- Goal: predict risk of death by 30 days
- N = 40,830
- Event = 2,851 death by 30 days (overall risk = 0.07)
- Predictors:
  - Sex (0 = male, 1 = female)
  - Age (years)
  - Hypertension (0 = no, 1 = yes)
  - Hypotension (0 = no, 1 = yes)
  - Tachycardia (0 = no, 1 = yes)
  - Previous Myocardial Infarction (0 = no, 1 = yes)
  - ST Elevation on ECG (number of leads).

# Experiment 2: Other cases for instability

## Unpenalized logistic regression forcing in seven predictors

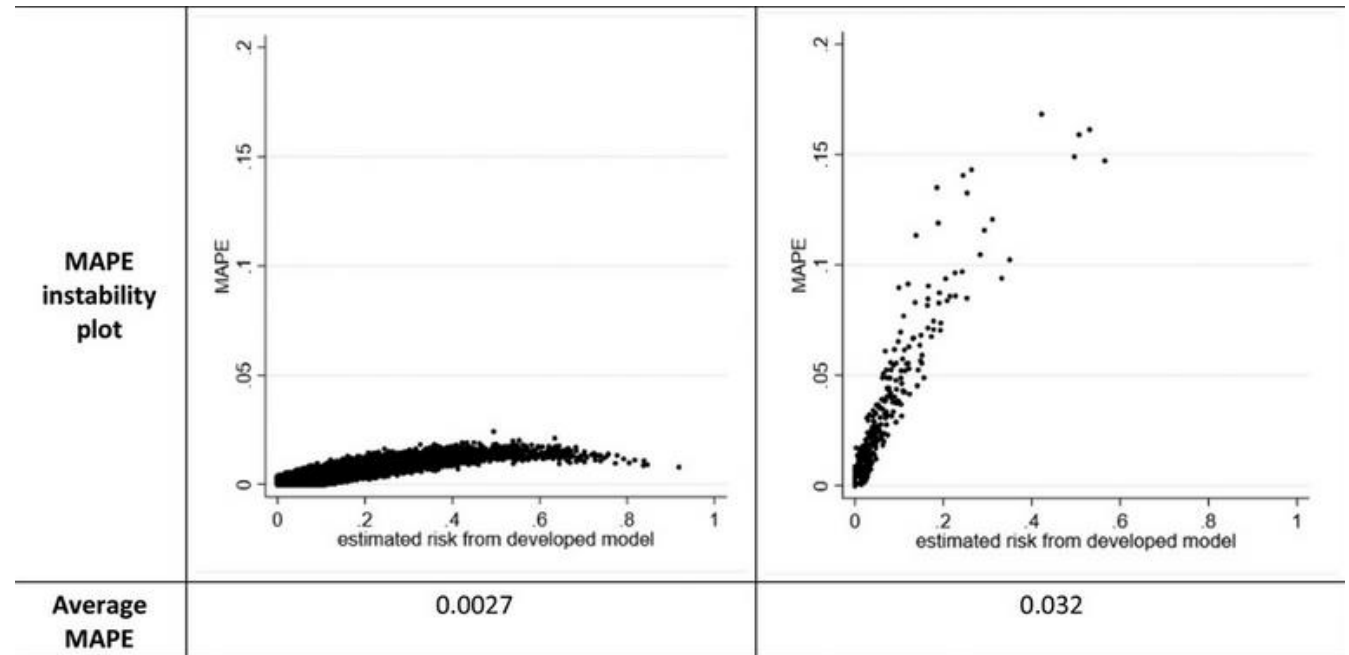
- Large sample size
  - 40,830 participants
  - 2851 deaths
  - 407 Events per Predictor Parameter (EPP)
- Small sample size
  - 300 participants
  - 21 deaths
  - 3 EPP



# Experiment 2: Other cases for instability

## Unpenalized logistic regression forcing in seven predictors

- Large sample size
  - 40,830 participants
  - 2851 deaths
  - 407 Events per Predictor Parameter (EPP)
- Small sample size
  - 300 participants
  - 21 deaths
  - 3 EPP

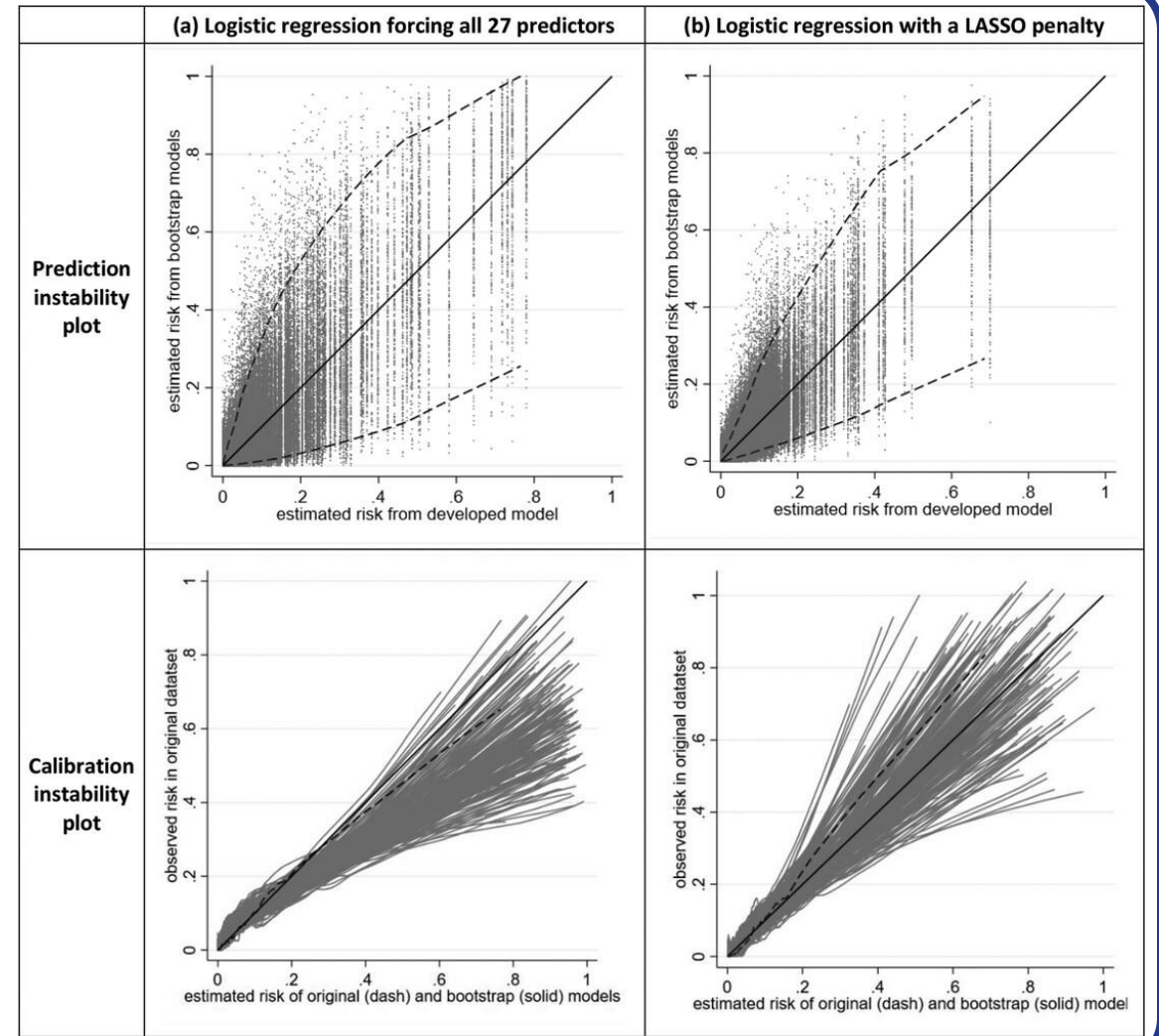




## Experiment 2: Other cases for instability

### Many noise variables and the LASSO

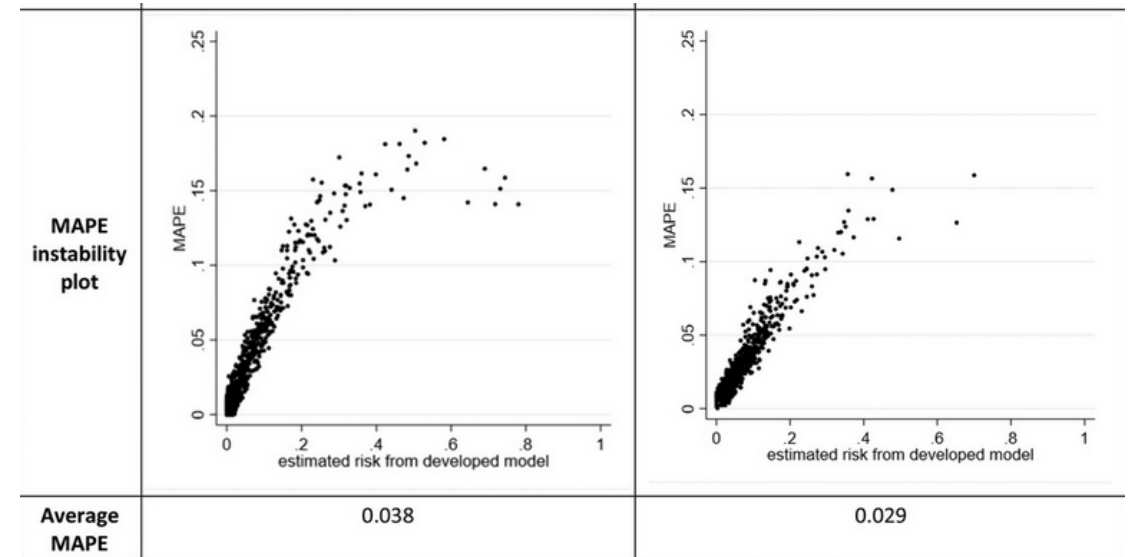
- In this experiment, additional 20 noise variables generated from  $N(0,1)$  were added to the model (total 27 predictors).
- $N = 752$  participants, 53 deaths, 2 EPP
- LASSO has been employed to reduce overfitting in high-dimensional data with low EPP. However, in this case, even with LASSO, the variability was still huge.



## Experiment 2: Other cases for instability

### Many noise variables and the LASSO

- In this experiment, additional 20 noise variables generated from  $N(0,1)$  were added to the model (total 27 predictors).
- $N = 752$  participants, 53 deaths, 2 EPP
- LASSO has been employed to reduce overfitting in high-dimensional data with low EPP. However, in this case, even with LASSO, the variability was still huge.





# Experiment 2: Other cases for instability

## LASSO and uniform shrinkage with a minimum sample size

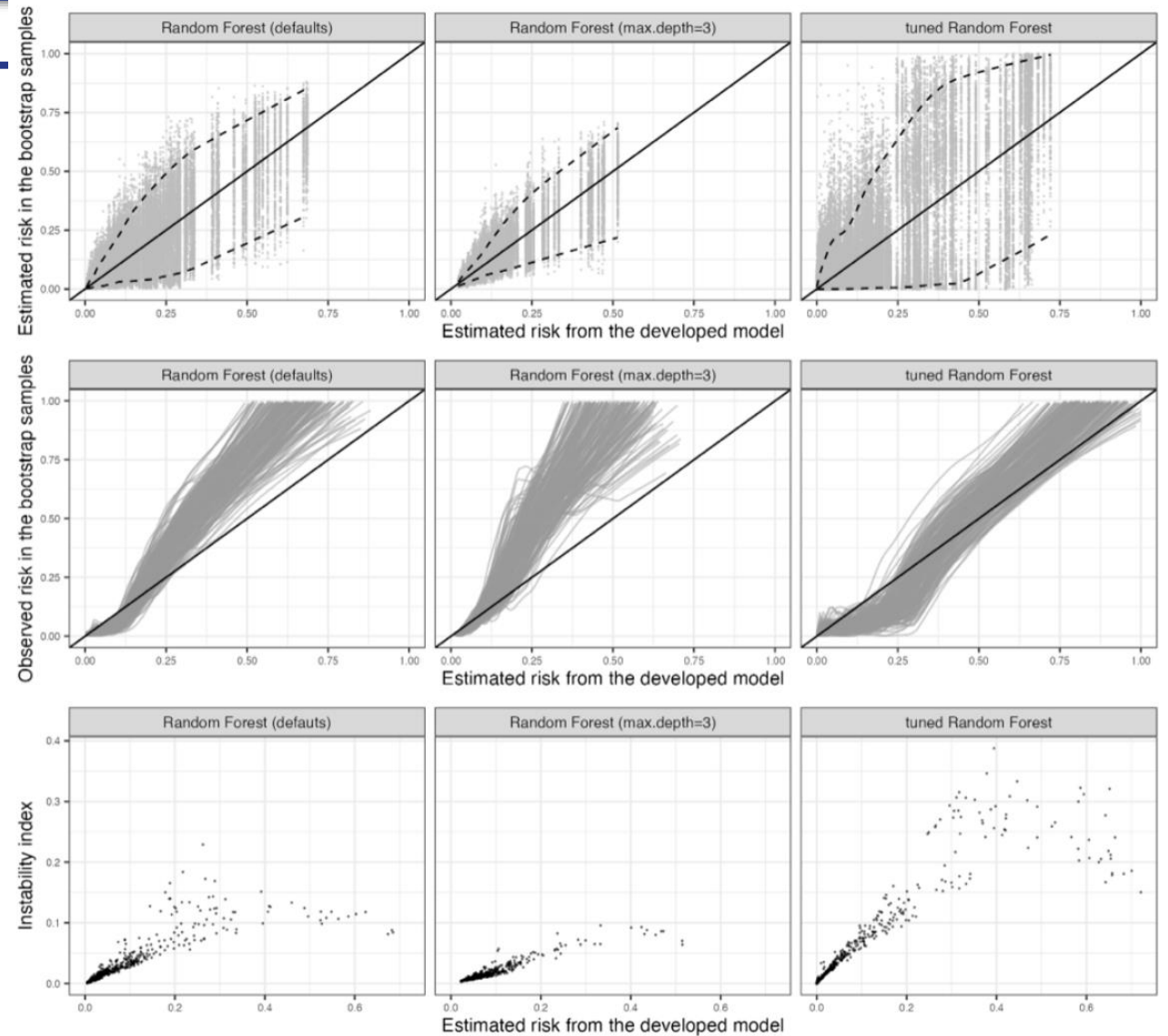
- Using minimum sample size criteria (Riley et al., 2020) with 7 predictors
  - Assuming a Cox–Snell  $R^2$  of 0.08
  - Targeting a uniform shrinkage factor of 0.9
  - $N = 752$  participant, 53 deaths, 7.5 EPP
- The two models, demonstrate similar variability, suggesting that the choice of penalization approach is less important as the sample size increases.

	(a) logistic regression with LASSO	(a) unpenalised logistic regression followed by uniform shrinkage of predictor effects
Prediction instability plot		
MAPE instability plot		
Average MAPE	0.019	0.018

## Experiment 2: Other cases for instability

### Random forests and hyperparameter tuning

- Instability also applies to modeling approaches other than (penalized) regression, such as random forests.
- Use the same sample as previously
  - N = 752 participant, 53 deaths, 7.5 EPP
- With the same number of trees (100), limiting the depth to 3 reduce instability.
- Automate hyperparameter tuning (third row) increase instability as compared to pre-chosen parameter.

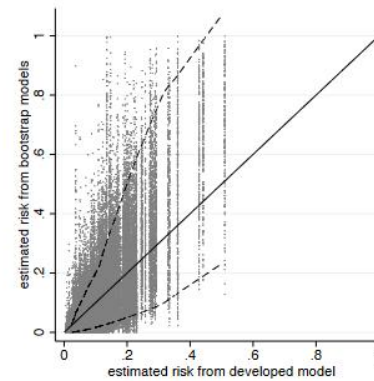


## Experiment 2: Other cases for instability

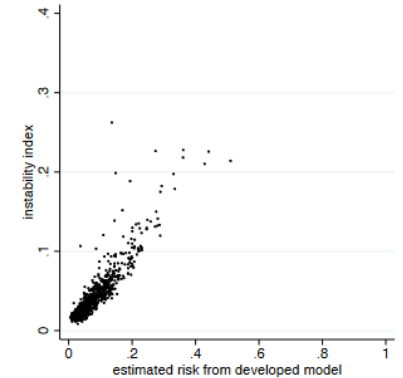
### Impact of data splitting for recalibration

- A holdout dataset is sometimes used to recalibrate developed model.
- To examine stability of this approach, we randomly split the 752 participants into two parts
  - N = 452 for developing the random forest (Tree number = 100, depth = 3)
  - Then follow by recalibration with N = 300
- The bootstrap approach include the recalibration process.
- MAPE increase from 0.019->0.045, mainly due to splitting the dataset, which reduces the sample size.

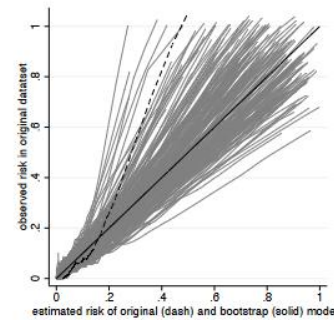
Prediction instability plot



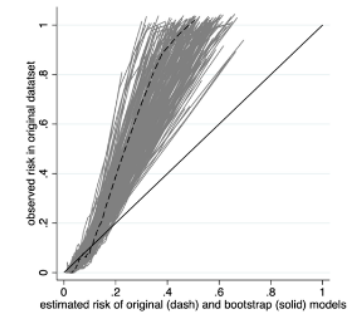
MAPE (mean 0.045)



Calibration instability plot



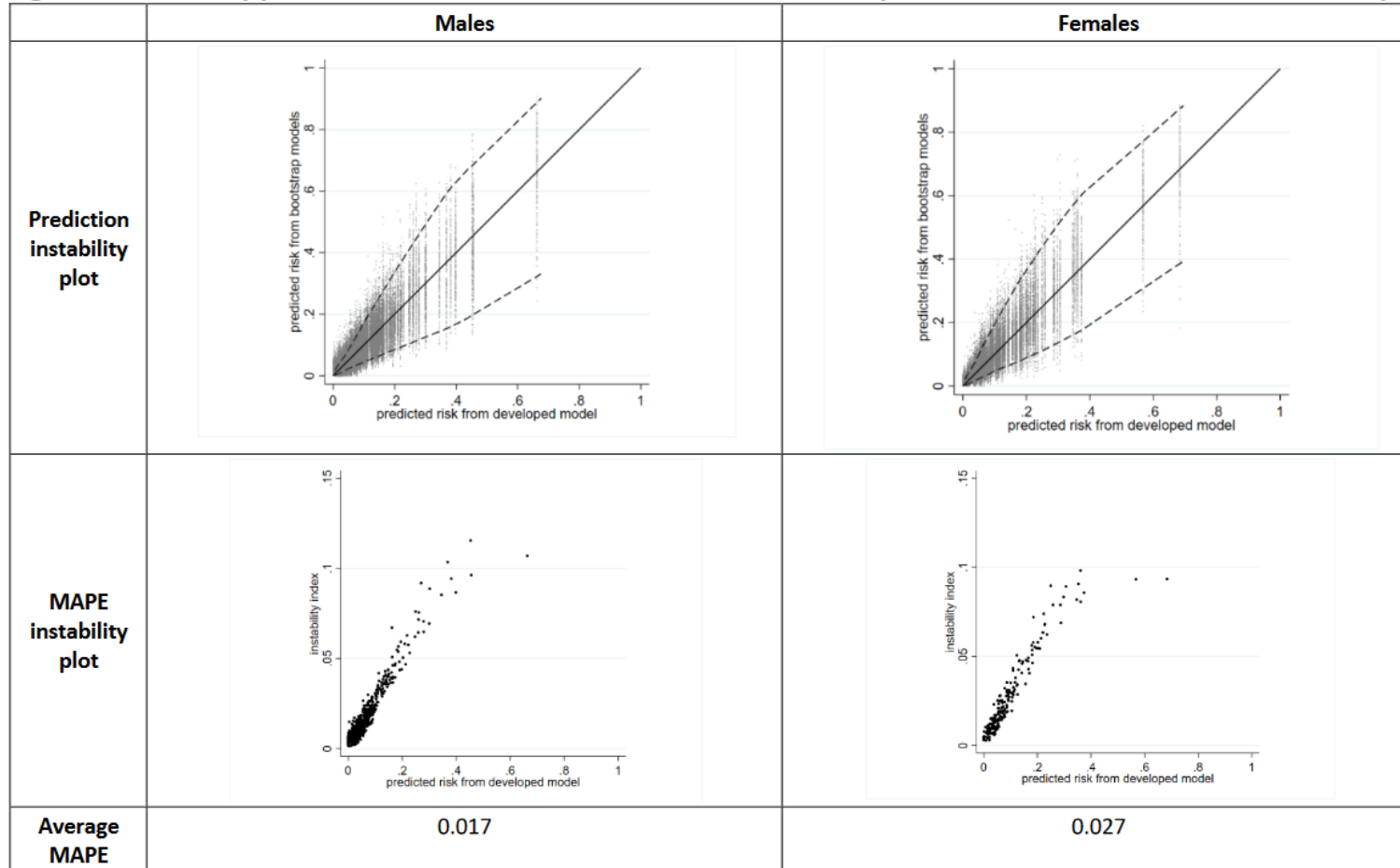
Calibration instability plot when using all 752 participants for model development without recalibration



# Further role of stability assessment

## examining stability in subgroups

Figure S5: Instability plots and measures to examine fairness of a LASSO prediction model in males and females separately (see Section 5.1)

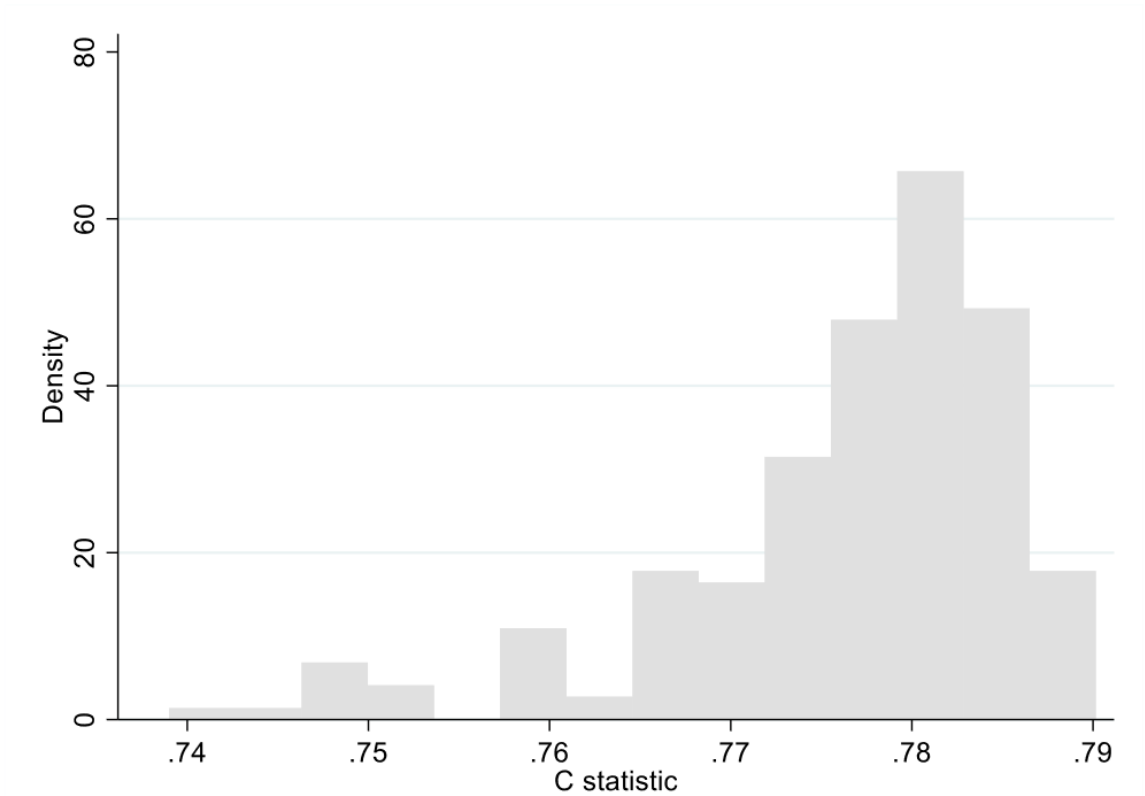


# Further role of stability assessment

## c-statistic

- c-statistic can also be calculated and plotted for each bootstrap model to show instability.

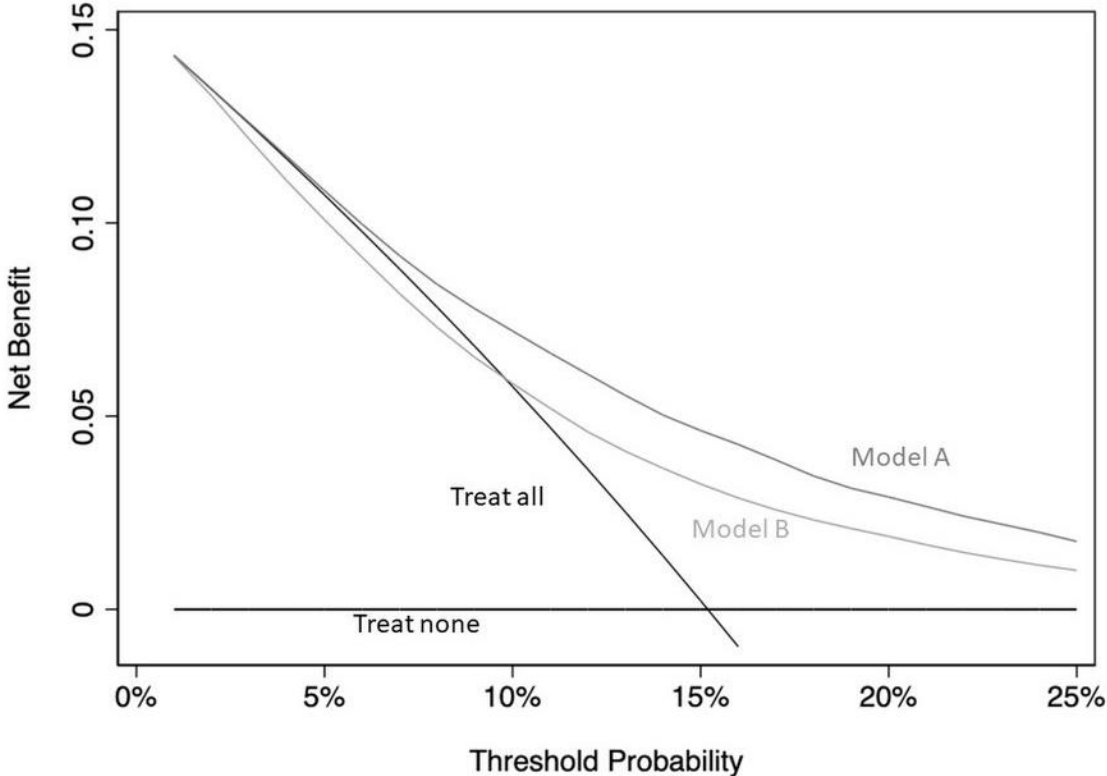
Figure S6: Histogram of C-statistics from applying the LASSO model in Section 5.2 to the bootstrap samples



# Further role of stability assessment

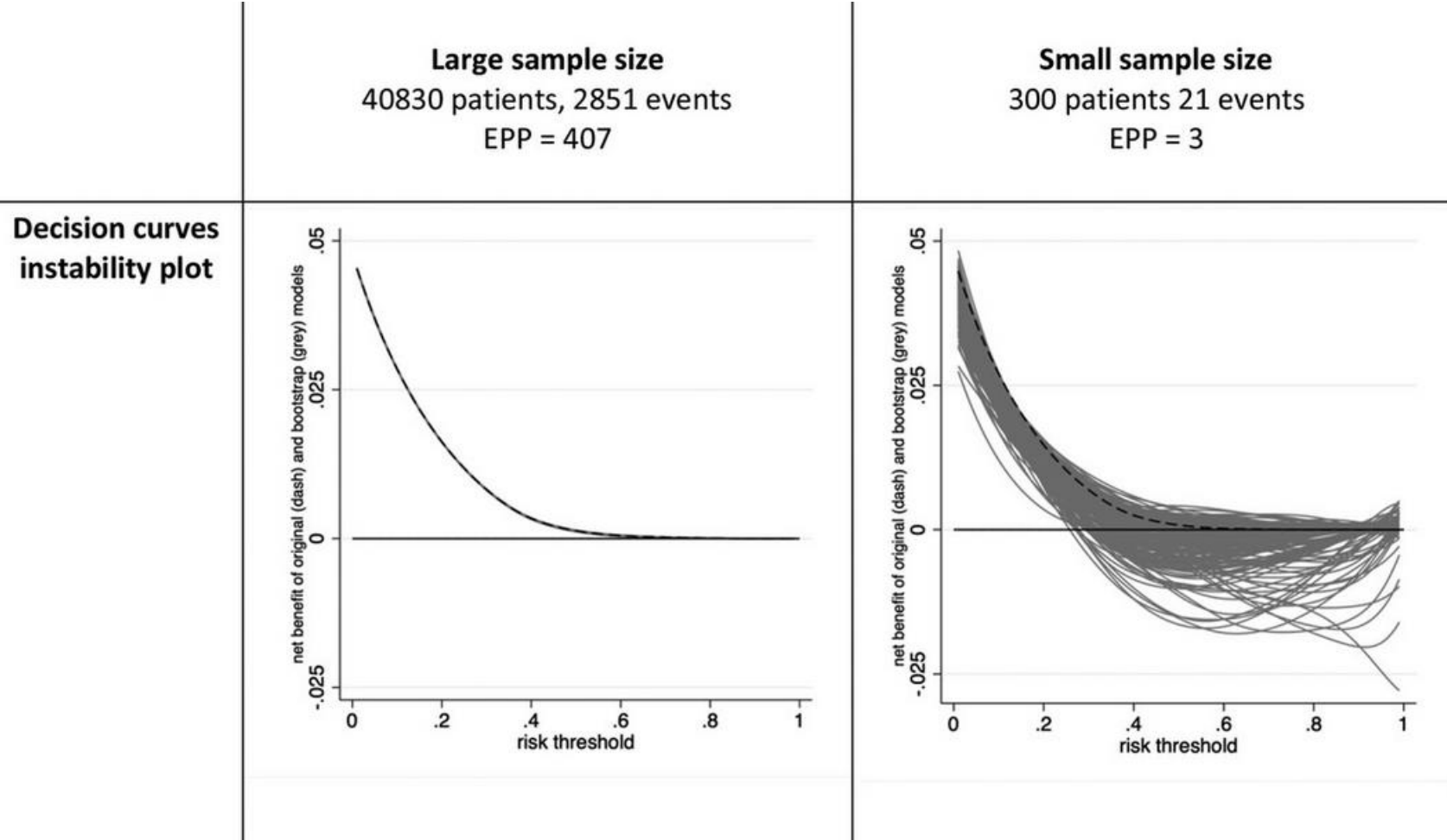
## Clinical utility

- For directing clinical decision making, a certain **threshold value** will be decided which when exceeded may take some action (such as treatment).
- A **decision curve** can be used to display a model's net benefit across the range of chosen threshold values.



# Further role of stability assessment

Clinical utility

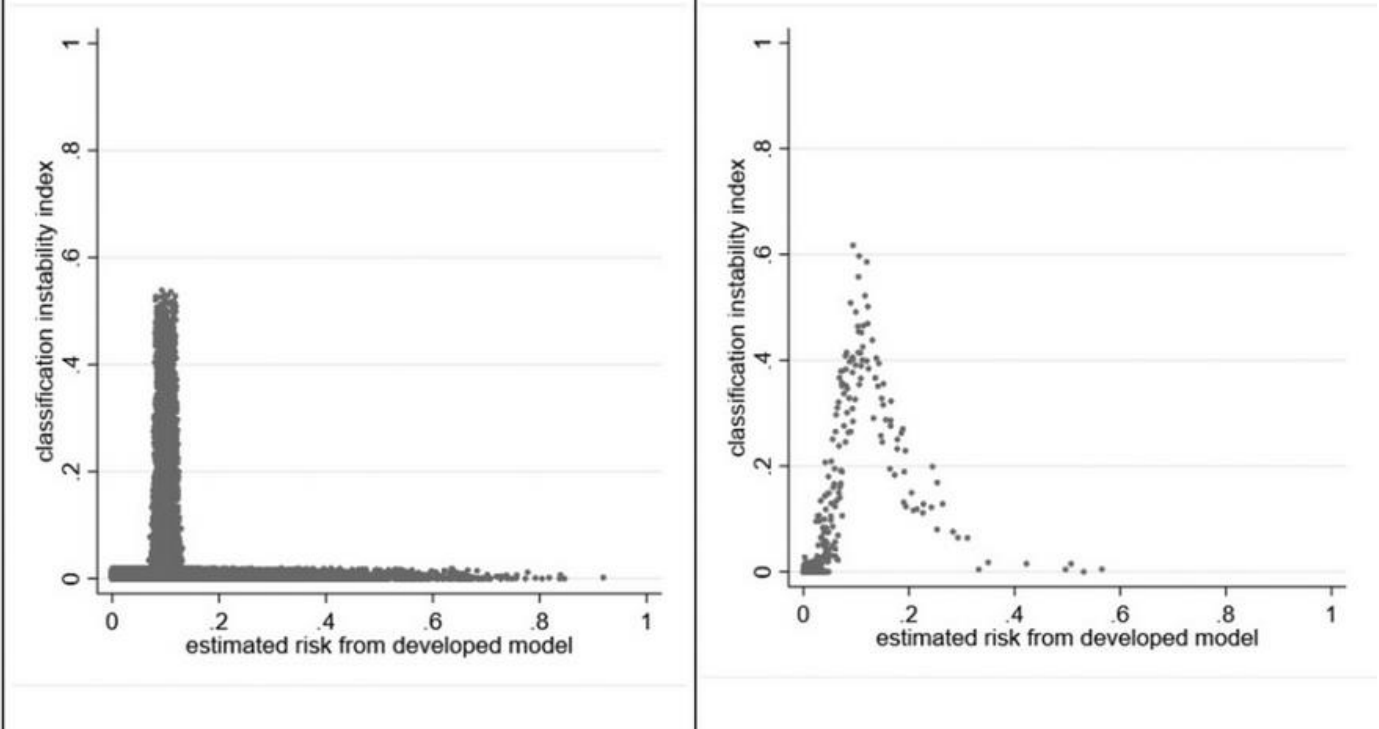


# Further role of stability assessment

## Classification and risk grouping

- Classification index for each individual can be calculated from the proportion of bootstrap models that give a different classification (i.e., above rather than below the threshold, or below rather than above the threshold) than the original model. Classification stability plot plots the index with the predicted risk from the original model.

**Classification instability plot (based on classifying individuals as  $<0.1$  or  $\geq 0.1$ )**

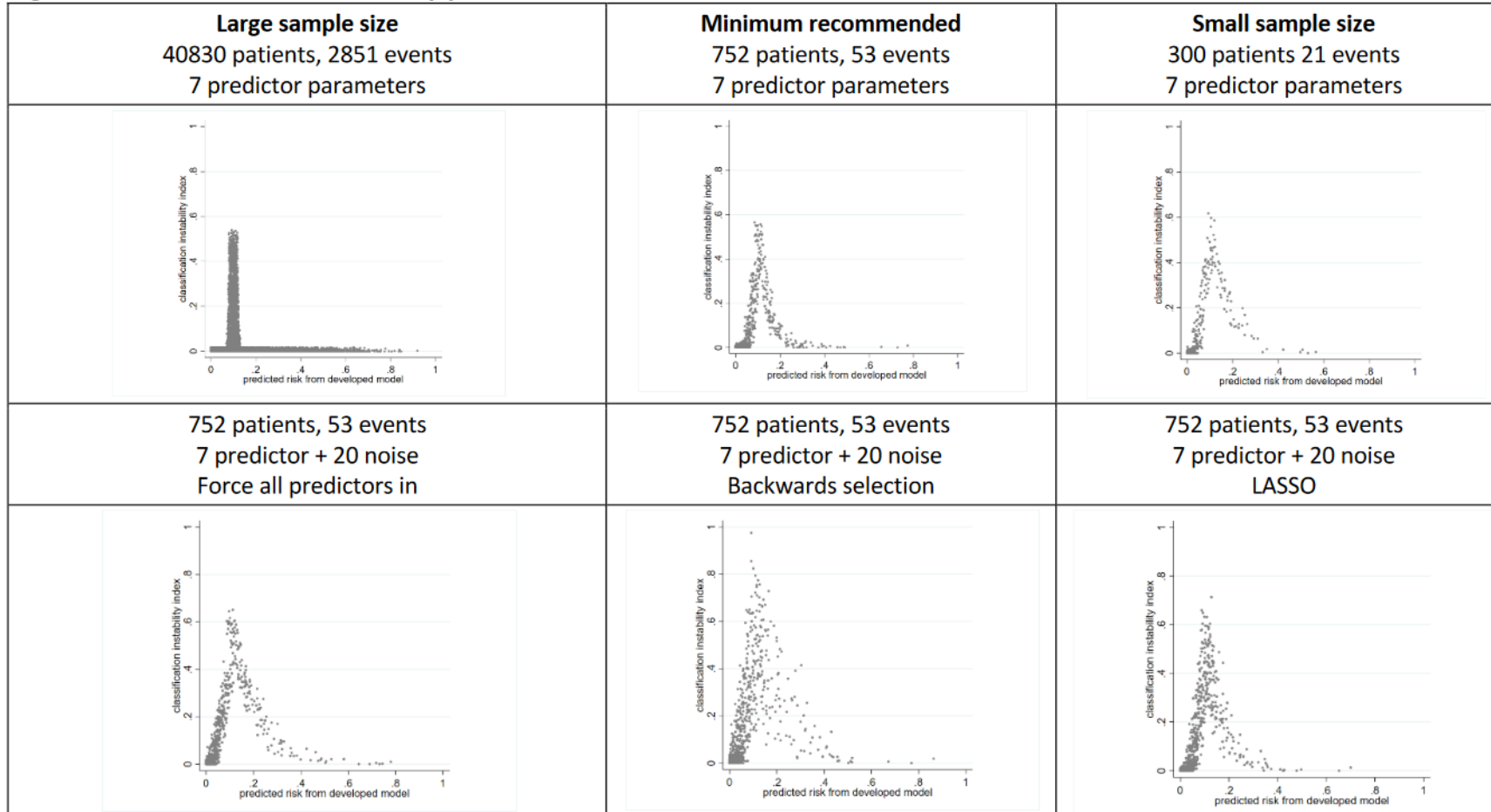




# Further role of stability assessment

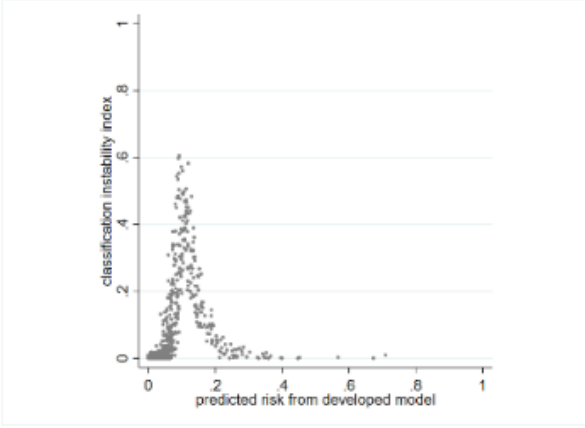
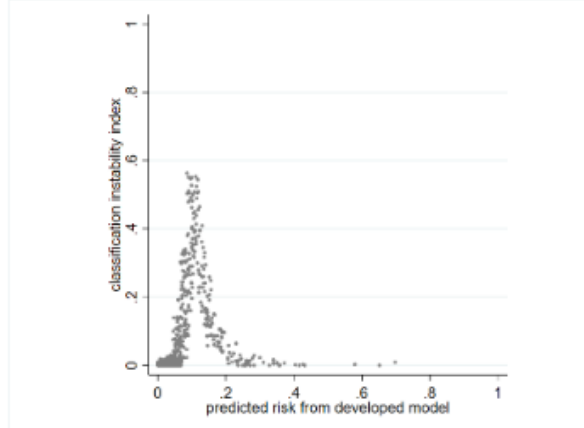
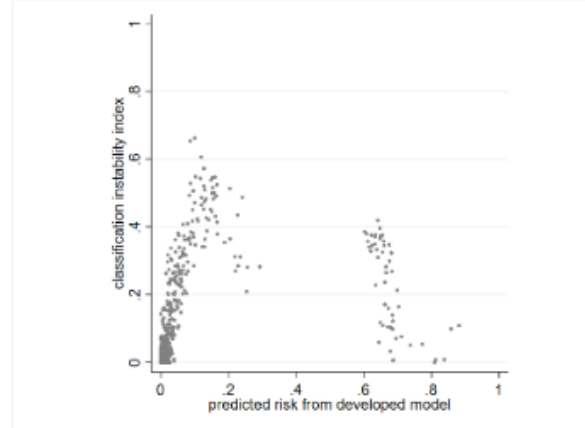
## Classification and risk grouping

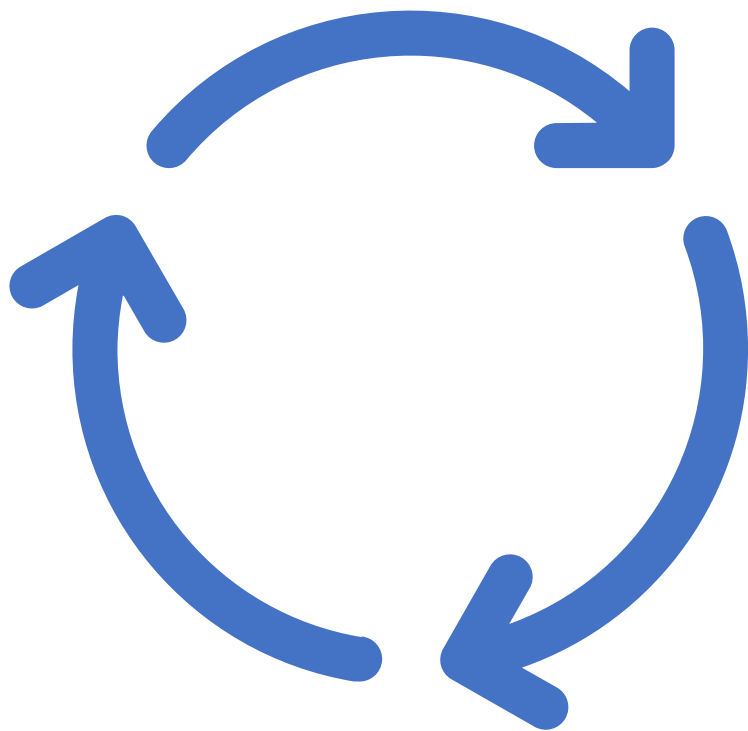
**Figure S7: Classification instability plots for various models from the case studies of Section 4.**



# Further role of stability assessment

## Classification and risk grouping

<b>Minimum recommended</b> 752 patients, 53 events 7 predictor parameters LASSO	<b>Minimum recommended</b> 752 patients, 53 events 7 predictor parameters Uniform	<b>Minimum recommended</b> 752 patients, 53 events 7 predictor parameters Random forest
		



# Discussion

## Discussion

### Bootstrap quality

- Popular model development methods like the LASSO **do not resolve issues of small sample sizes and low EPP.**
- Bootstrapping is an important method for checking stability
  - However, **bootstrap only examines instability in the population that the development dataset was sampled from.**
  - **If the sample size is too small, impact of outliers in the bootstrap process may be large.**
  - **Some aspect of model development can also be hard to automated.**

## Discussion

### Level of instability

- At the bare minimum, a model should demonstrate stability at levels 1 and 2
  - **Mean estimated risk**
  - **Distribution of estimated risk**
- Minimum sample size (and number of predictor parameters) for model development should target **precise estimation of the overall risk in the population, low overfitting, and small average MAPE.**
- For level 3 and 4 (**group** and **individual stability**), large samples size are required, which is not always possible for rare outcomes.
  - **Data sharing** and individual participant data meta-analysis may help to address this
- Regardless of sample size, **stability checks should always be undertaken and reported.**

## Discussion

### Stability in the context of clinical decision

- Instability in individual-level predictions and classifications is **inevitable and to be expected**.
- A model may still have population-level benefit even with instability at the individual level.
- We may desire greatest stability in regions of risk relevant to clinical decision making and be willing to accept lower stability in other regions where miscalibration is less important.
  - **Example: recurrence of venous thromboembolism**
    - Predicted risks between about 0.03 - 0.20 have been suggested to warrant clinical action.
    - Slight to moderate instability in ranges of highest risk (0.5–1) is potentially acceptable
    - **Decision curve stability plot** can help with this.