

# Prediction models in healthcare: a playground for researchers?

**Richard D. Riley**  
*Professor of Biostatistics*

Institute of Applied Health Research  
University of Birmingham, UK

e-mail: [r.d.riley@bham.ac.uk](mailto:r.d.riley@bham.ac.uk)

X: [@Richard\\_D\\_Riley](#)

BlueSky: [@richarddriley](#)

**BIG THANKS: Gary Collins (Oxford)**

**FUNDING: NIHR Birmingham Biomedical Research Centre,  
ESPRC grant for AI to accelerate healthcare research**

# Prediction models in healthcare: a playground for researchers

**Richard D. Riley**  
*Professor of Biostatistics*

Institute of Applied Health Research  
University of Birmingham, UK

e-mail: [r.d.riley@bham.ac.uk](mailto:r.d.riley@bham.ac.uk)

X: [@Richard\\_D\\_Riley](#)

BlueSky: [@richarddriley](#)

**BIG THANKS: Gary Collins (Oxford)**

**FUNDING: NIHR Birmingham Biomedical Research Centre,  
ESPRC grant for AI to accelerate healthcare research**

*“Patient trust was essential in the healing process. It could be won by a punctilious bedside manner, by meticulous explanation, **and by mastery of prognosis, an art demanding experience, observation and logic”***

*Galen, 2<sup>nd</sup> Century AD*

PORTER, R. 1999. *The greatest benefit to mankind : a medical history of humanity from Antiquity to the present*, London, FontanaPress

Part 1

# **INTRODUCTION TO PREDICTION MODELS IN HEALTHCARE**

# Prediction model research


- **Prediction models utilise multiple prognostic factors (predictors, features) to estimate the risk of a particular outcome in individuals**
- A **useful** model provides accurate predictions that:
  - reliably inform patients & health professionals about outcome risks
  - guide healthcare decisions that improve outcomes
  - improve clinical research (e.g. trial randomisation)
- **Crucial: focus is estimating (predicting) values for individuals**
- **Based on (penalised) regression models, random forests, neural networks etc**

# Example: outcome risk in traumatic brain injury

Web-tool below used to calculate 14 day mortality risk, & 6-month unfavourable outcome risk *for an individual based on multiple prognostic factors in combination*

## Head injury prognosis

These prognostic models may be used as an aid to estimate mortality at 14 days and death and severe disability at six months in patients with traumatic brain injury (TBI). The predictions are based on the average outcome in adult patients with Glasgow coma score (GCS) of 14 or less, within 8 hours of injury, and can only support - not replace - clinical judgment. Although individual names of countries can be selected in the models, the estimates are based on two alternative sets of models (high income countries or low & middle income countries).

Country	Australia
Age, years	47
Glasgow coma score	9
Pupils react to light	One
Major extra-cranial injury? 	No
CT scan available?	<input type="checkbox"/>

### Prediction

Risk of 14 day mortality (95% CI)	14.2% (9.6 - 20.5)
Risk of <u>unfavourable outcome</u> at 6 months	48.9% (39.0 - 58.9)

# Prediction models are hot topic - inform clinical & public health guidelines

- **Framingham Risk Score & QRISK2 (NICE CG67)**
  - 10-year CVD risk
- **Nottingham Prognostic Index (NICE CG80)**
  - Recurrence & survival in breast cancer patients
- **FRAX & QFracture (NICE CG146)**
  - 10-year osteoporotic and hip fracture risk
- **GRACE/PURSUIT/PREDICT/TIMI (NICE CG94)**
  - Adverse CV outcomes in patients with UA/NSTEMI
- **APGAR (NICE CG132/2)**
  - Newborn prognosis
- **SAPS & APACHE (NICE CG50)**
  - ICU scoring systems
- **Leicester Diabetes Risk Score, QDSCORE, Cambridge Risk score (NICE PH38)**
  - Type 2 diabetes

# What do we need?

- Predictions should be accurate and clinically useful
- We should know the model's predictive performance
- Does it give estimated risks that,
  - calibrate closely with observed risks?
  - discriminate (separate) those who do & do not develop the outcome?
  - provide clinical utility (e.g. guide decisions at particular risk thresholds)
- Has the model been shown to work in intended populations and settings of interest? (validation studies)



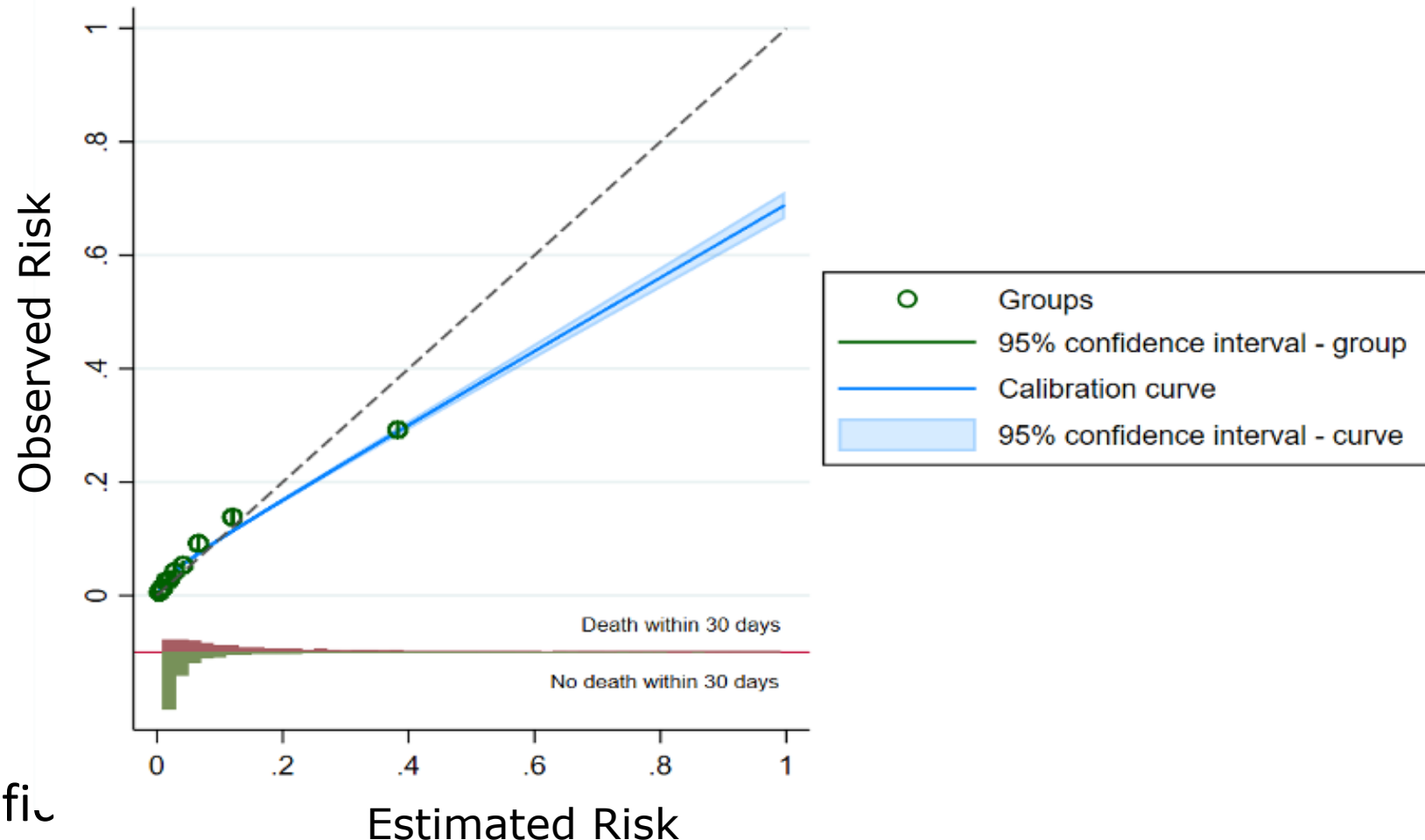
# What do we need?

- Predictions should be accurate and clinically useful
- We should know the model's predictive performance
- Does it give estimated risks that,
  - calibrate closely with observed risks? **CALIBRATION PLOTS & STATISTICS**
  - discriminate (separate) those who do & do not develop the outcome?  
**AUROC / C-STATISTIC**
  - provide clinical utility (e.g. guide decisions at particular risk thresholds)  
**NET BENEFIT & DECISION CURVES**
- **Requires careful statistical modelling and assessment**

# Calibration plots especially important

Example: Prediction model for 30-day mortality following acute MI

- Dotted line is ideal
- Calibration curve important (often just get green groupings)
- O/E = 1.01 (ideal 1)
- Cal slope = 0.72 (ideal 1)
- C-statistic (AUROC) = 0.81
- C range is 0.5 to 1
- But a 'good' C is context specific



Part 2

# **THE PLAYGROUND OF PREDICTION MODEL RESEARCH**

“We have to reduce our expectations of England  
and we have the players to do it”

***Steve McLaren***  
***(England Football Manager)***

# Landscape of clinical prediction models

- 408 models for COPD prognosis (Bellou, 2019)
- 363 models for cardiovascular disease general population (Damen, 2016)
- 263 prognosis models in obstetrics (Kleinrouweler, 2016)
- 258 models mortality after general trauma (Munter, 2017)
- 232 models related to COVID-19 (Wynants, 2020)
- 160 female-specific models for cardiovascular disease (Baart, 2019)
- 119 models for treatment response in pulmonary TB (Peetluk, 2021)
- 101 models for in vitro fertilisation (Ratna, 2020)
- 99 models for stroke in type-2 diabetes (Chowdhury, 2019)
- 81 models for graft failure in kidney transplantation (Kabore, 2017)
- 74 models for length of stay in ICU (Verburg, 2016)
- 73 models for low back pain (Haskins, 2015)
- 68 models for pediatric early warning systems (Trubey, 2019)
- 67 models for treatment response in pulmonary TB (Peetluk, 2021)
- 64 models for in vitro fertilisation (Ratna, 2020)
- 61 models for stroke in type-2 diabetes (Chowdhury, 2019)
- 58 models for graft failure in kidney transplantation (Kabore, 2017)
- 52 models for length of stay in ICU (Verburg, 2016)
- 52 models for low back pain (Haskins, 2015)
- 48 models for pediatric early warning systems (Trubey, 2019)
- 46 models for treatment response in pulmonary TB (Peetluk, 2021)
- 46 models for in vitro fertilisation (Ratna, 2020)
- 43 models for stroke in type-2 diabetes (Chowdhury, 2019)
- 42 models for graft failure in kidney transplantation (Kabore, 2017)
- 40 models for length of stay in ICU (Verburg, 2016)

- Very few have been 'validated' in new data & compared

- Calibration & clinical utility rarely assessed

- Models are easy to create

- was there any intention for them to be used?

- or just extra line on a CV?

- **Should we trust them? Mostly no!**

- **Could they be harmful? Yes!**

- (decisions based on unreliable predictions)**

- 40 models for incident heart failure (Sahle, 2017)

Thanks to Maarten van Smeden for this slide

# COVID19 PANDEMIC - An opportunity to take centre(ish) stage

## RESEARCH

 OPEN ACCESS

 Check for updates

 **FAST TRACK**

## Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Laure Wynants,<sup>1,2</sup> Ben Van Calster,<sup>2,3</sup> Gary S Collins,<sup>4,5</sup> Richard D Riley,<sup>6</sup> Georg Heinze,<sup>7</sup> Ewoud Schuit,<sup>8,9</sup> Marc M J Bonten,<sup>8,10</sup> Darren L Dahly,<sup>11,12</sup> Johanna A A Damen,<sup>8,9</sup> Thomas P A Debray,<sup>8,9</sup> Valentijn M T de Jong,<sup>8,9</sup> Maarten De Vos,<sup>2,13</sup> Paula Dhiman,<sup>4,5</sup> Maria C Haller,<sup>7,14</sup> Michael O Harhay,<sup>15,16</sup> Liesbet Henckaerts,<sup>17,18</sup> Pauline Heus,<sup>8,9</sup> Nina Kreuzberger,<sup>19</sup> Anna Lohmann,<sup>20</sup> Kim Luijken,<sup>20</sup> Jie Ma,<sup>5</sup> Glen P Martin,<sup>21</sup> Constanza L Andaur Navarro,<sup>8,9</sup> Johannes B Reitsma,<sup>8,9</sup> Jamie C Sergeant,<sup>22,23</sup> Chunhu Shi,<sup>24</sup> Nicole Skoetz,<sup>19</sup> Luc J M Smits,<sup>1</sup> Kym I E Snell,<sup>6</sup> Matthew Sperrin,<sup>25</sup> René Spijker,<sup>8,9,26</sup> Ewout W Steyerberg,<sup>3</sup> Toshihiko Takada,<sup>8</sup> Ioanna Tzoulaki,<sup>27,28</sup> Sander M J van Kuijk,<sup>29</sup> Florian S van Royen,<sup>8</sup> Jan Y Verbakel,<sup>30,31</sup> Christine Wallisch,<sup>7,32,33</sup> Jack Wilkinson,<sup>22</sup> Robert Wolff,<sup>34</sup> Lotty Hooft,<sup>8,9</sup> Karel G M Moons,<sup>8,9</sup> Maarten van Smeden<sup>8</sup>

For numbered affiliations see end of the article

Correspondence to: L Wynants  
laure.wynants@maastrichtuniversity.nl  
(ORCID 0000-0002-3037-122X)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2020;369:m1328  
<http://dx.doi.org/10.1136/bmj.m1328>

### ABSTRACT OBJECTIVE

To review and appraise the validity and usefulness of published and preprint reports of prediction models for diagnosing coronavirus disease 2019 (covid-19) in patients with suspected infection, for prognosis of patients with covid-19, and for detecting people in the general population at increased risk of becoming infected with covid-19 or being admitted to hospital with the disease.

### STUDY SELECTION

Studies that developed or validated a multivariable covid-19 related prediction model.

### DATA EXTRACTION

At least two authors independently extracted data using the CHARMS (critical appraisal and data extraction for systematic reviews of prediction modelling studies) checklist; risk of bias was assessed using PROBAST (prediction model risk of bias assessment tool)

# Aims of our review

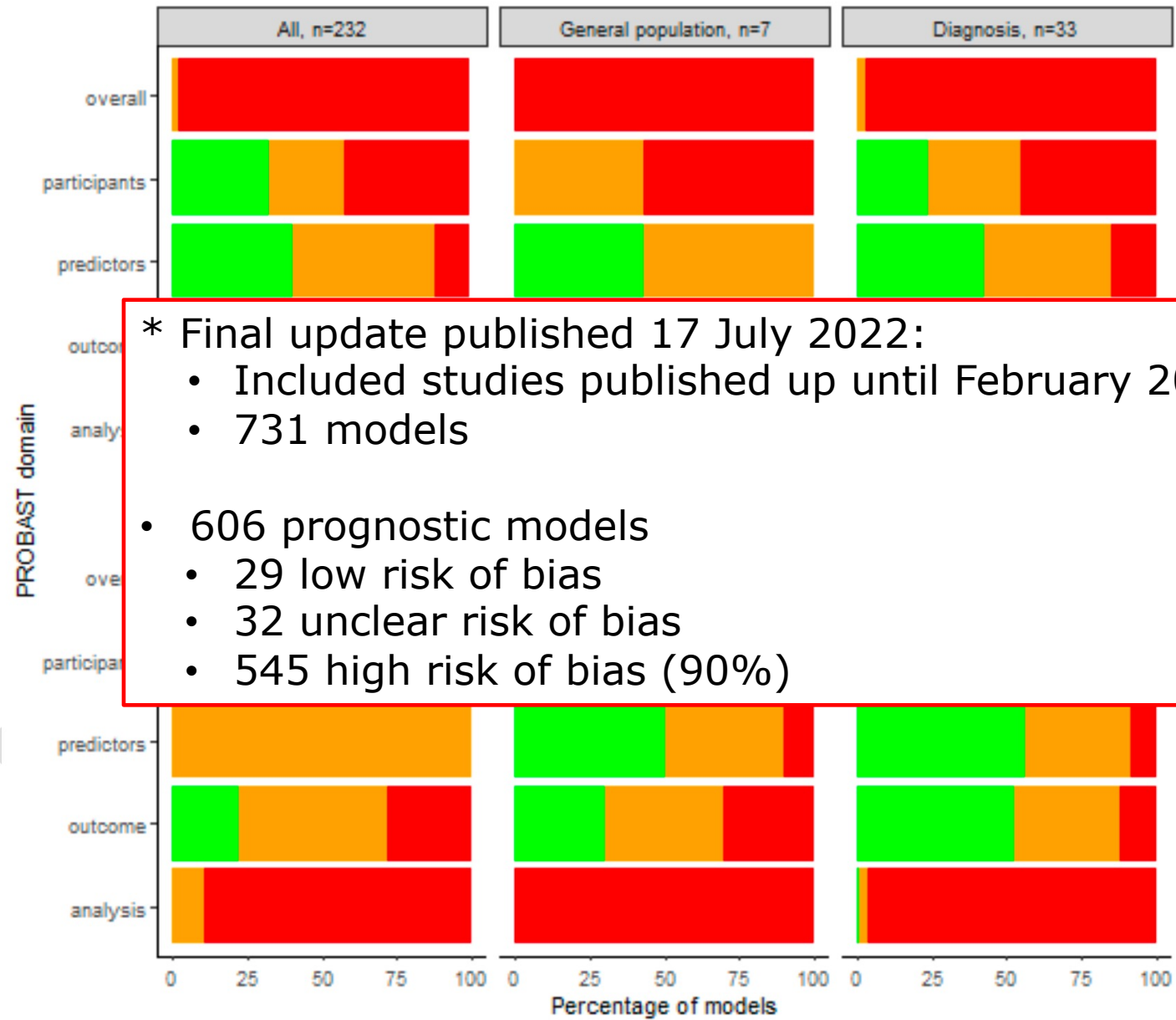
- **To review and critically appraise published reports (and preprint reports) of prediction models for**
  - Diagnosing covid-19 in patients with suspected infection
  - Prognosis of patients with covid-19 infection
  - Identifying people in general population at increased risk of infection & hospital admission

**By July 2020... 169 studies identified, proposing 232 prediction models**

# Our risk of bias (quality) summary

- **Participants domain: 98/232 (42%) at high risk of bias; 58 (25%) unclear**
  - Non-representative of the target population (e.g., non-consecutive patients)
  - e.g. no COVID19 patients included (even just simulated data)
- **Predictors domain: 15/232 (6%) at high risk of bias; 135 (35%) unclear**
  - Predictors not available at time of intended model use
- **Outcome domain: 50/232 (22%) at high risk of bias; 87 (38%) unclear**
  - Subjective or proxy outcomes
  - Predictors part of the outcome definition
- **Analysis domain: 218/232 (94%) at high risk of bias; 13 (6%) unclear**
  - Small sample size (->overfitting & no adjustment), dichotomisation of continuous predictors, incomplete reporting of model performance (e.g., no calibration), not accounting for censoring, no external validation

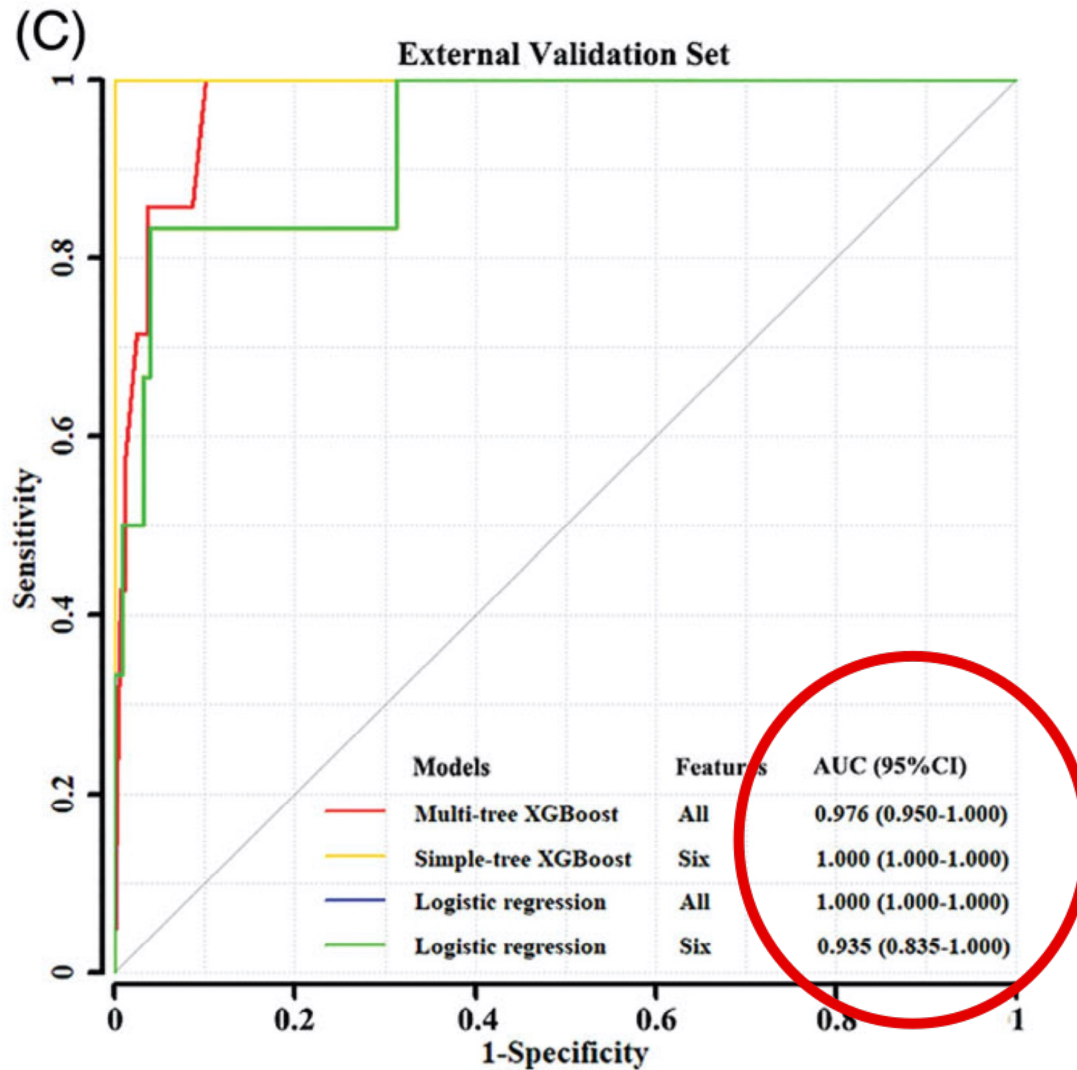




\* Final update published 17 July 2022:

- Included studies published up until February 2021
- 731 models
- 606 prognostic models
  - 29 low risk of bias
  - 32 unclear risk of bias
  - 545 high risk of bias (90%)

# Example: Guan et al. (2021)



## Prognostic model for risk of death from covid19

“Simple-tree XGBoost model conducted by these features can predict death risk accurately”

## Sample size

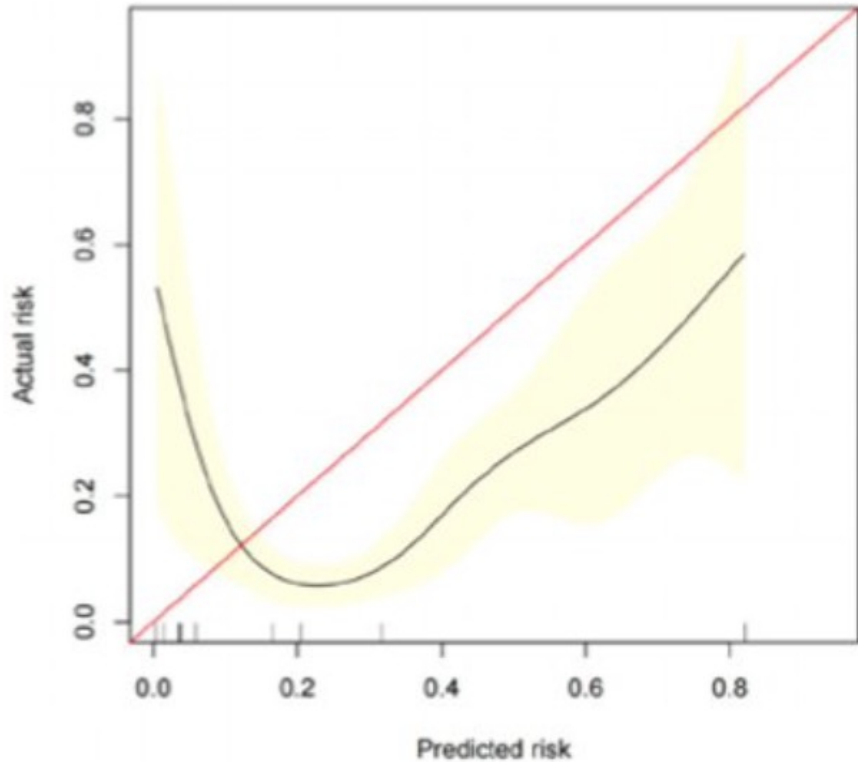
Internal: 217 participants (16 events)

External: 279 participants (7 events)

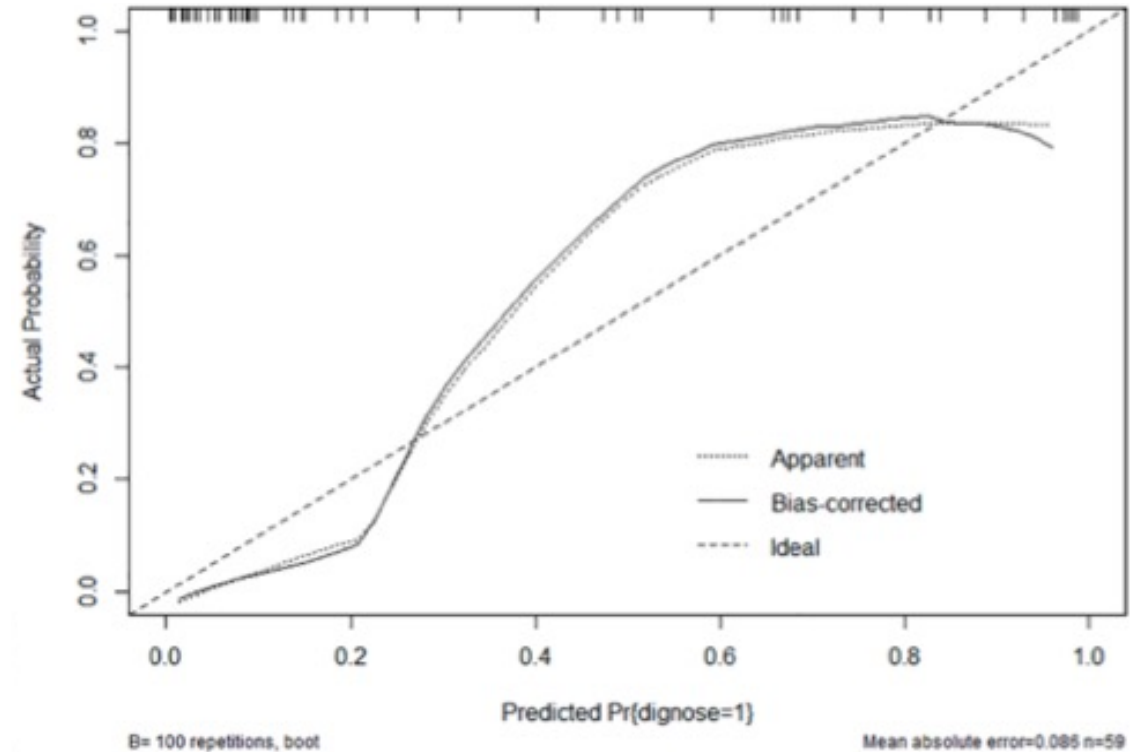
## No calibration checks

\*Guan et al, Ann Med 2021

# Miscalibration & spin



“The calibration curve showed a good agreement between the predictive risk and the actual probability”



“Good calibration”  
“Hosmer-Lemeshow Test: p-value = 1.0”

# A good prediction model study

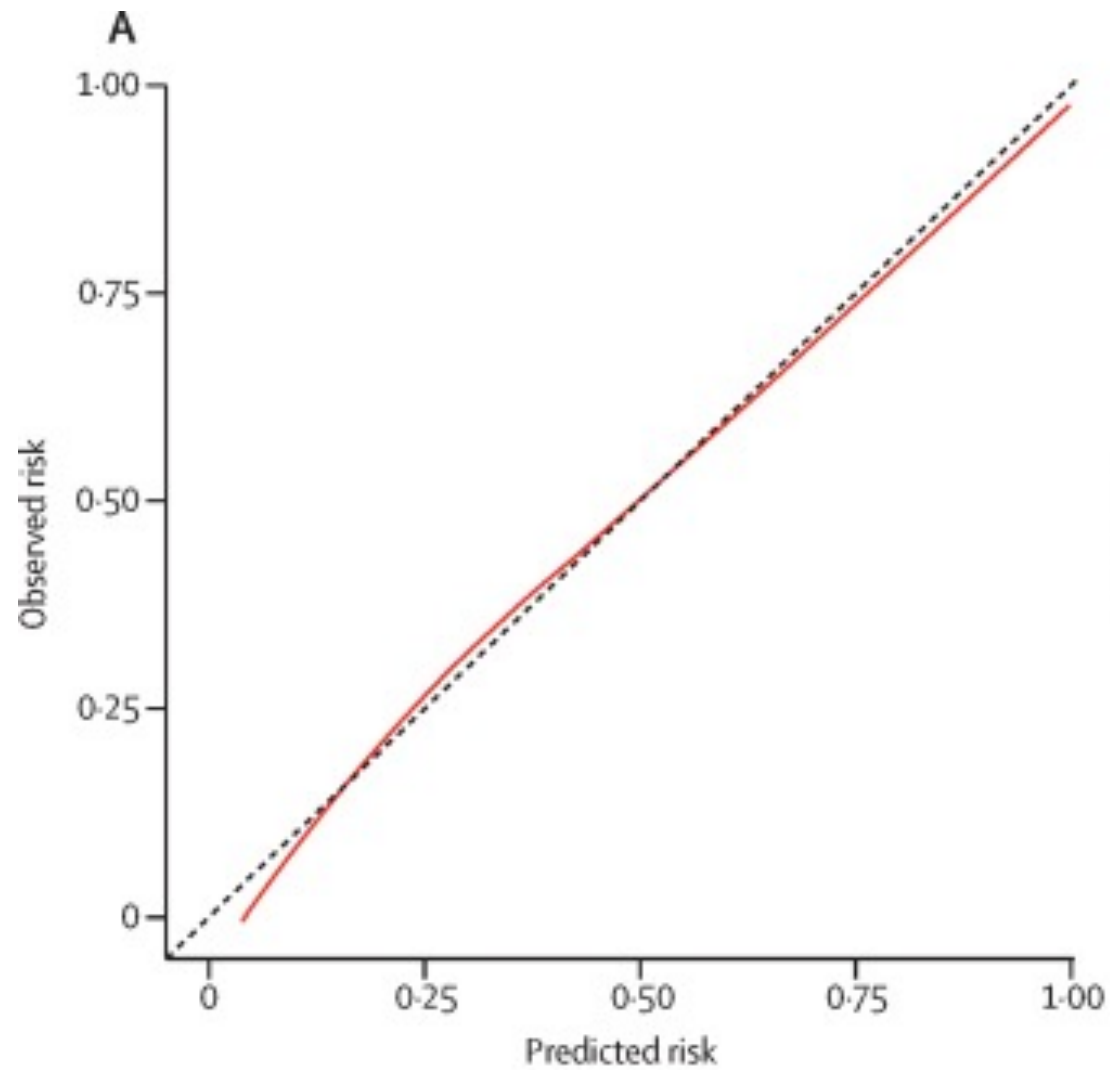
---

## Development and validation of the ISARIC 4C Deterioration model for adults hospitalised with COVID-19: a prospective cohort study

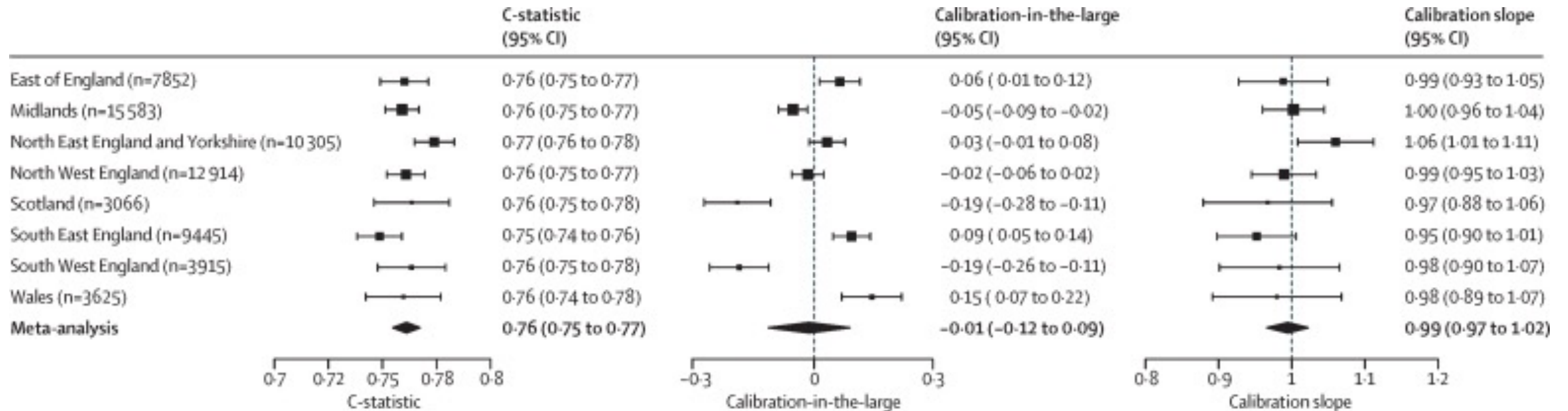
*Rishi K Gupta, Ewen M Harrison, Antonia Ho, Annemarie B Docherty, Stephen R Knight, Maarten van Smeden, Ibrahim Abubakar, Marc Lipman, Matteo Quartagno, Riinu Pius, Iain Buchan, Gail Carson, Thomas M Drake, Jake Dunning, Cameron J Fairfield, Carrol Gamble, Christopher A Green, Sophie Halpin, Hayley E Hardwick, Karl A Holden, Peter W Horby, Clare Jackson, Kenneth A Mclean, Laura Merson, Jonathan S Nguyen-Van-Tam, Lisa Norman, Piero L Olliaro, Mark G Pritchard, Clark D Russell, James Scott-Brown, Catherine A Shaw, Aziz Sheikh, Tom Solomon, Cathie Sudlow, Olivia V Swann, Lance Turtle, Peter J M Openshaw\*, J Kenneth Baillie\*, Malcolm G Semple\*, Mahdad Noursadeghi\*, on behalf of the ISARIC4C Investigators*

- Extensive internal and external validation in very large samples
- Assessed discrimination, calibration & clinical utility
- Performance showed good generalisability

# A good prediction model study



# A good prediction model study



Part 3

# HOW CAN WE DO BETTER?

# Hard to stop making predictions

- *Paul Gascoigne (footballer):*

“I never make predictions & I never will”



# Hard to stop making predictions

- ***Paul Gascoigne (footballer):***

“I never make predictions & I never will”

- ***Andrea Leadsom (MP), 5<sup>th</sup> December 2018***

“I have never, and will not, start predicting the future... I don't do predictions ever”

# Hard to stop making predictions

- ***Paul Gascoigne (footballer):***

“I never make predictions & I never will”

- ***Andrea Leadsom (MP), 5<sup>th</sup> December 2018***

“I have never, and will not, start predicting the future... I don't do predictions ever”  
*(a few hours later ...)*

“I am a very strong arch Brexiteer, I genuinely believe that we have a bright future ahead of us when we leave the EU”

**So, if we are to keep making predictions, let's improve our methodology standards ...**

# We must do better

## Challenge for us all:

### - strive for better prediction model research

- Register projects, e.g. *clinicaltrials.gov*
- Publish protocols, e.g. *Diagnostic & Prognostic Research*
- Validate existing models (no need for a new model?)
- **Include statisticians & health data experts from outset**
- **Work with clinical experts to understand why the model is needed**
- Clearly report your project methods & findings

# TRIPOD reporting guideline

- **“Good reporting is not an optional extra; it is an essential component of research”** - Altman et al. Open Med 2008
- **TRIPOD: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis**
- 32 items covering 22 ‘topics’ for model development & validation
- Soon to be updated to TRIPOD+AI
- [www.tripod-statement.org](http://www.tripod-statement.org)
  - includes extensions to systematic reviews, clusters, protocols, ...

# Do not dichotomise continuous predictors

## Dichotomisation is biologically implausible

- e.g. dichotomise age into two groups:  $<65$ , or  $\geq 65$ 
  - Individuals aged 64 and 65 considered different
  - Individuals aged 23 and 64 considered the same

## Dichotomisation leads to worse performance & data-dredging

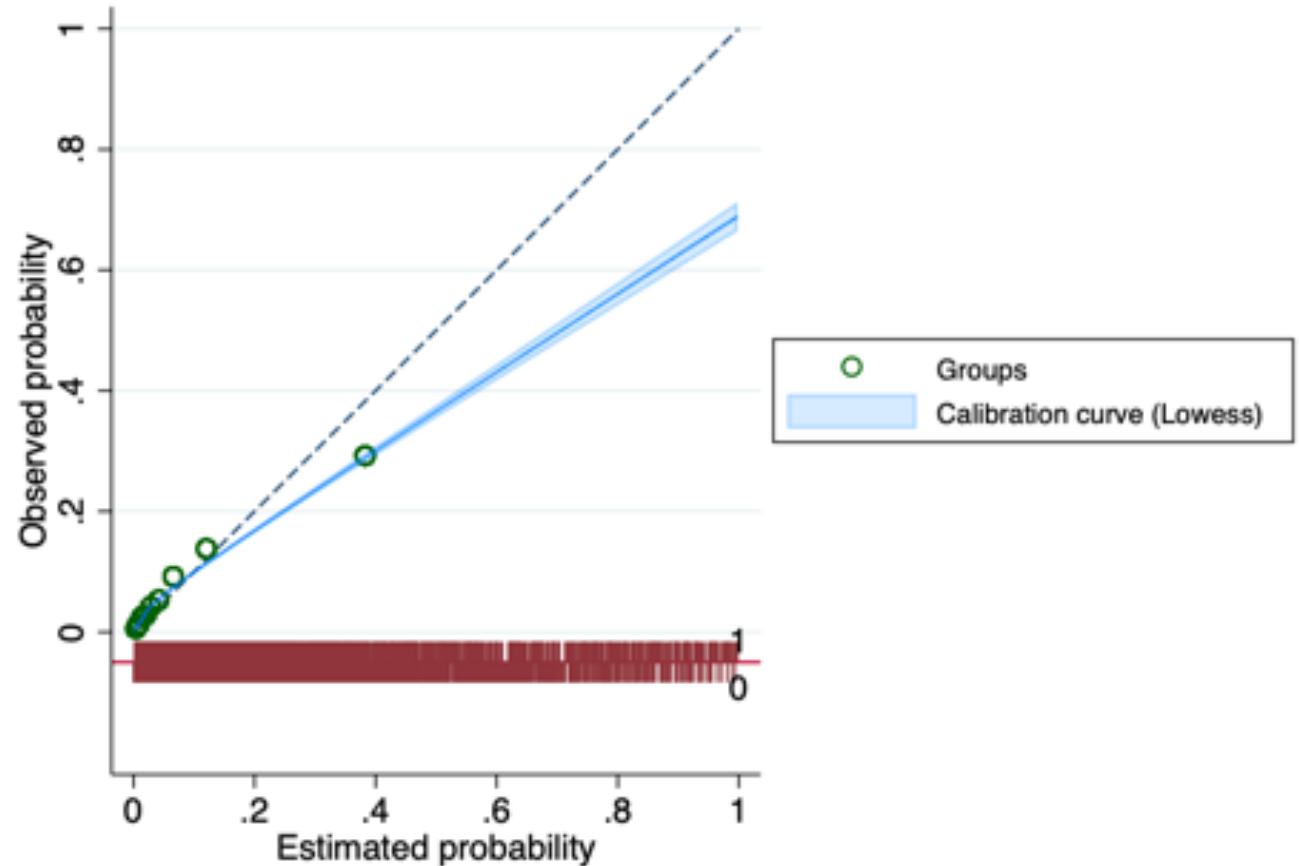
- e.g. selection of 'optimal' cut-points to maximise statistical significance

## Rather model non-linear relationships (e.g., splines, polynomials)

Thresholds for decision making can be defined **AFTER** analysis on a relevant scale (e.g. based on predicted risk)

# Beyond calibration & discrimination

- Example: Prediction of 30-day mortality in acute MI patients
- Calibration slope = 0.72
- Calibration is not perfect but miscalibration mainly in areas above risks of about 15-20% i.e. driven by those with the highest risk (where the under-prediction of risks may not matter)
- Model could still be clinically useful

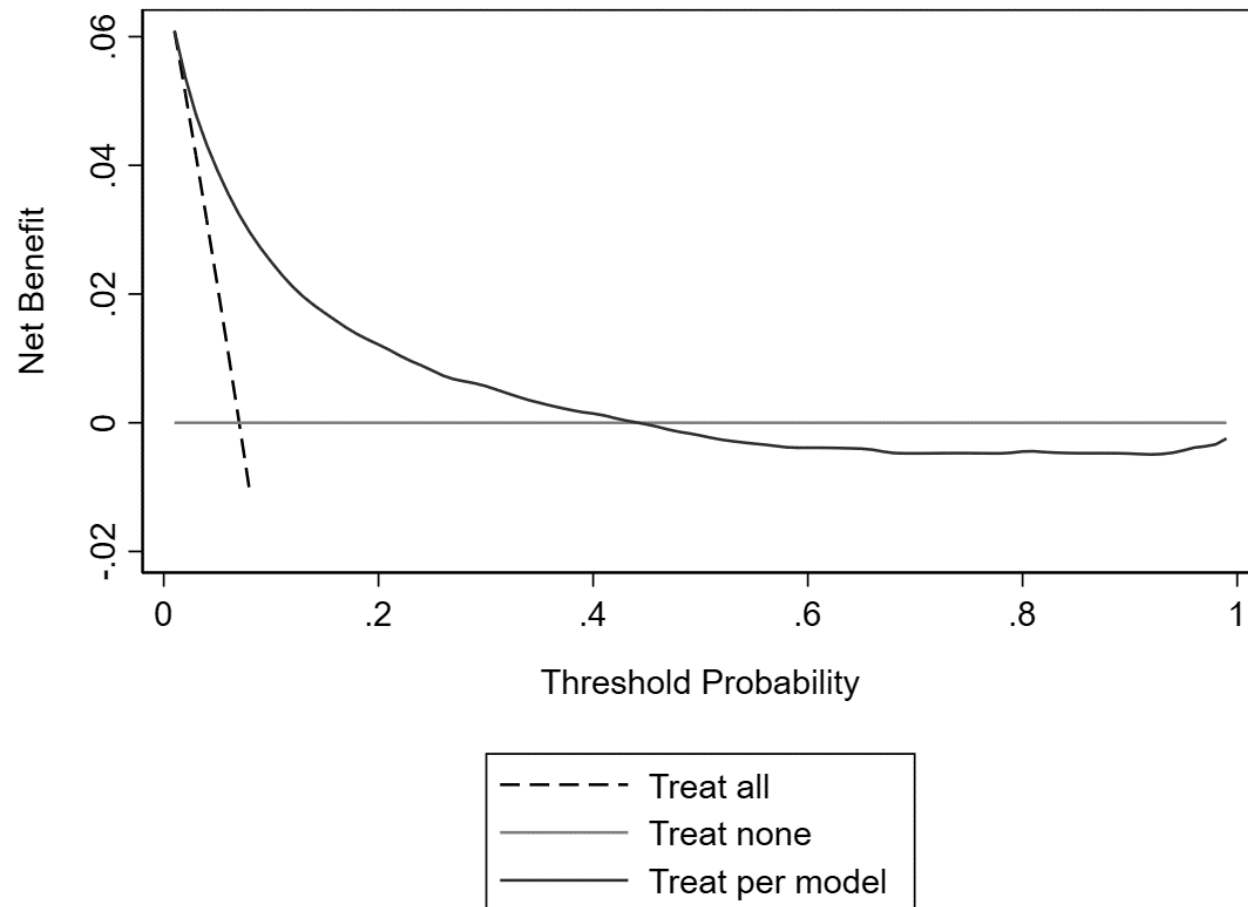


# Evaluating clinical impact

- Cost-effectiveness modelling
  - simulate patients and pathways, conditional on model predictions
  - are outcomes improved and process cost-effective?
- Randomised trial
  - one group uses the model (+ usual care); other group uses usual care only
  - are patient outcomes improved?
- Decision Analysis and evaluation of 'net benefit' (clinical utility)
  - weigh benefits (improved outcomes) vs. harms (worse outcomes, costs)
  - depends on potential risk thresholds (identified in advance of analysis)  
e.g. 10% threshold: willing to 'treat' 10 individuals so that 1 benefits

# Revisit the acute MI model

- Despite miscalibration, still potential clinical utility
- Very dependent on the (range of) thresholds deemed relevant





# Aim to develop a stable model

- Models more reliable & stable when developed using
  - large sample sizes representative of target population
  - appropriate no. predictors relative to no. events
  - approaches to ‘address’ overfitting (e.g. lasso, ensemble methods)
  - resampling (e.g. cross-validation, bootstrapping) to examine/adjust optimism
- **Concerns of an unreliable model exposed by examining *model instability***
  - bootstrapping: develop multiple models & see how predictions change
  - you may be shocked what you find ...

Riley & Collins, 2023 – Biometrical Journal

Is your model stable? Most models are like ...



# Instability

- **What do we mean by instability?**
  - Idea that your developed model (e.g. regression, forest) may be different if it were developed again in exactly same way in a different sample of same size from same population  
e.g. different intercept estimate, different selected predictors, different trees & predictor effects

# Instability

- **What do we mean by instability?**
  - Idea that your developed model (e.g. regression, forest) may be different if it were developed again in exactly same way in a different sample of same size from same population  
e.g. different intercept estimate, different selected predictors, different trees & predictor effects
- **Instability in a model leads to instability in predictions**
  - Predictions from your model are different to predictions from another (hypothetical) model
  - e.g. Sam obtains an estimated risk of 0.2 in your model, but 0.7 in another model

# Instability

- **What do we mean by instability?**
  - Idea that your developed model (e.g. regression, forest) may be different if it were developed again in exactly same way in a different sample of same size from same population  
e.g. different intercept estimate, different selected predictors, different trees & predictor effects
- **Instability in a model leads to instability in predictions**
  - Predictions from your model are different to predictions from another (hypothetical) model
  - e.g. Sam obtains an estimated risk of 0.2 in your model, but 0.7 in another model
- **The larger the instability concern, the greater the threat a model is unreliable**
- Large instability => poor internal validity (in the development population)
- We should always examine & report instability after developing our models ...

# Quantifying instability using bootstrapping

- **Use bootstrapping with replacement** (i.e. resample from the model development data)
- Generate 1000 bootstrap samples, each of same size as original dataset. Then ...
  - 1) in each bootstrap sample, develop new model using same model development steps
    - this produces 1000 bootstrap models

# Quantifying instability using bootstrapping

- Use bootstrapping with replacement (i.e. resample from the model development data)
- Generate 1000 bootstrap samples, each of same size as original dataset. Then ...
  - 1) in each bootstrap sample, develop new model using same model development steps
    - this produces 1000 bootstrap models
  - 2) in original sample, calculate predictions for each individual for each bootstrap model
    - leads to 1000 predicted values (estimated risks) for each individual



# Quantifying instability using bootstrapping

- Use bootstrapping with replacement (i.e. resample from the model development data)
- Generate 1000 bootstrap samples, each of same size as original dataset. Then ...
  - 1) in each bootstrap sample, develop new model using same model development steps
    - this produces 1000 bootstrap models
  - 2) in original sample, calculate predictions for each individual for each bootstrap model
    - leads to 1000 predicted values (estimated risks) for each individual
  - 3) present a “[prediction instability plot](#)”
    - bootstrap model predictions (y-axis) vs. original model prediction (x-axis).



# Quantifying instability using bootstrapping

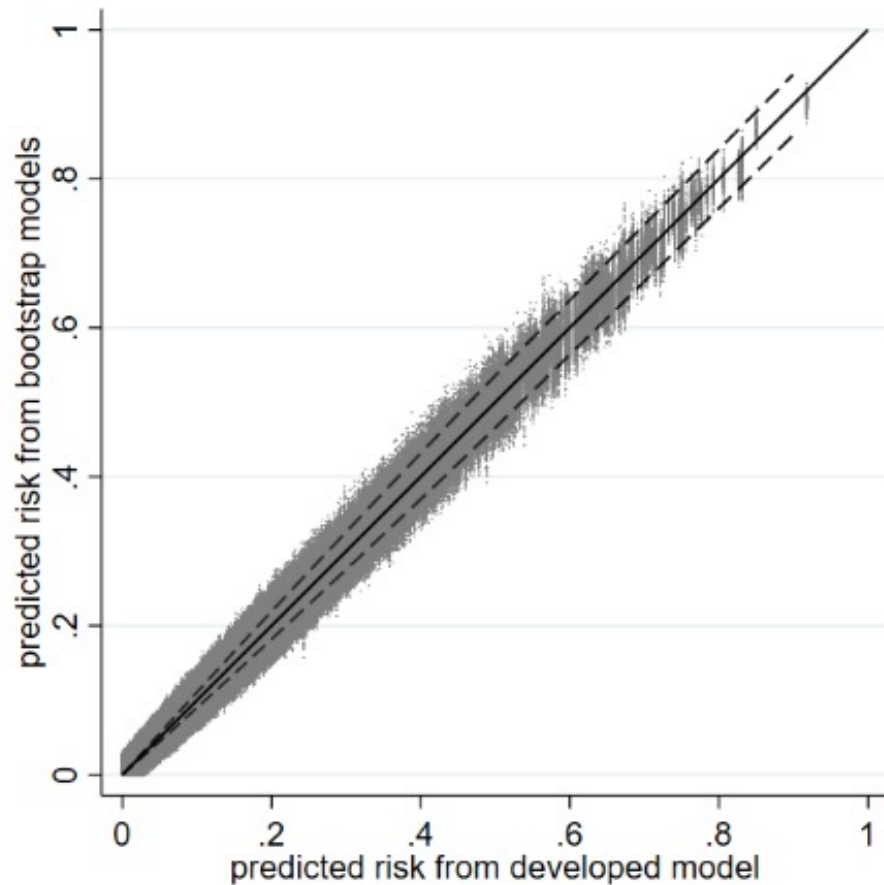
- Use bootstrapping with replacement (i.e. resample from the model development data)
- Generate 1000 bootstrap samples, each of same size as original dataset. Then ...
  - 1) in each bootstrap sample, develop new model using same model development steps
    - this produces 1000 bootstrap models
  - 2) in original sample, calculate predictions for each individual for each bootstrap model
    - leads to 1000 predicted values (estimated risks) for each individual
  - 3) present a “[prediction instability plot](#)”
    - bootstrap model predictions (y-axis) vs. original model prediction (x-axis).
  - 4) present other measures, such as
    - “[classification instability plot](#)” and “[calibration instability plot](#)”
    - MAPE: mean absolute difference between original and bootstrap model predictions

# Real example

- Develop a prediction model for risk of death by 30 days after acute myocardial infarction
- Use GUSTO-I dataset (freely available - acknowledge Duke Clinical Research Institute)
- In full dataset: 40830 participants & 2851 deaths by 30 days
- Overall risk is about 7%
- Eight predictors are of interest:
  - Sex, Age, Hypertension, Hypotension, Tachycardia, Previous Myocardial Infarction, ST Elevation on ECG, and systolic blood pressure.
- A lasso logistic regression fitted to the full dataset gives C-statistic of 0.80
- Let's apply bootstrapping to examine instability of this model ...

# Example 1: lasso logistic regression

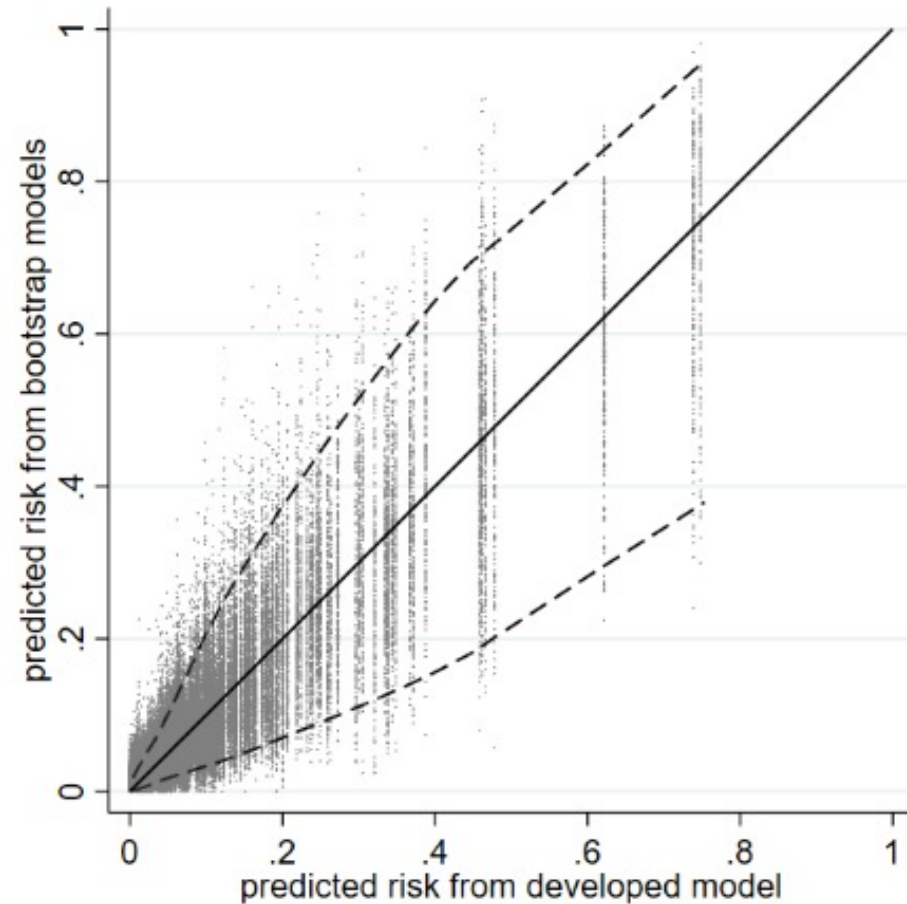
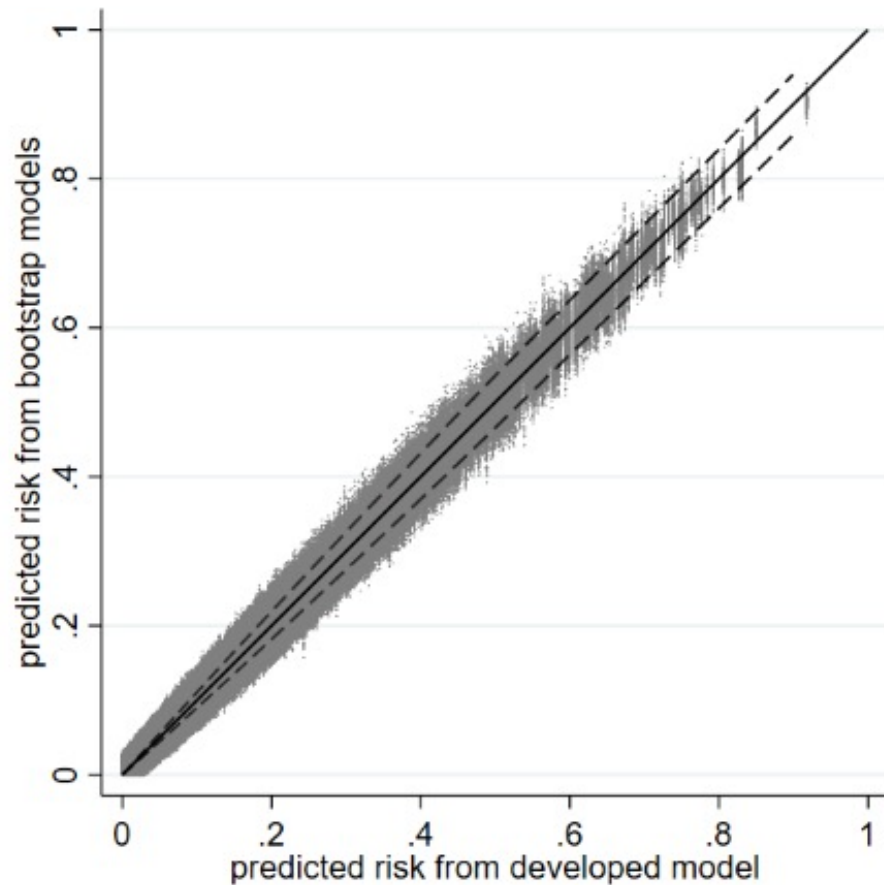
- **FULL**: 40,830 patients, 2851 events, 407 events per predictor
- Average MAPE = 0.0027 (largest 0.027)



# Example 1: lasso logistic regression

- **FULL**: 40,830 patients, 2851 events, 407 events per predictor
- Average MAPE = 0.0027 (largest 0.027)

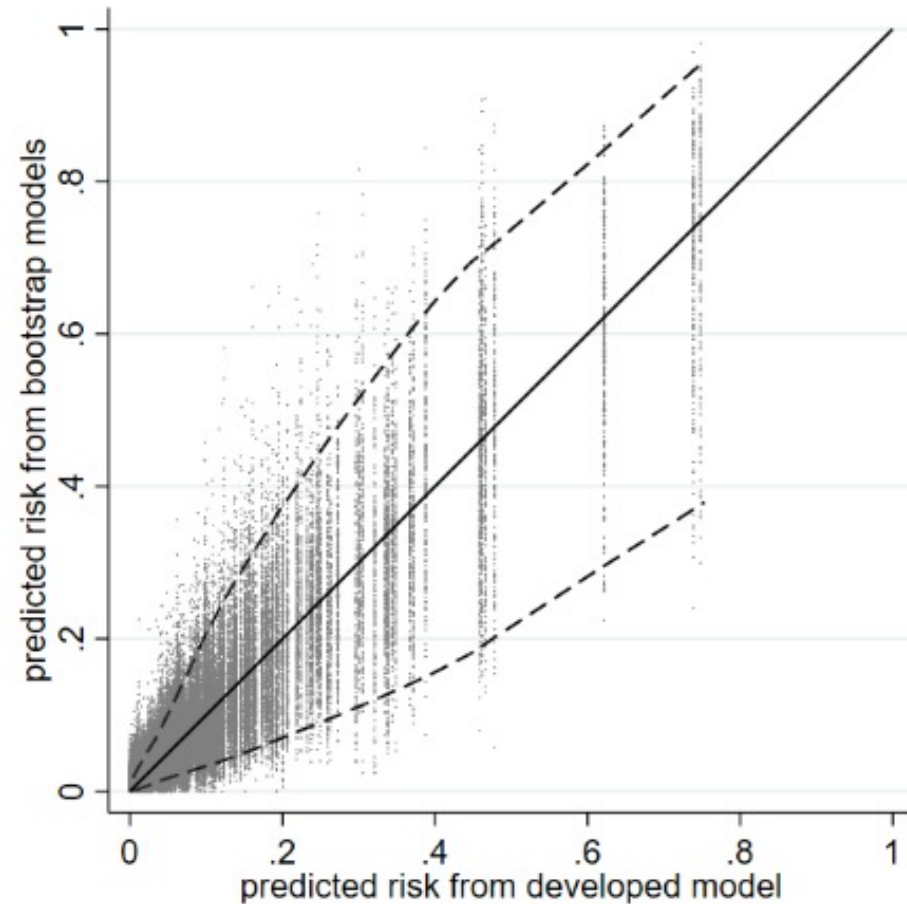
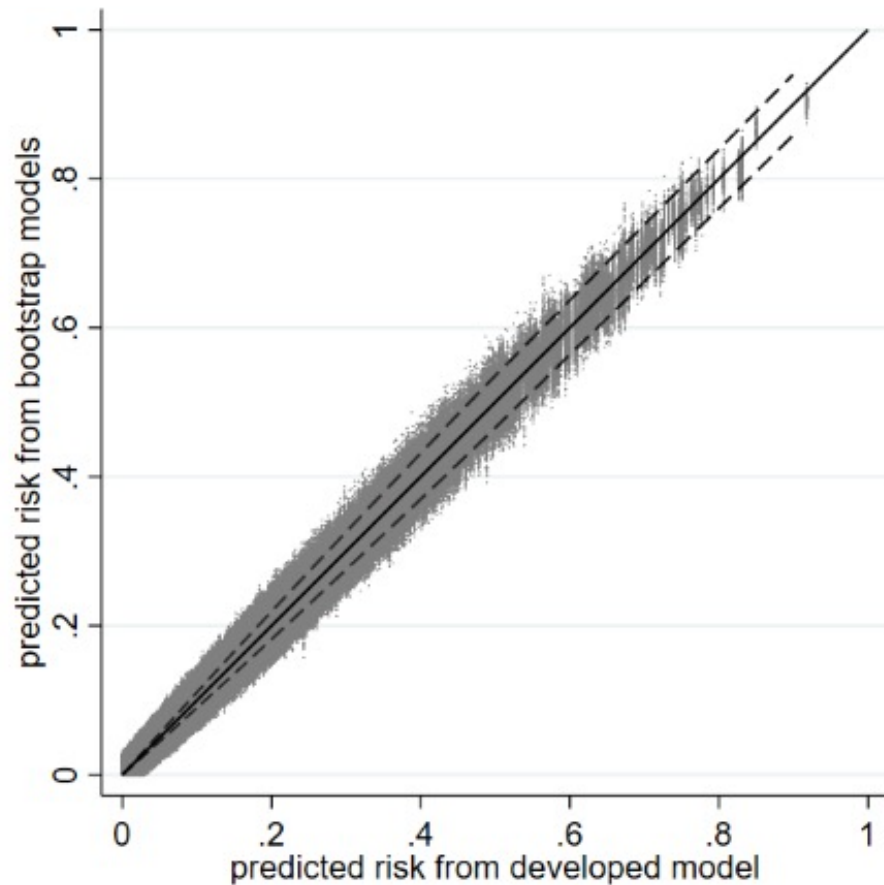
- **SMALL**: 500 patients, 35 events, 4 events per predictor
- Average MAPE = 0.023 (largest 0.14)



# Example 1: lasso logistic regression

- **FULL**: 40,830 patients, 2851 events, 407 events per predictor
- Average MAPE = 0.0027 (largest 0.027)

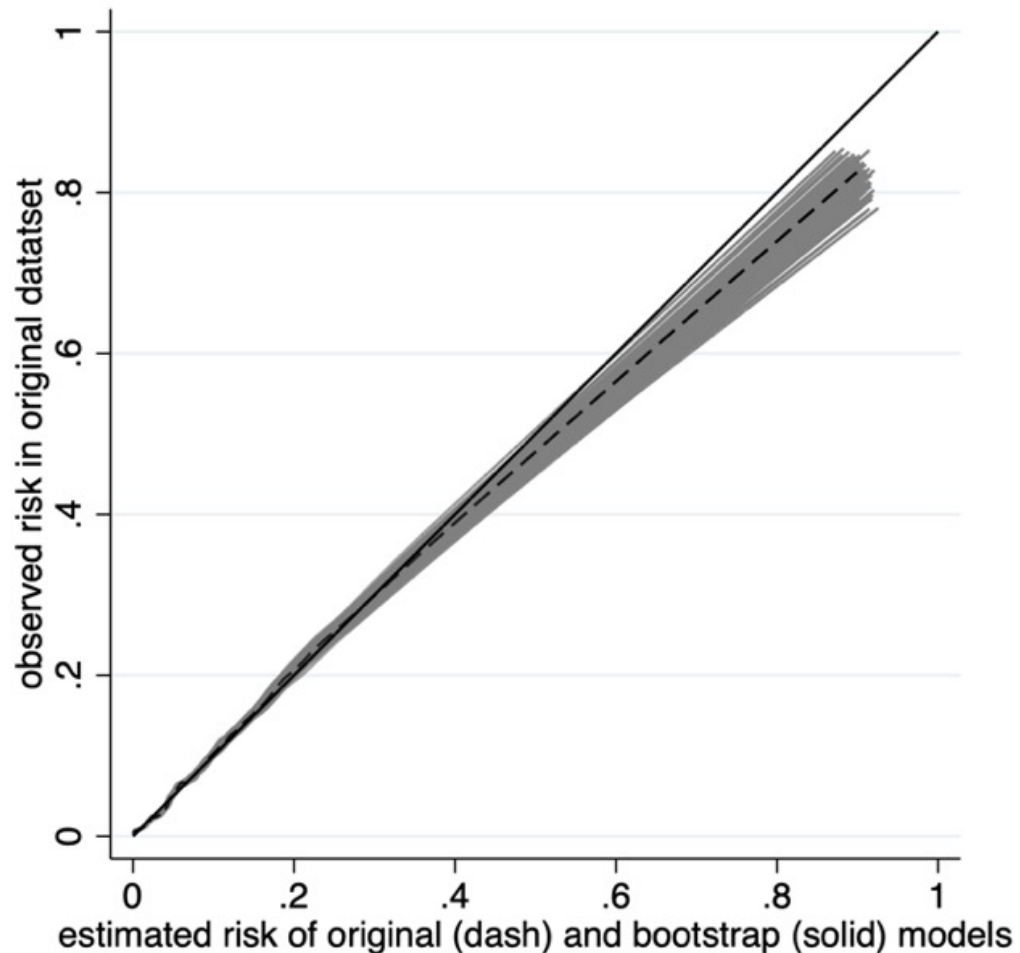
- **SMALL**: 500 patients, 35 events, 4 events per predictor
- **C-STATISTIC RANGES FROM 0.77 to 0.83**



# Example 1: lasso logistic regression

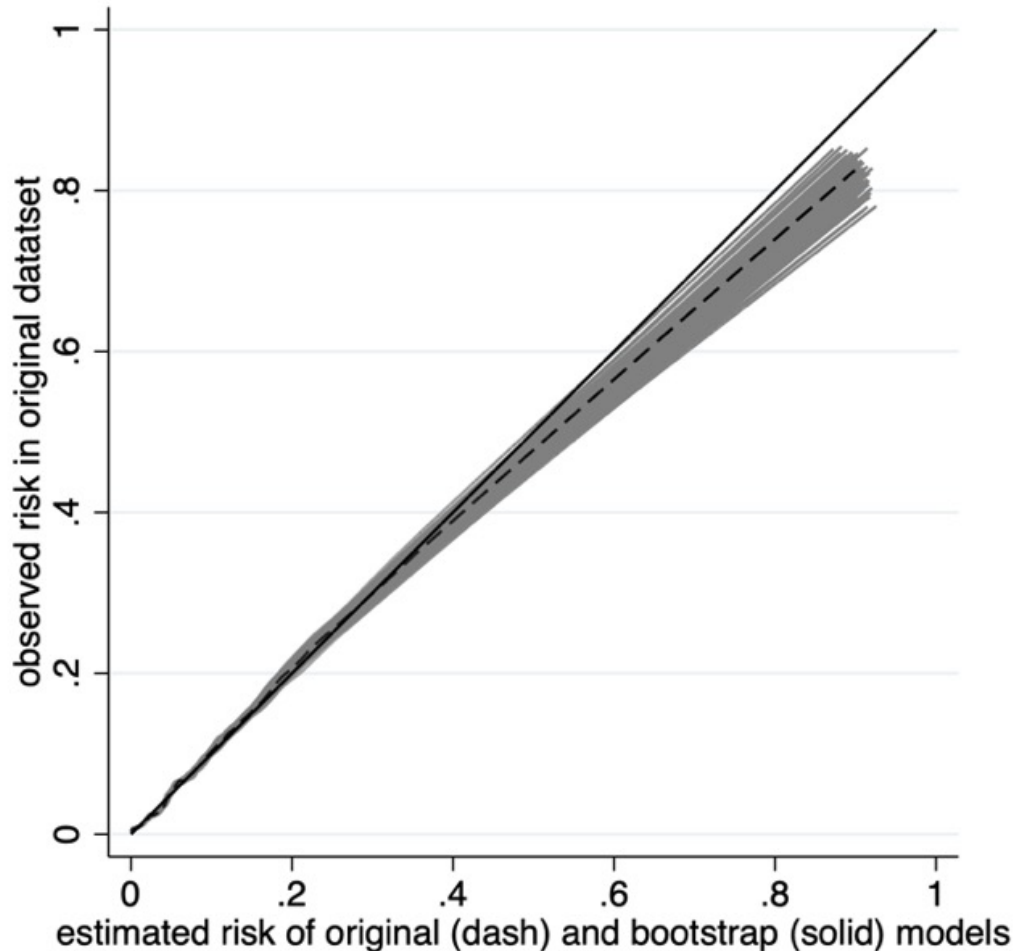
- **FULL**: 40,830 patients, 2851 events, 407 events per predictor
- CALIBRATION INSTABILITY

- **SMALL**: 500 patients, 35 events, 4 events per predictor
- CALIBRATION INSTABILITY

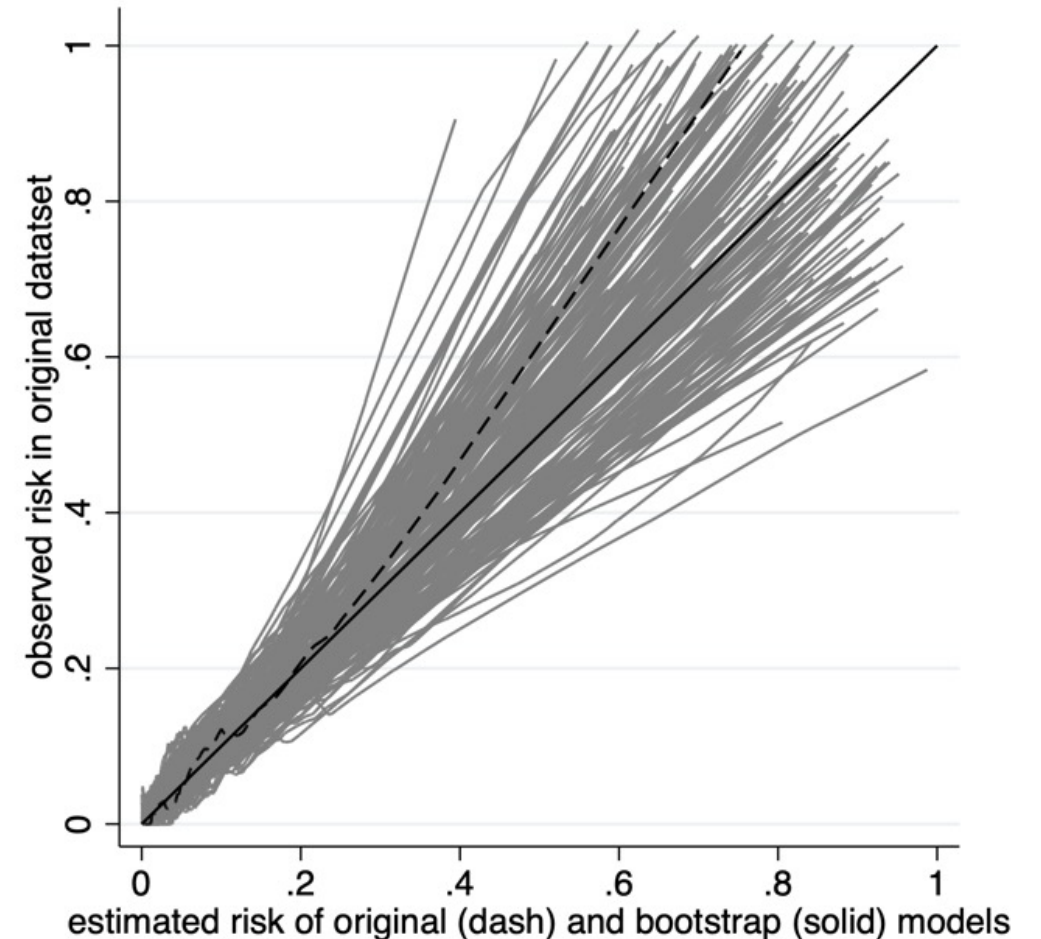


# Example 1: lasso logistic regression

- **FULL**: 40,830 patients, 2851 events, 407 events per predictor
- CALIBRATION INSTABILITY



- **SMALL**: 500 patients, 35 events, 4 events per predictor
- CALIBRATION INSTABILITY



# Hang on ... don't AI methods resolve this?

- Lots of work to improve stability of models
- **Modern methods aim to reduce variance in the bias-variance trade off**
  - e.g. repeated cross-validation to estimate penalty factors in penalized regression
- Machine learning (AI) focuses on **ensemble methods** and **super learners**
  - these approaches aggregate predictions over many models
  - recommended to 'address' instability concerns & improve upon single model

e.g. Random forest is a popular ensemble method

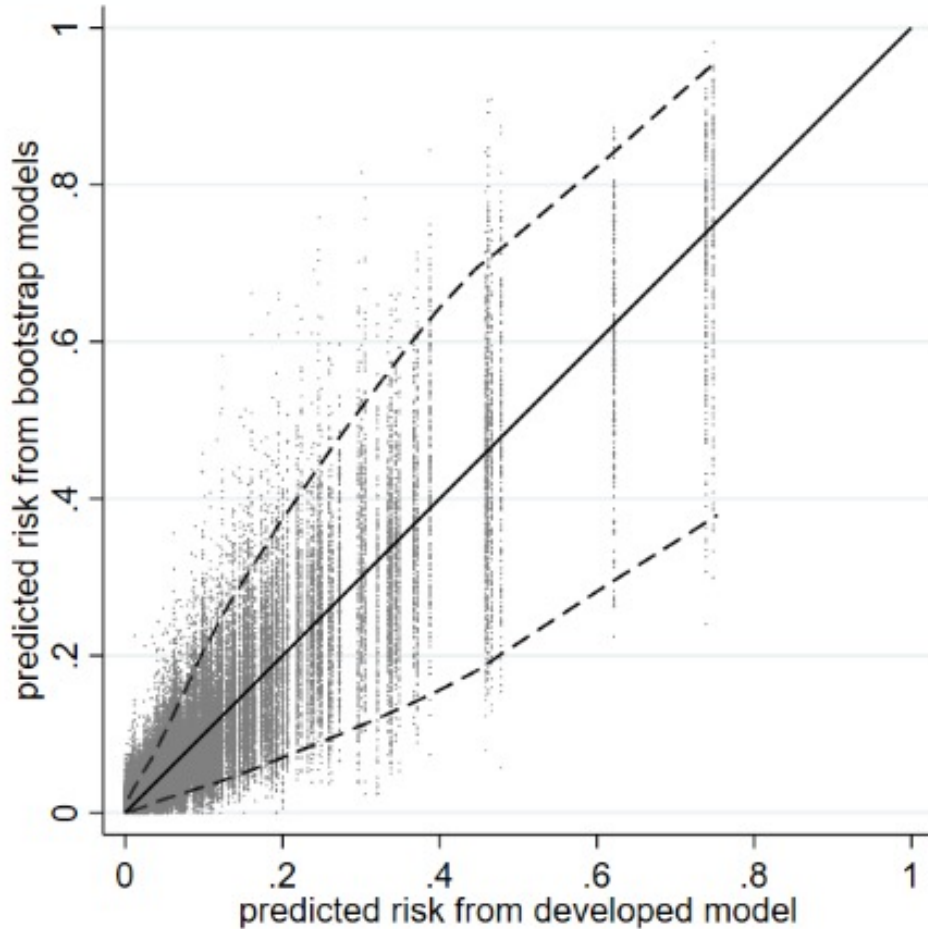
- Uses BAGGING (bootstrap aggregating) to generate predictions over multiple models
- Includes a random element to feature selection, to limit the correlation across models
- Let's see how random forest performs here ...



# Example 2: lasso vs. random forests

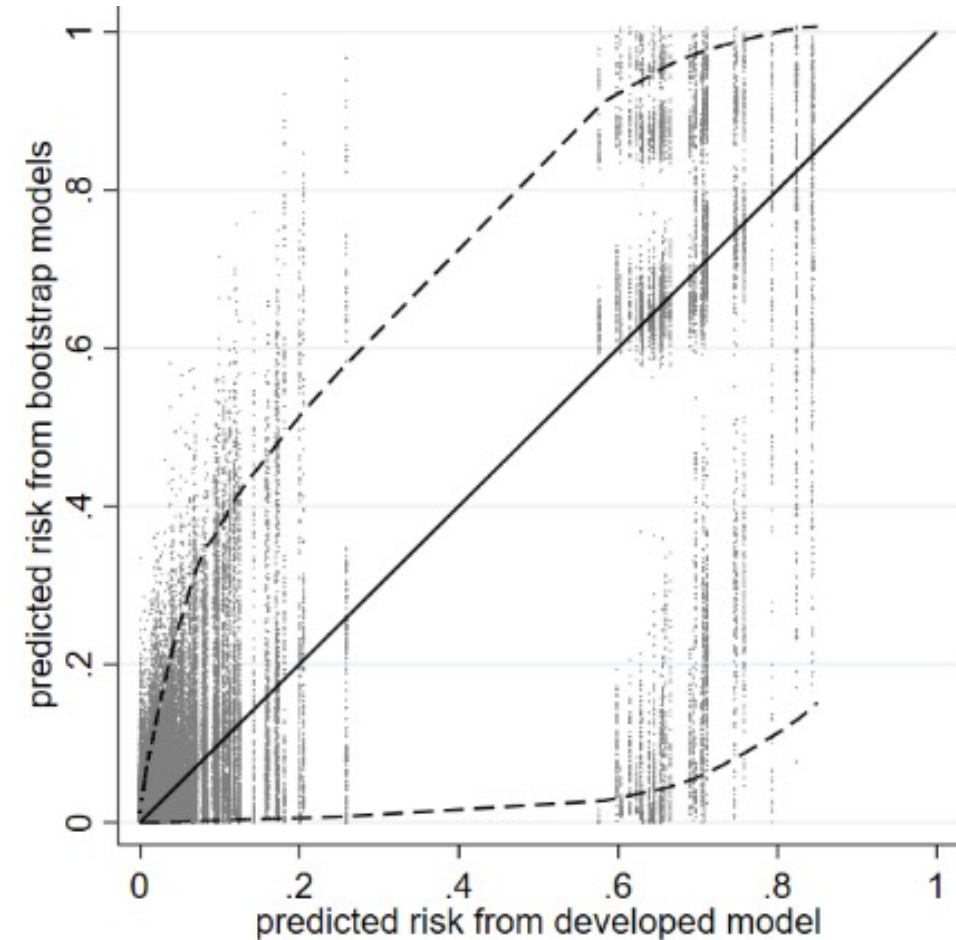
- SMALL: 500 patients, 35 events, 4 events per predictor

**LASSO logistic regression:**



- SMALL: 500 patients, 35 events, 4 events per predictor

**RANDOM FOREST: (100 trees – default settings):**



# ASIDE: ML versus statistical methods

- This is not the debate!
- **Rather: identify the right method to answer the right research question**
- Machine learning methods have much potential, but
  - usually require a (much) larger sample size for stability
  - black-box aspect concerning for transparent, shared decision making
- Thus, consider stability and transparency when choosing development method
- Explainable AI and fairness checks futile when there is instability!

# A note on sample size for model development

- **Without a decent sample size, you're in trouble**
- We proposed guidance for (penalized) regression approaches – see refs at end
  - Target precise estimation of the overall risk (or mean value)
  - Target small MAPE (mean absolute prediction error)
  - Target small amount of overfitting (e.g. shrinkage of  $< 10\%$ )
- Available in the **pmsampsize** package for R or Stata
- Focused on (penalized) regression models – but still relevant for machine learning
- Provides 'minimum' required
  - still check stability though (often not small)

# Stata module by Ensor: PMSAMPSIZE

**Binary outcome example: Cox-Snell R2 0.2, outcome 50%, p = 30**

```
. pmsampsize, type(b) rsquared(0.2) parameters(30) prev(0.5)
```

	Samp_size	Shrinkage	Parameter	Rsqr	Max_Rsqr	EPP
<b>Criteria 1</b>	1194	.9	30	.2	.75	<b>19.9</b>
Criteria 2	701	.842	30	.2	.75	11.68
Criteria 3	385	.9	30	.2	.75	6.42
Final	1194	.9	30	.2	.75	19.9

**Minimum sample size required = 1194, with 597 events**

**EPP = 19.9**

# Stata module by Ensor: PMSAMPSIZE

Binary outcome example: Cox-Snell **R2 0.5**, outcome 50%, p = 30

```
. pmsampsize, type(b) rsquared(0.5) parameters(30) prev(0.5)
```

	Samp_size	Shrinkage	Parameter	Rsq	Max_Rsq	EPP
Criteria 1	370	.9	30	.5	.75	6.17
<b>Criteria 2</b>	556	.93	30	.5	.75	<b>9.27</b>
Criteria 3	385	.93	30	.5	.75	6.42
Final	556	.93	30	.5	.75	9.27

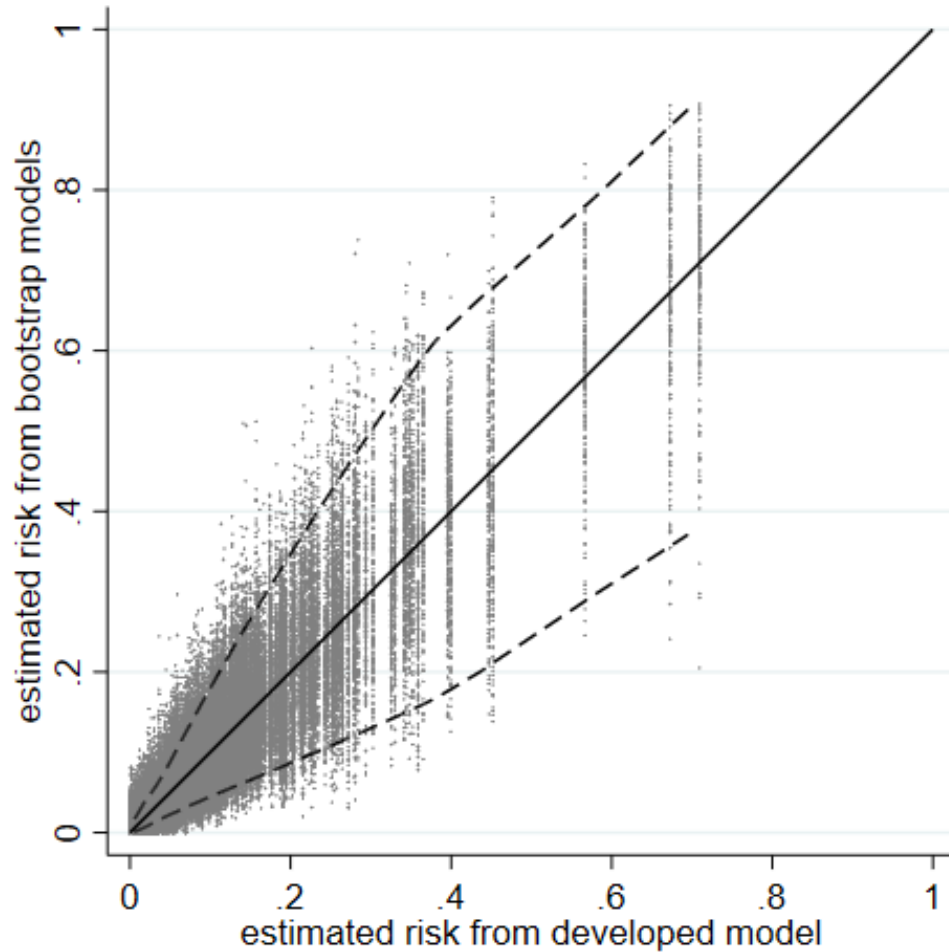
Minimum sample size required = 556, with 278 events

**EPP = 9.27**

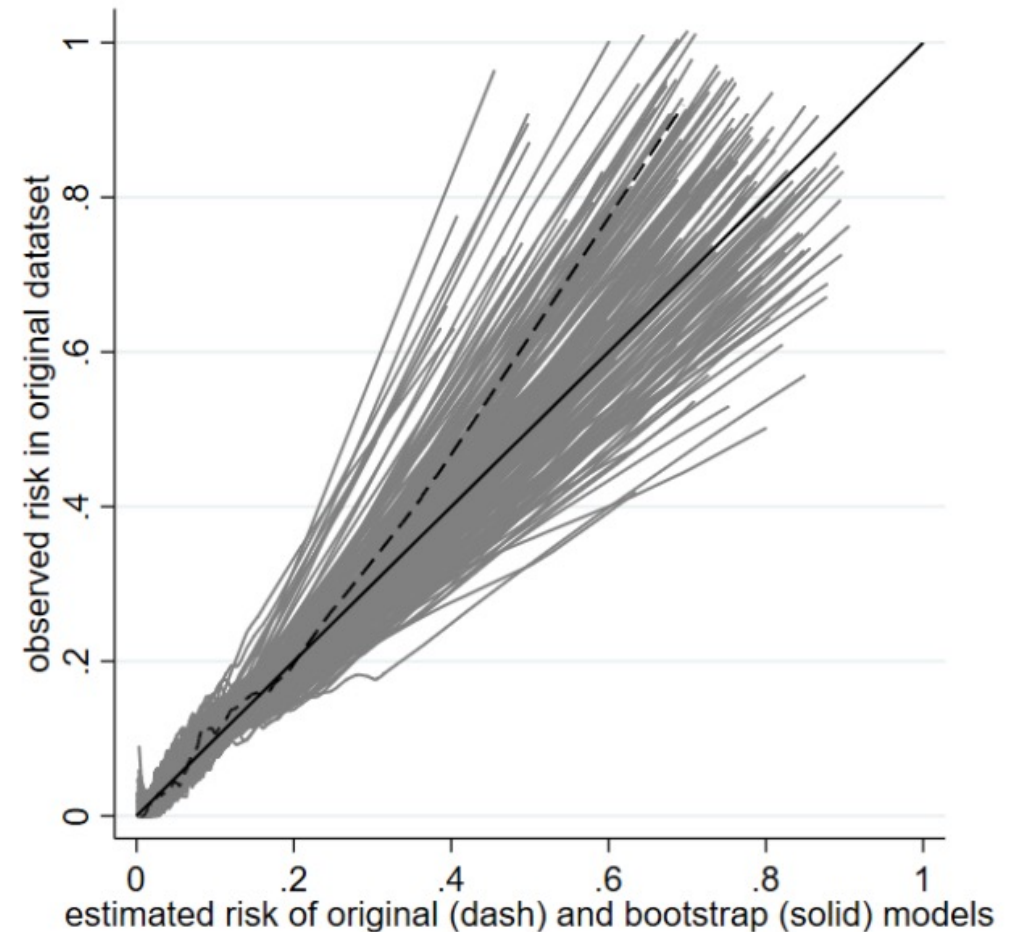
# Example using minimum sample size: lasso

- MINIMUM: 752 participants (53 events), 7 predictors

## PREDICTION INSTABILITY PLOT



## CALIBRATION INSTABILITY PLOT



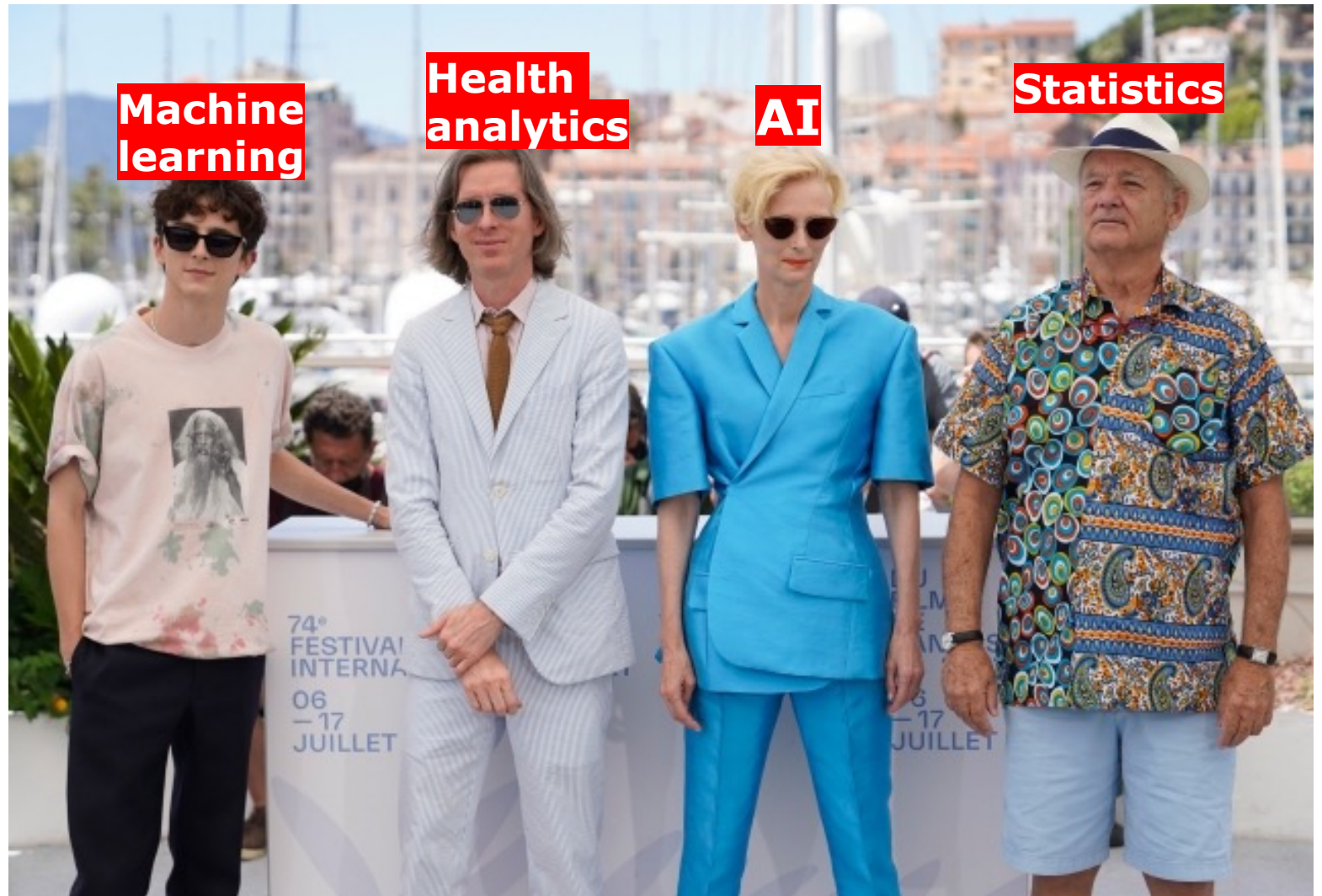
# A note on sample size for model validation

- For external validation, the focus is on estimation of predictive performance (e.g. calibration, discrimination & clinical utility)
- **Thus, minimum sample size should target precise estimates of performance**
- We proposed guidance for continuous, binary and survival outcomes
- User must input
  - linear predictor distribution
  - outcome proportion (mean outcome value)
  - target confidence interval widths
  - calibration performance (e.g. slope & O/E = 1)
- **pmvalsampsize (just released!)**



# Summary

- A diverse set of researchers are working on prediction models in healthcare
- Current standards very poor
- Regardless of analytic skills & background, we need to
  - be better trained
  - educate others
  - enforce high standards
  - aim for stable models
  - consider sample size
  - focus more on validation
  - target clinical impact





## SELECTED REFERENCES: [r.d.riley@bham.ac.uk](mailto:r.d.riley@bham.ac.uk) @Richard\_D\_Riley

### STABILITY

- Riley RD, Collins GS. Quantifying instability of a clinical prediction model developed using statistical or machine learning methods. *Biometrical Journal* 2023; 65: e2200302
- Pate A, Emsley R, Sperrin M, et al. Impact of sample size on the stability of risk scores from clinical prediction models: a case study in cardiovascular disease. *Diag Prog Res* 2020;4:14

### SAMPLE SIZE

- Riley RD, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;368:m441.
- Riley RD, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part II - binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276-96.
- Riley RD, et al. Minimum sample size for ... Part I - Continuous outcomes. *Stat Med*. 2019;38(7):1262-75
- Archer L, Snell KIE, Ensor J, et al. 2020. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. ***Stat Med* 2021; 40: 133-146**
- Riley R, Debray TP, Collins GS et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. ***Stat Med* 2021;40:4230-4251**
- Riley RD, Collins GS, Ensor J, et al. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. ***Stat Med* 2022**

\*\*\* CHECK OUT [www.prognosisresearch.com](http://www.prognosisresearch.com) \*\*\*

**Aiming to improve prognosis research in healthcare.**

Disseminating good practice, latest methods, introductory videos, software, training courses and more

### COURSES AT BIRMINGHAM (online)

- Statistical methods for risk prediction & prognostic models
- Statistical Methods for IPD Meta-analysis