UMC Utrecht
Julius Center

External validation of prediction models in big datasets

Thomas Debray, PhD
Julius Center for Health Sciences and Primary Care
Cochrane Netherlands

Cochrane

University Medical Center Utrecht

Background

Key references

- Riley et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016.

- Debray et al. Individual Participant Data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. PLOS MED 2015.

Why do we need external validation?

- The predictive performance of a model estimated on the development data is often too optimistic
- A prognostic model should provide predictions that are valid outside the specific context of the sample that was used for model development 5 months later or different hospital etc.
- How a model was derived is of little importance if it performs well.

What do we mean by external validation?

- Apply the clinical prediction model (CPM) to new individuals new populations
  - Temporal validation same population but 5 years later
  - Geographical validation new region or different setting (secondary care instead of primary care)
  - Domain validation

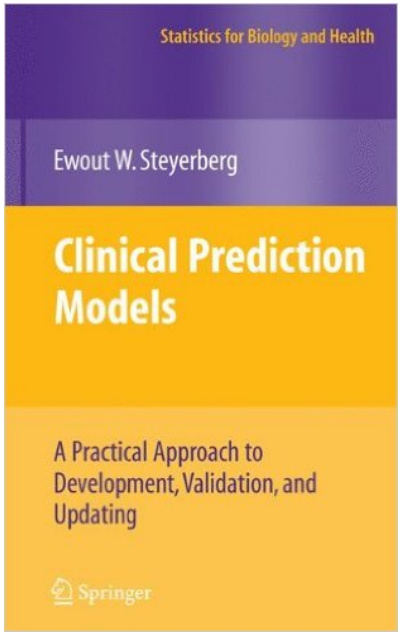
- Evaluate the predictive accuracy
  - Overall performance
  - Calibration typically quantified with these
  - Discrimination

Measures of prediction model performance

Overall performance

- Amount of explained variation (R^2)
- Brier score (for binary outcomes)

hard to interpret



Ref: Steyerberg. Clinical prediction models: a practical approach to development, validation and updating. Springer 2009.

Measures of prediction model performance

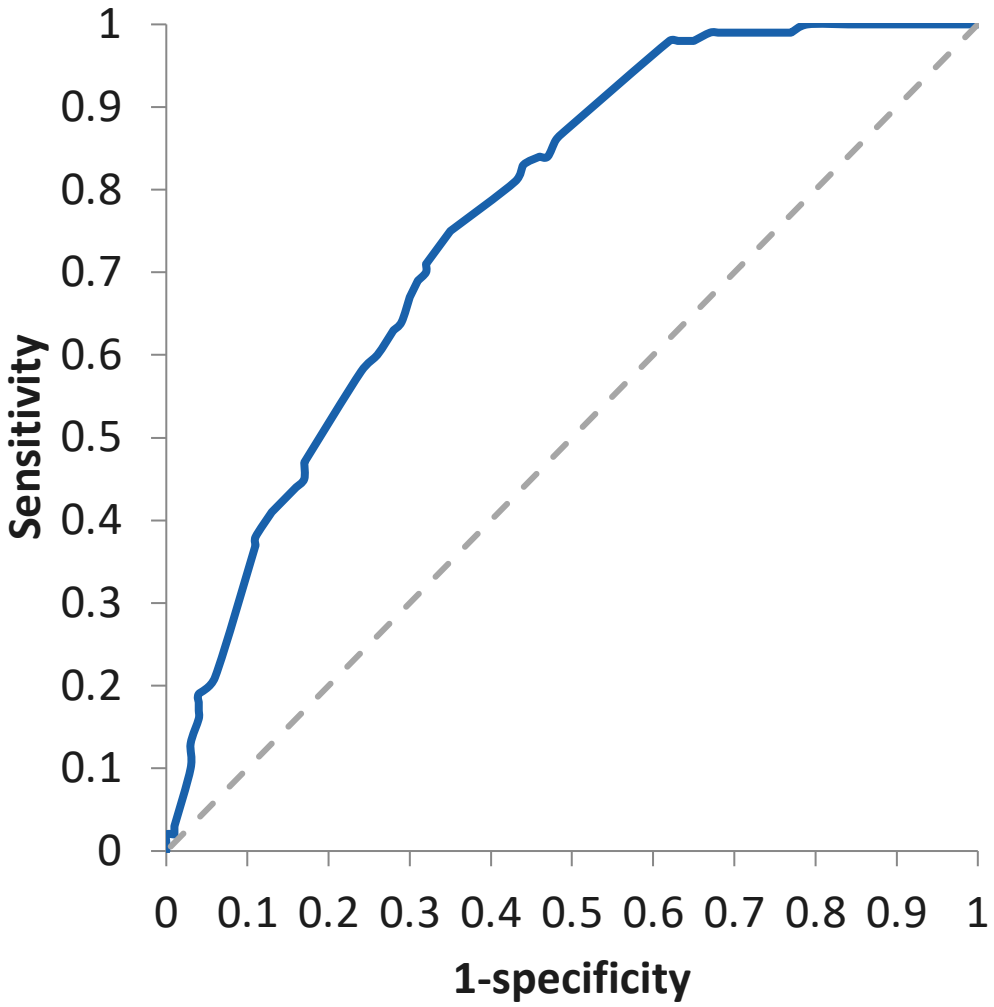
Discrimination

Quantifies the model's extent to distinguish between events and non-events

- Summary statistics
  - Concordance (c) index
  - Area under the ROC curve (AUC)
  - Discrimination slope
- Visual inspection
  - Receiving Operating Characteristics (ROC) curve

Measures of prediction model performance

Discrimination



AUC = 0.77

sensitivity and specificity - contrasts for different threshold (1 is perfect discrimination = ideal)

Measures of prediction model performance

Calibration

Agreement between observed outcomes and predictions

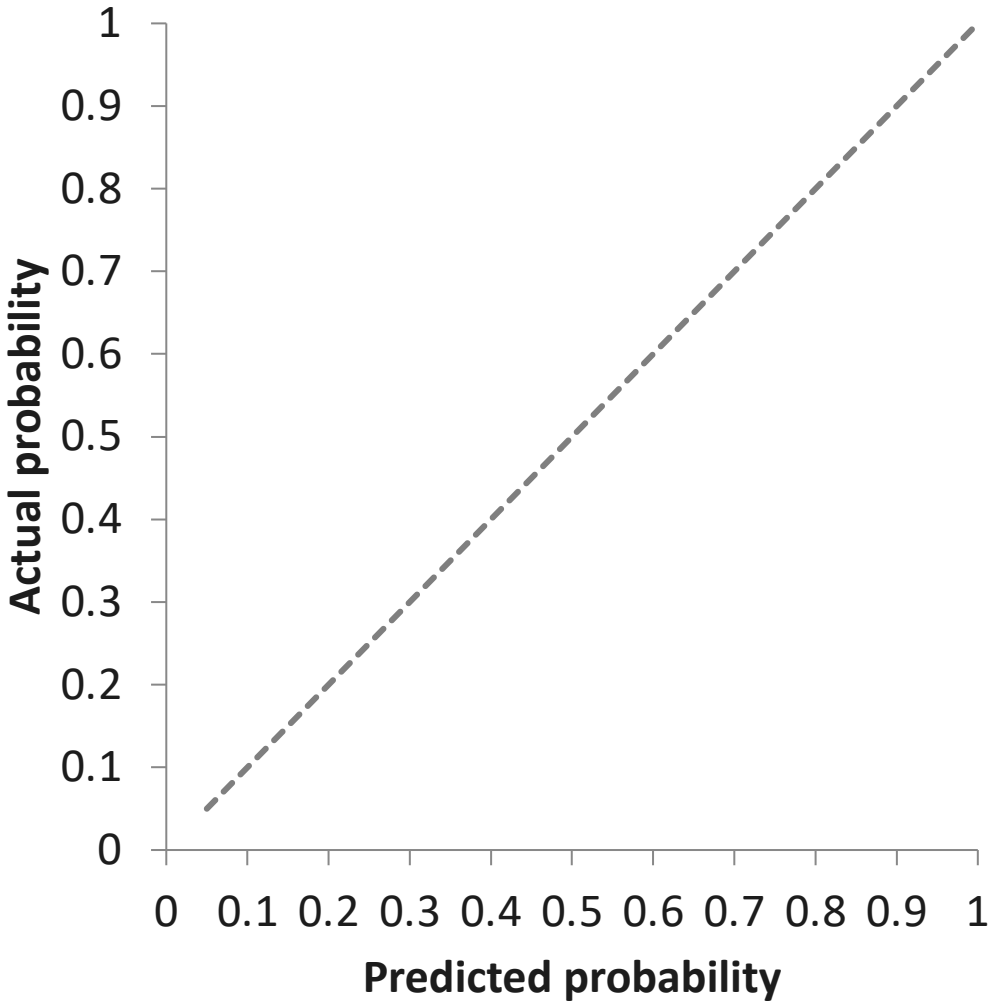
- Visual inspection
  - Calibration plot
- Summary statistics
  - O:E statistic (#observed events / #predicted events)
  - Calibration-in-the-large
  - Calibration slope
  - Hosmer-Lemeshow goodness-of-fit test

often explored visually with plot but also with statistics

Measures of prediction model performance

Calibration plot – good model?

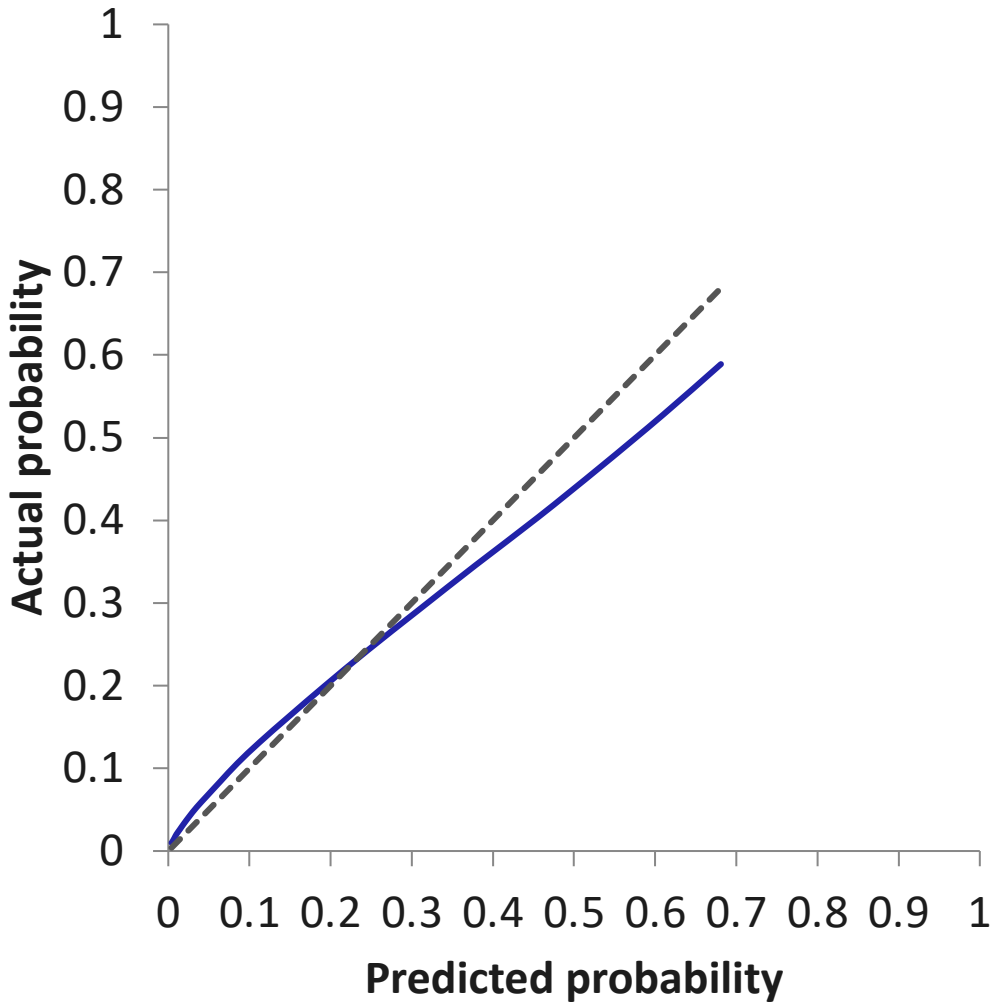
perfect mode = predicted probability = actual probability so OE =1 and slope = 1



Ideal calibration
O:E = 1
Slope = 1

Measures of prediction model performance

Calibration plot – good model?



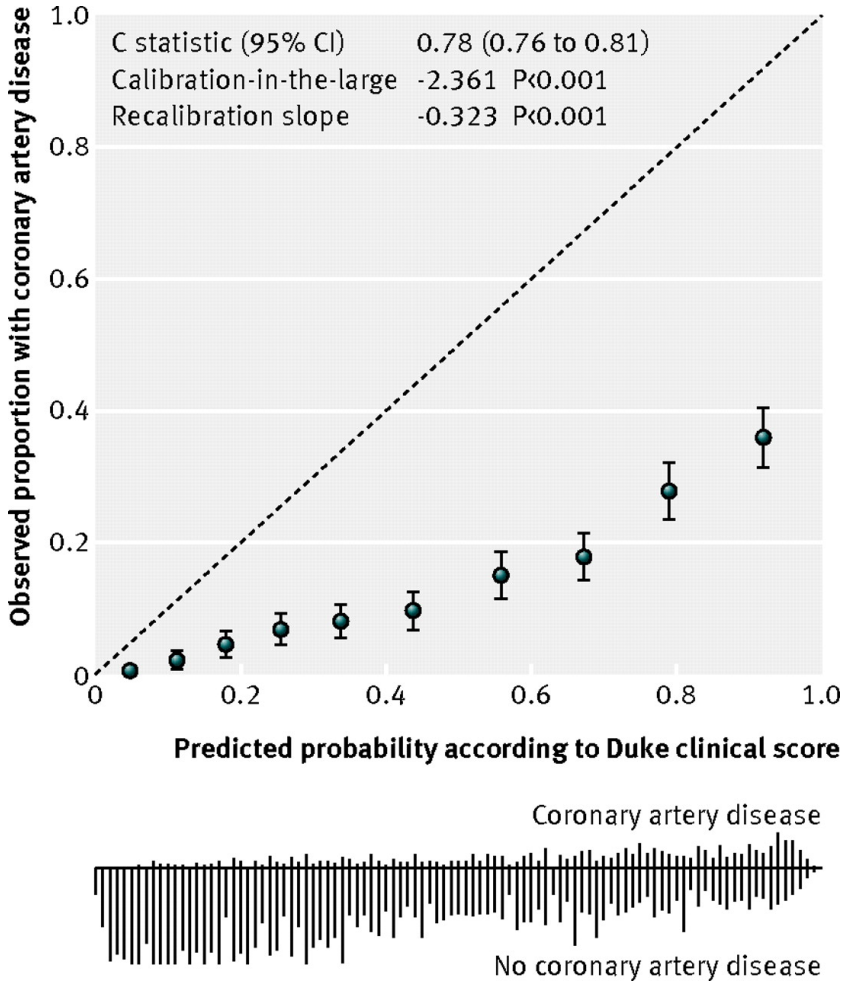
reality = rarely perfect here predicted probability are too extreme common example of overfitting

O:E = 1
Slope = 0.8

Sub-obtimal slope because curve does not follow reference line

Measures of prediction model performance

Calibration plot – good model?



in this validation more serious problem as S shape implies that there is over-fitting and predicted risk varies too much or is too extreme

all predictions - observed risk is much lower than predicted risk

predicted risk are systematically too high

Ref: Genders et al. Prediction model to estimate presence of coronary artery disease: retrospective pooled analysis of existing cohorts. BMJ 2012

Example study

Development and validation of a prediction model for the presence of serious bacterial infections in children with fever.

Development of the model

- Population: 379 children between 1 month and 36 months of age referred to the Emergency Department from a hospital in NL (75 events)
- Analysis: logistic regression with forward stepwise variable selection (57 variables -> 9 predictors)
- Internal validation:
  - AUC = 0.825 (95% CI: 0.78 – 0.87)
  - Bootstrap-corrected AUC = 0.756

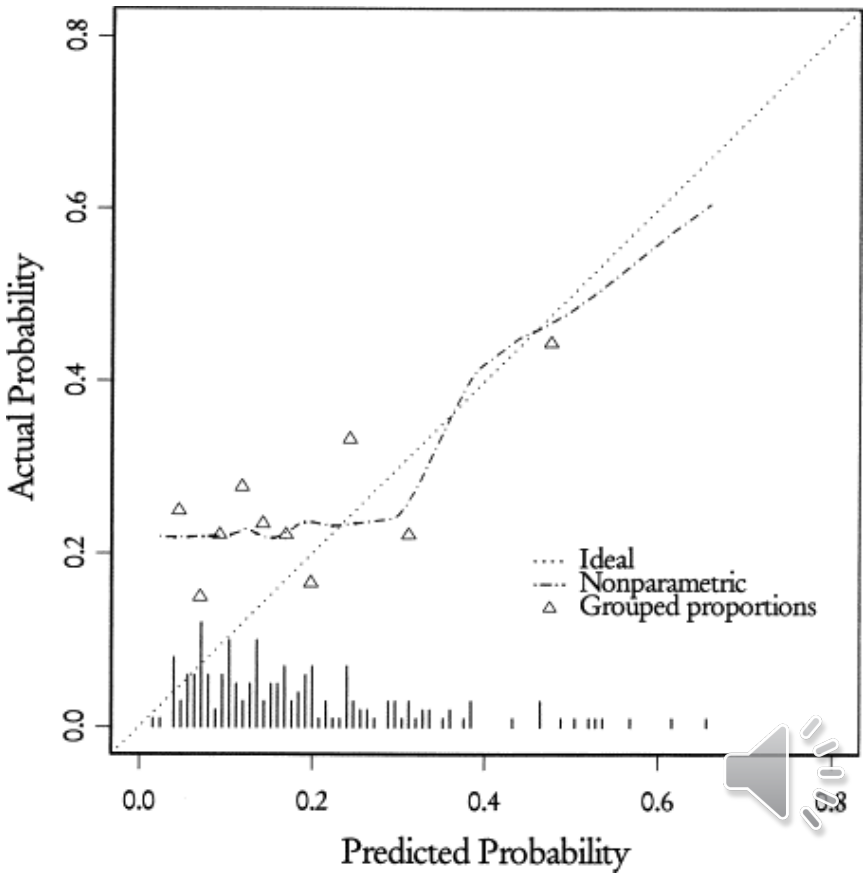
internal validation very optimistic

Example study

but when it was externially be validated the ROC decreased drastically

External validation of the model

- Population: 179 children from a different time period and another hospital from a different city in NL
- Very similar in- and exclusion criteria and variable def.
- AUC = 0.57
  (95% CI: 0.47 – 0.67)



Ref: Bleeker SE et al. External validation is necessary in prediction research: a clinical example. Journal of Clinical Epidemiology 2003.

Current shortcomings of validation studies

- External validation requires sufficient data large amount of data
  - Recommendations: >100 events and >100 non-events
  - Less data available for model development

- Not all validation studies are equally informative.
  - To what extent do individuals from the validation sample represent the target population?
  - To what extent are estimates of model performance affected by flaws in the design and analysis of the validation study?
  - To what extent can the CPM be implemented across different populations and settings?

Current shortcomings of validation studies

Why do we need big datasets for external validation?

- The predictive performance of a model tends to vary across settings, populations and periods

- Multiple external validation studies are needed to fully appreciate the generalizability of a prediction model

  one study doesn't say much about generalization

- Heterogeneity in model performance is rarely assessed, but investigating its extent is crucial to evaluate the model's potential generalizability and clinical usefulness.

  how likely is it to perform in new setting

Causes of heterogeneity in model performance

Invalid predictor effects

- Over-fitting of the prediction model to the development study (sometimes avoided using *penalization*)

  or model mist interaction or non-linear associations

- Biased estimates of predictor effects
  (e.g. due to flaws in the development study)

- Missed interactions or non-linear associations

Causes of heterogeneity in model performance

Discrepancies in outcome and predictor assessement

- Different measurement method for predictors
  (e.g. using equipment from different manufacturers)
- Different recording time of predictors
  (e.g. before or after surgery)
- Different quantification of predictors
  (e.g. use of cut-points may vary)
- Different disease and outcome definitions
- Different follow-up lengths

Causes of heterogeneity in model performance

Differences between study characteristics

Open access Research

BMJ Open

Empirical evidence of the impact of study characteristics on the performance of prediction models: a meta-epidemiological study

Johanna A A G Damen,[1,2] Thomas P A Debray,[1,2] Romin Pajouheshnia,[2] Johannes B Reitsma,[1,2] Rob J P M Scholten,[1,2] Karel G M Moons,[1,2] Lotty Hooft[1,2]

Causes of heterogeneity in model performance

Case-mix variation (spectrum effect)

- Different distribution of predictor values
- Different standards of care and treatment strategies
- Different starting points
  (e.g. earlier diagnosis due to screening program)
- Different outcome prevalence or incidence
- Different participant or setting characteristics

Case-mix variation can lead to genuine differences in the performance of a prediction model, even when the predictor effects remain ‘‘correct” in the validation study

Interpretation of model performance

Need to disentangle case-mix differences from differences in predictor effects!

across studies
important to differentiate

Journal of Clinical Epidemiology

Journal of Clinical Epidemiology 68 (2015) 279–289

ORIGINAL ARTICLES

A new framework to enhance the interpretation of external validation studies of clinical prediction models

Thomas P.A. Debray[a,*], Yvonne Vergouwe[b], Hendrik Koffijberg[a], Daan Nieboer[b], Ewout W. Steyerberg[b,1], Karel G.M. Moons[a,1]

[a]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Str. 6.131, PO Box 85500, 3508GA Utrecht, The Netherlands*

[b]*Department of Public Health, Erasmus Medical Center, Rotterdam, The Netherlands*

Accepted 30 June 2014; Published online 30 August 2014

Interpretation of model performance

large datasets with patients form different setting etc. clustering especially important

Need to adjust for clustering!

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
0(0) 1–14

DOI: 10.1177/0962280216668555
smm.sagepub.com
SAGE

Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study

L Wynants,[1,2] Y Vergouwe,[3] S Van Huffel,[1,2] D Timmerman[4] and B Van Calster[3,4]

Abstract

Clinical risk prediction models are increasingly being developed and validated on multicenter datasets. In this article, we present a comprehensive framework for the evaluation of the predictive performance of prediction models at the center level and the population level, considering population-averaged predictions, center-specific predictions, and predictions assuming an average random center effect. We demonstrated in a simulation study that calibration slopes do not only deviate from one because of over- or underfitting of patterns in the development dataset, but also as a result of the choice of the model (standard versus mixed effects logistic regression), the type of predictions (marginal versus conditional versus assuming an average random effect), and the level of model validation (center versus population). In particular, when data is heavily clustered (ICC 20%), center-specific predictions offer the best predictive performance at the population level and the center level. We recommend that models should reflect the data structure, while the level of model validation should reflect the research question.

Examining heterogeneity and improving model performance

Recommendations

- Calculate key performance statistics in each cluster (e.g. study or hospital)
- Summarize the performance measures by applying (multivariate) random effects meta-analysis
- Quantify between-study heterogeneity in model performance using 95% prediction intervals

inspect generalizable

Examining heterogeneity and improving model performance

Guidance paper

RESEARCH METHODS AND REPORTING

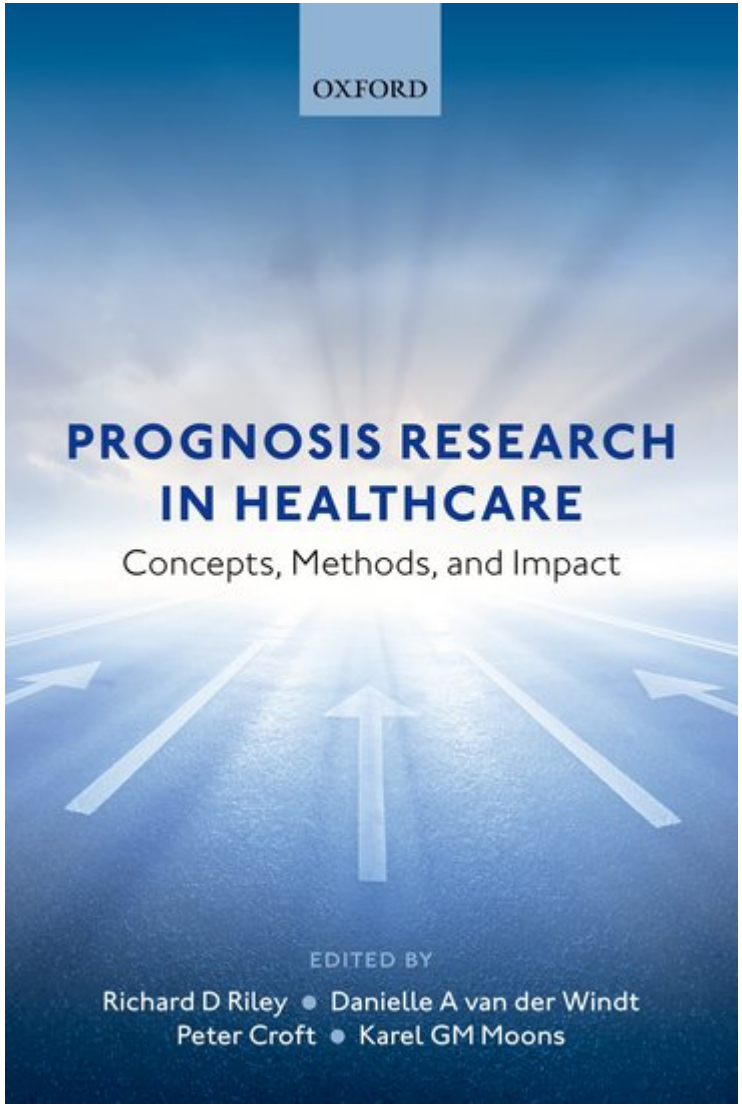
A guide to systematic review and meta-analysis of prediction model performance

Thomas P A Debray,[1,2] Johanna A A G Damen,[1,2] Kym I E Snell,[3] Joie Ensor,[3] Lotty Hooft,[1,2] Johannes B Reitsma,[1,2] Richard D Riley,[3] Karel G M Moons[1,2]

Examining heterogeneity and improving model performance

More guidance

- Prognostic model research
- Systematic reviews and meta-analysis of prognosis research studies
- Individual participant data meta-analysis of prognosis studies
- Electronic healthcare records and prognosis research



Example 1

Diagnosis of deep vein thrombosis in patients suspected of DVT

Previously developed prediction model for diagnosing DVT

- Logistic regression analysis
- Three predictors
  - *Sex*
  - *Surgery*
  - *Calf difference*

$$Pr(\text{DVT}) = \frac{1}{1 + \exp(-(\alpha + \beta_1 \text{sex} + \beta_2 \text{surg} + \beta_3 \text{cdif})}$$

Example 1

External validation in 12 studies

- Sample size: 153 – 1768 (total N=10014) both varies
- Event occurrence: 8% - 39% (total E=1897)
- Results (95% confidence interval)
  - Calibration-in-the-large: -0.004 (-0.313; 0.305) close to 0
  - Calibration slope: 0.980 (0.853; 1.107) close to 1
  - E:O ratio: 1.02 (0.81; 1.28) ratio of expected versus observed close to 1
  - C-statistic: 0.687 (0.669; 0.705)

Does the model predict well? What about generalizability?

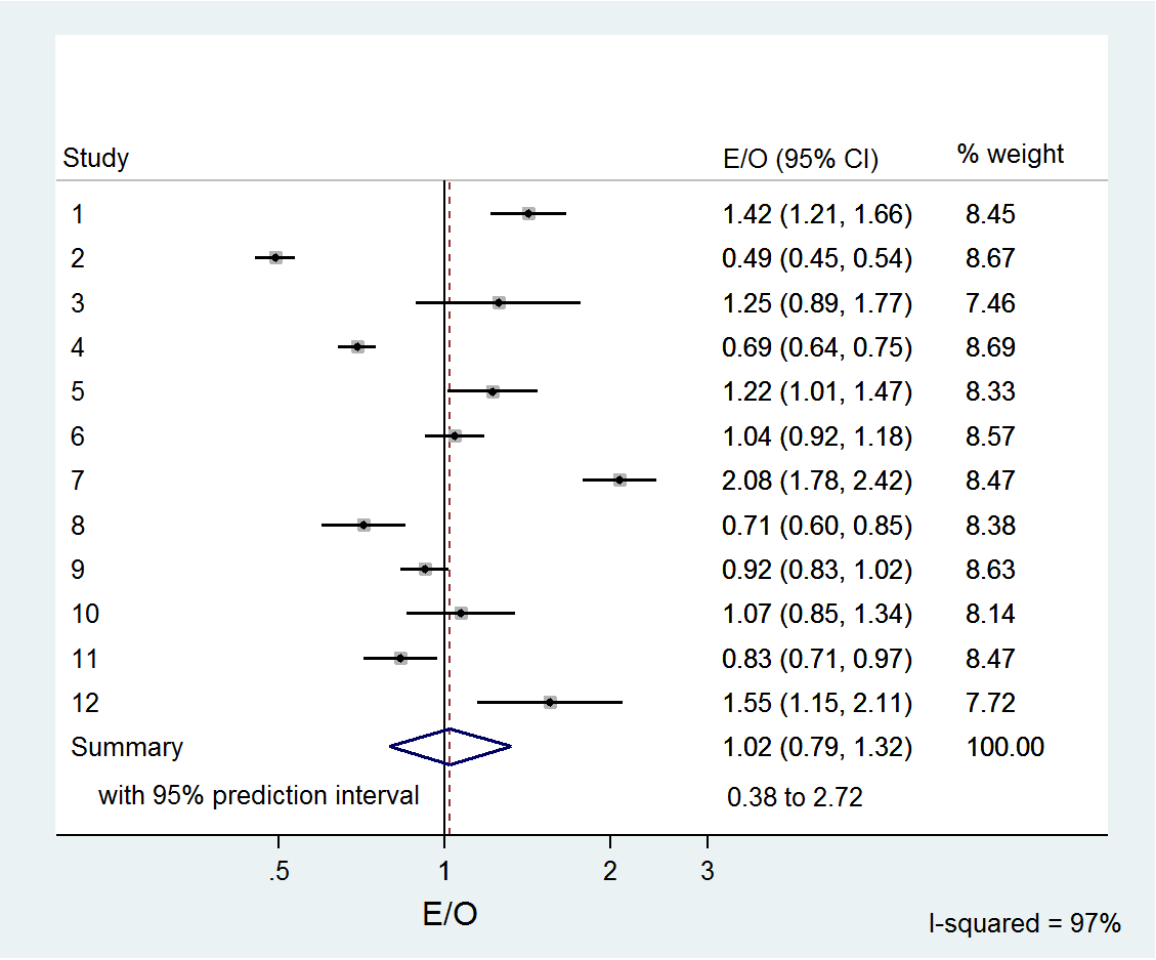
on average the prediction model yields adequate calibration BUT doesn't tell us much about generalizability

how does it perform in individual studies?

Example 1

External validation in 12 studies

forest plot



performs well on average but not in individual studies (over or underestimated outcome)

Example 1

External validation in 12 studies

- Substantial between-study heterogeneity! can be quantifies with prediction intervals
- Approximate 95% prediction intervals:
  - Calibration slope = 0.59 to1.38 quite substantial
  - E:O ratio= 0.38 to 2.72
  - c-statistic = 0.64 to 0.73

we can expect quite the deficiencies for individual use - needs to be improves

The model requires improvements to improve discrimination and to be clinically useful!

Example 2

Prognosis of cardiovascular disease in patients from general practice using QRISK2

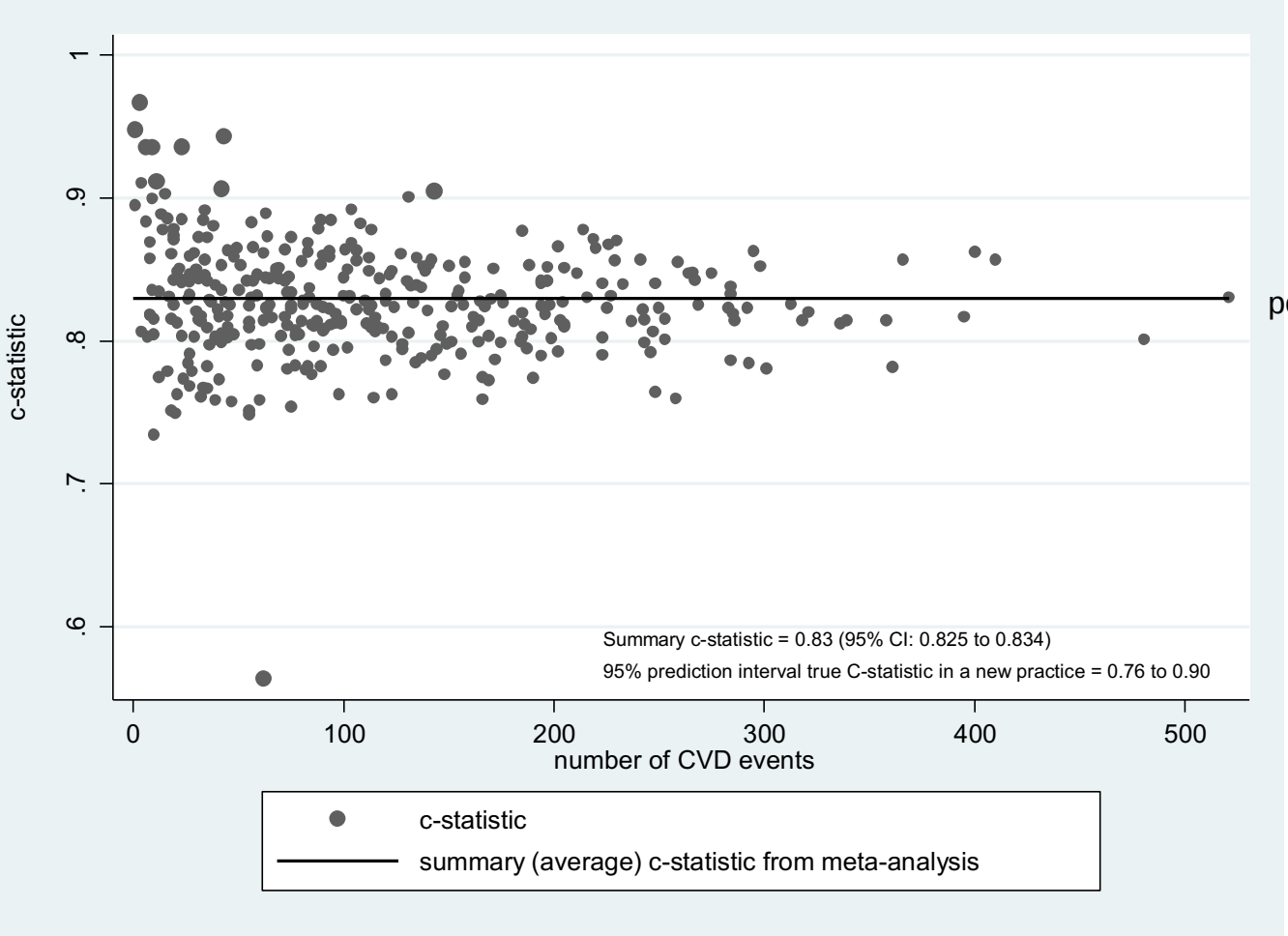
- 364 practices from the UK
- Total sample size N=2,084,445
- Total number of events E=93,564

Again, each cluster might be viewed as a different external validation study!

Example 2

External validation in 364 practices

clusters therefore difficult to make forest plot

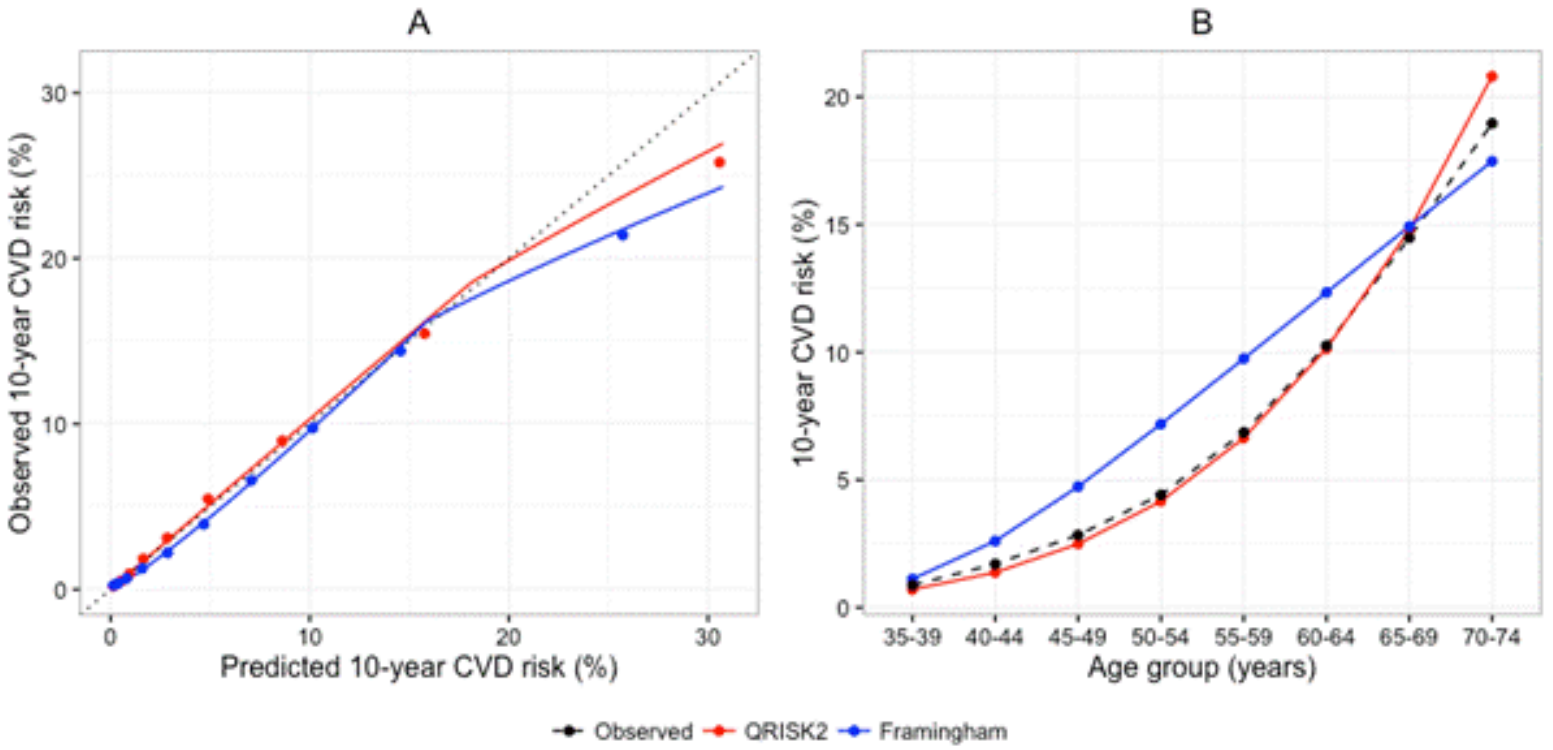


pooled c-statistic

Example 2

External validation in 364 practices

advantage of big data= can be stratified external validation against different risk groups



-> predicted and observed risk generally correspond to the model risk

Example 2

External validation in 364 practices

again with random effect meta-analysis

- Discrimination
  - Summary c-statistic: 0.83
  - 95% confidence interval: 0.825 to 0.834
  - 95% prediction interval: 0.76 to 0.90

- Calibration
  - E:O summary estimate: 1.01 close to 1 again
  - Slight over-prediction in women at higher CVD risk
  - QRISK2 appears to accurately predict 10-year CVD risk across all age groups.

Example 2

Investigating between-study heterogeneity

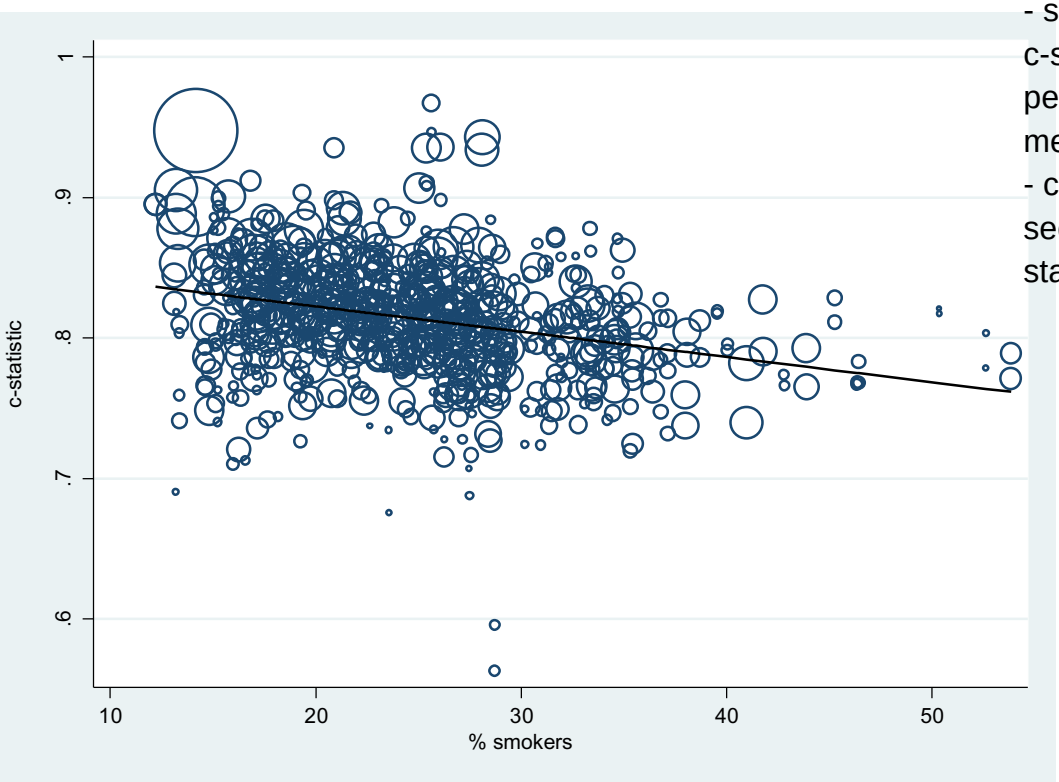
how is it affected by heterogeneity

- Recall that variation in case-mix severity and case-mix heterogeneity may affect model performance.
  - Larger case-mix variation is related to larger discrimination performance
  - Populations with a narrower case-mix tend to have worse discrimination performance

- Recall that discrimination of QRISK2 heterogeneous
  - Population mean age
  - Percentage smokers

Example 2

Investigating between-study heterogeneity



results of meta reression analysis
- shows that graphs with higher c-statistics have lower percentage of smokers which means:
- case mix of validation study seems to have an impact on c-statistics of prediction model

NB Circle size is weighted by the precision of the c-statistic estimate (i.e. larger circles indicate c-statistic estimates with smaller standard errors, and thus more weight in the meta-regression)

Example 3

Prognosis of coronary heart disease

- 37 prospective studies
- Sample size (total N=165,856)
- Event occurrence (total E=8,806)
- Average follow-up of 9.8 years

Ref: Pennells LS et al. Assessing Risk Prediction Models Using Individual Participant Data From Multiple Studies. AJE 2013.

Example 3

External validation in 37 studies

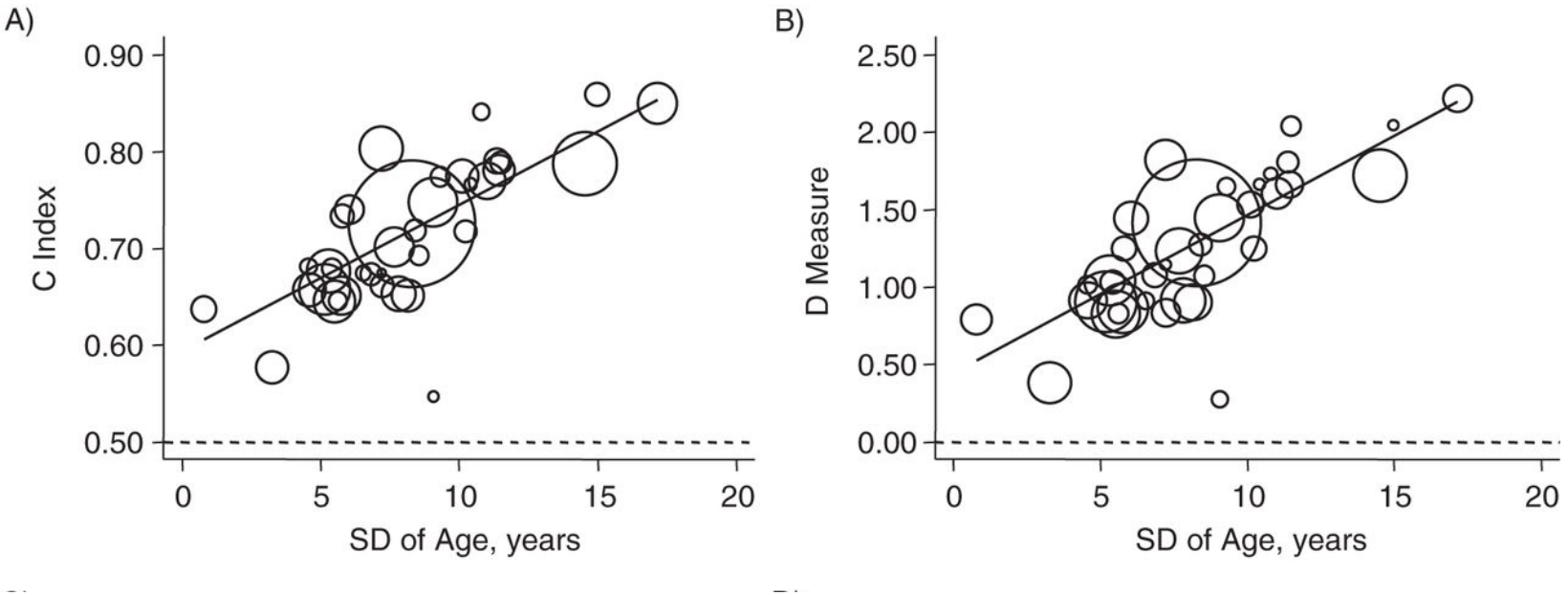
- Discrimination
  - Summary c-statistic: 0.715
  - 95% confidence interval: 0.694 to 0.736
  - 95% prediction interval: 0.59 to 0.84

very wide interval - heterogenity

Again, there is substantial heterogeneity in model performance!

Example 3

Investigating between-study heterogeneity



is the standard deviation of age associated with the discrimination in performance
larger SD age -> larger disc.

Practical and methodological challenges

Caution is warranted when interpreting summary estimates of model performance and between-study heterogeneity.

- **Data quality**
  - Missing predictor values
  - Non-standardised definitions of diagnoses and outcomes
  - Incomplete follow-up times and event dates
  - Lack of recording of novel/costly predictors
  - Risk of double entries
- **Data dredging**

Need for study protocols and quality appraisal tools!