A calibration hierarchy for risk prediction models

Ben Van Calster

Dept of Development and Regeneration, KU Leuven

Dept of Public Health, Erasmus MC, Rotterdam

Collaborators:

Daan Nieboer, Bavo De Cock,

Yvonne Vergouwe, Ewout Steyerberg

MEMTAB, Birmingham, July 20th 2016

Risk prediction models and calibration

Predict the risk of disease given a set of predictor variables

Yet most attention goes to discrimination

→ Do patients with disease get higher risks than patients without? (AUC, c)

Usually less attention goes to calibration

→ Are the predicted risks in fact accurate?

"For informing patients and medical decision making, calibration is the primary requirement" (Steyerberg, 2009)

"Well-calibratedness is more important because it indicates that the predictions have aggregate validity" (Kim & Simon, Biostatistics 2011)

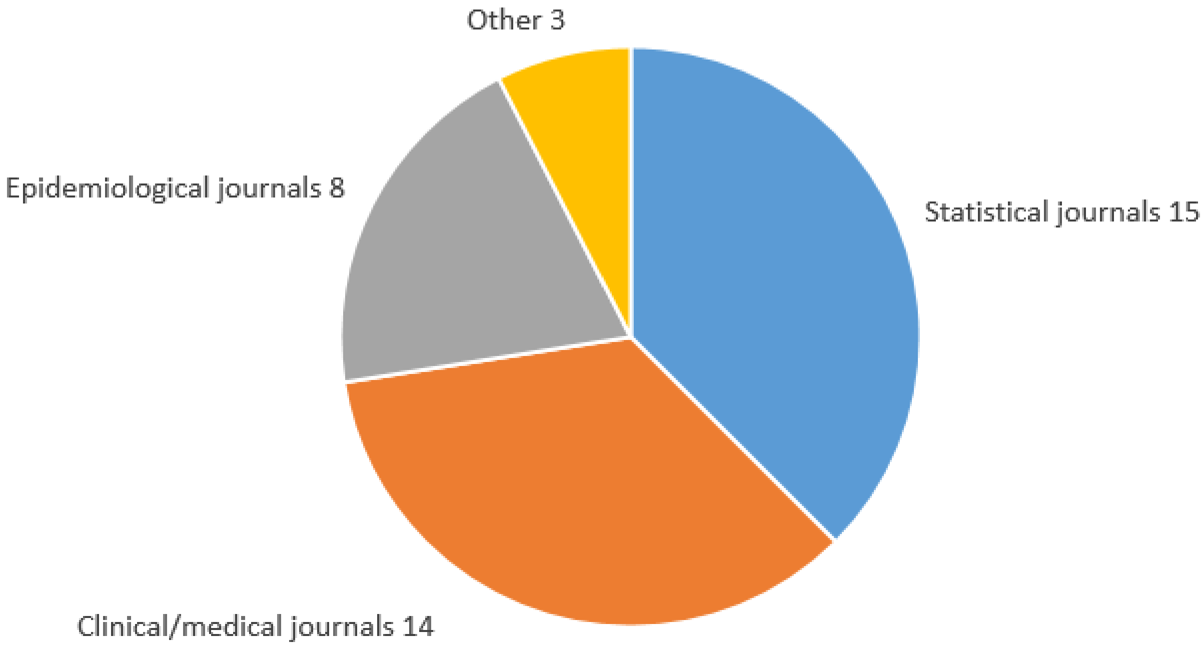
KU LEUVEN

What is calibration?

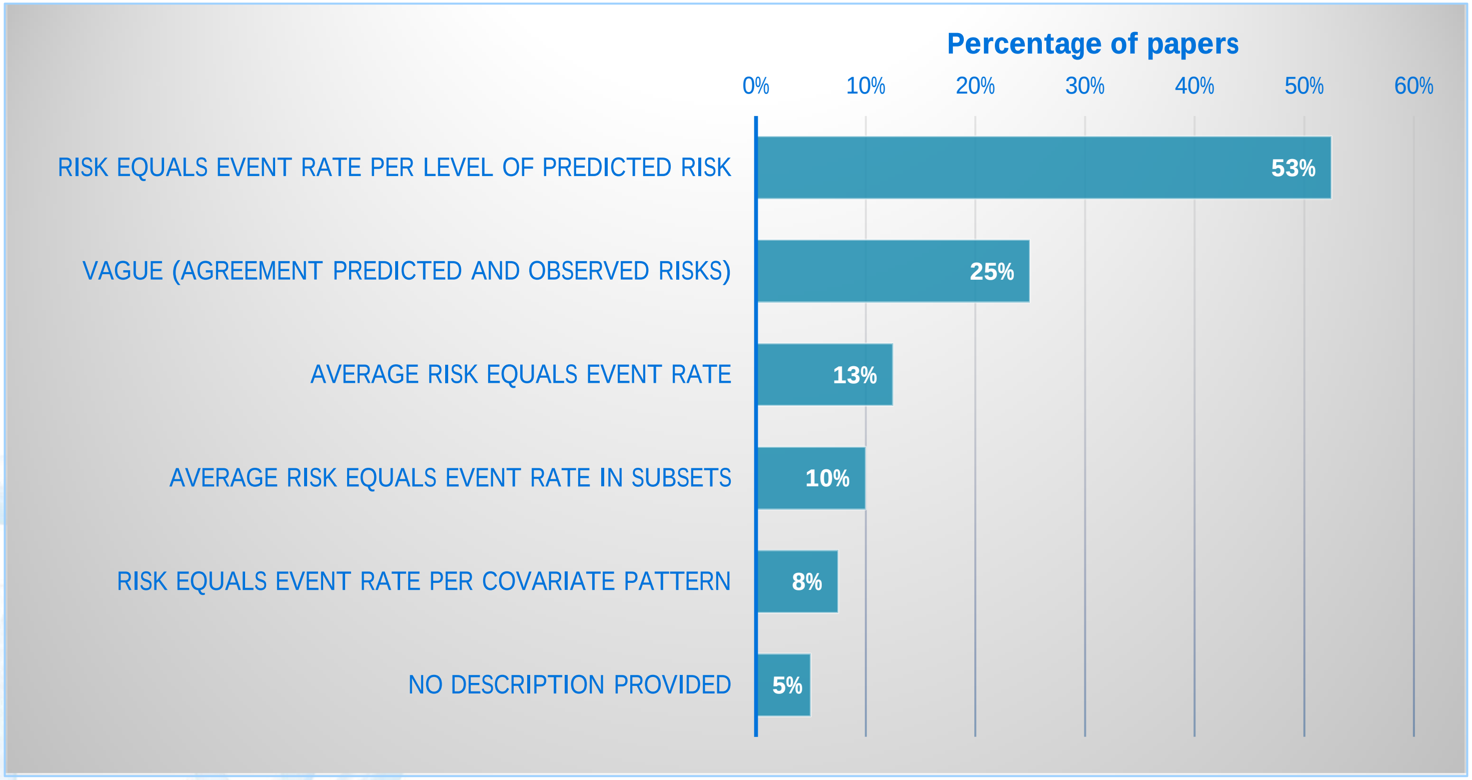
What is calibration?

Non-systematic review of methodological literature

40 papers

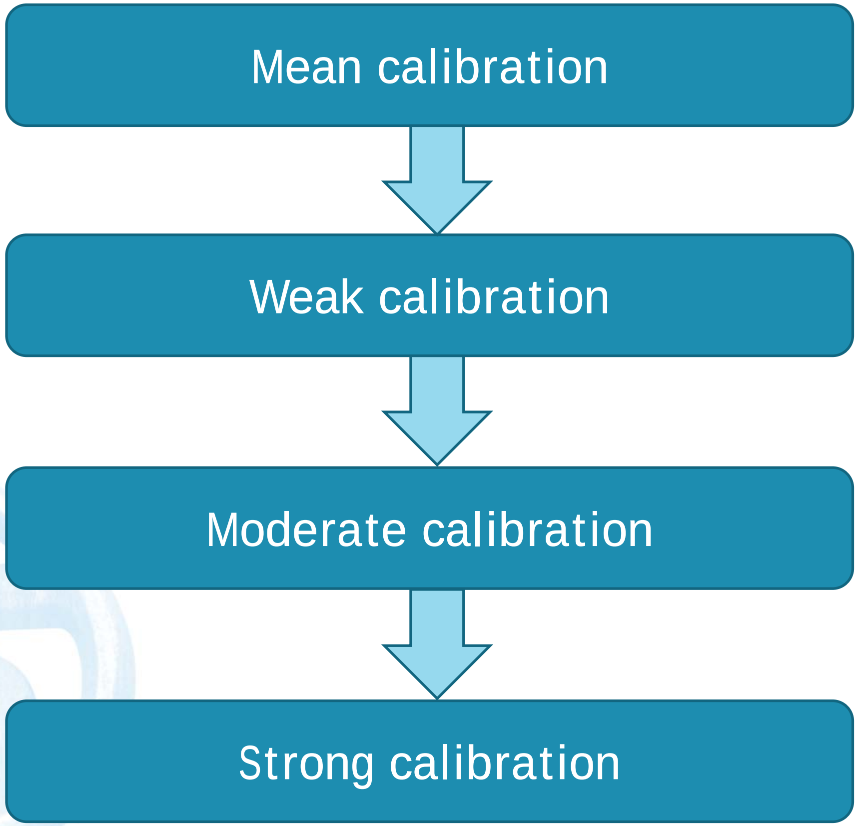


What is calibration?



KU LEUVEN

Hierarchy of calibration



Setting

- Binary outcome Y
- Dataset to evaluate calibration of a risk model
- Model based on logistic regression
  - Linear predictor $L = \hat{\alpha} + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$
- Covariate pattern: $\{x_1 \dots x_p\}$

KU LEUVEN

Methods: logistic recalibration (Cox, Biometrika 1958)

- $logit(Y) = a + b \times L$

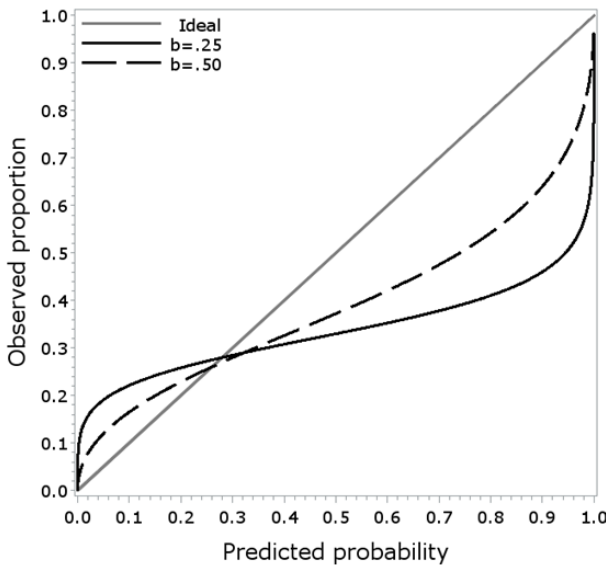
- b is the calibration slope
  - $b < 1$ suggests overfitting, risks are too extreme
  - $b > 1$ suggests underfitting

- a when fixing b to 1, $a_{b=1}$, is the calibration intercept
  - $a_{b=1} < 0$ indicates general overestimation of risks
  - $a_{b=1} > 0$ indicates general underestimation of risks

- Result of this model is an indirect estimation of observed event rate given predicted risk

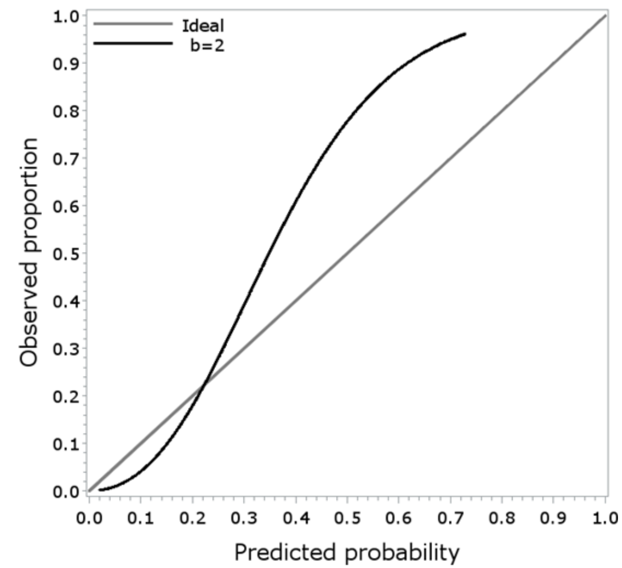
KU LEUVEN

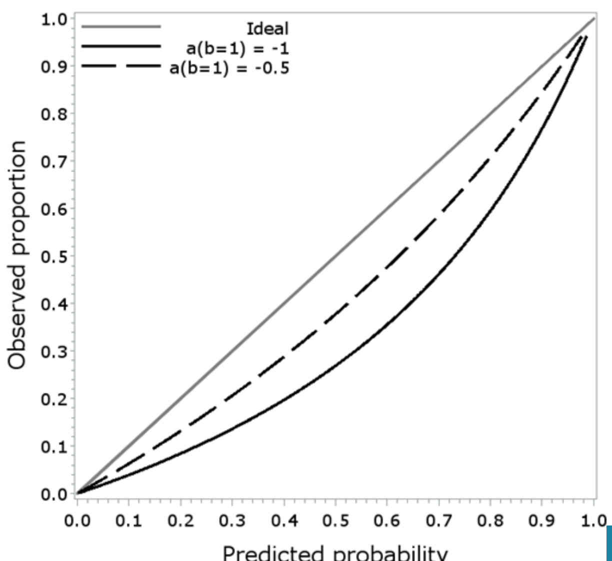
Methods: logistic calibration curves



$$b < 1$$
$$a_{b=1} = 0$$

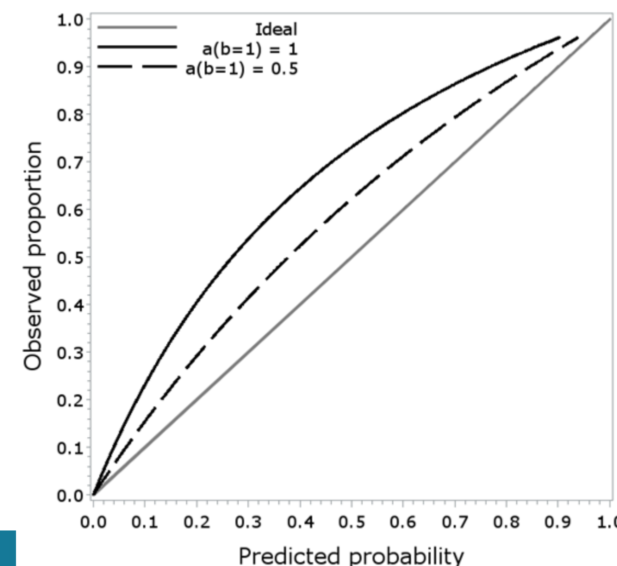
$$b > 1$$
$$a_{b=1} = 0$$





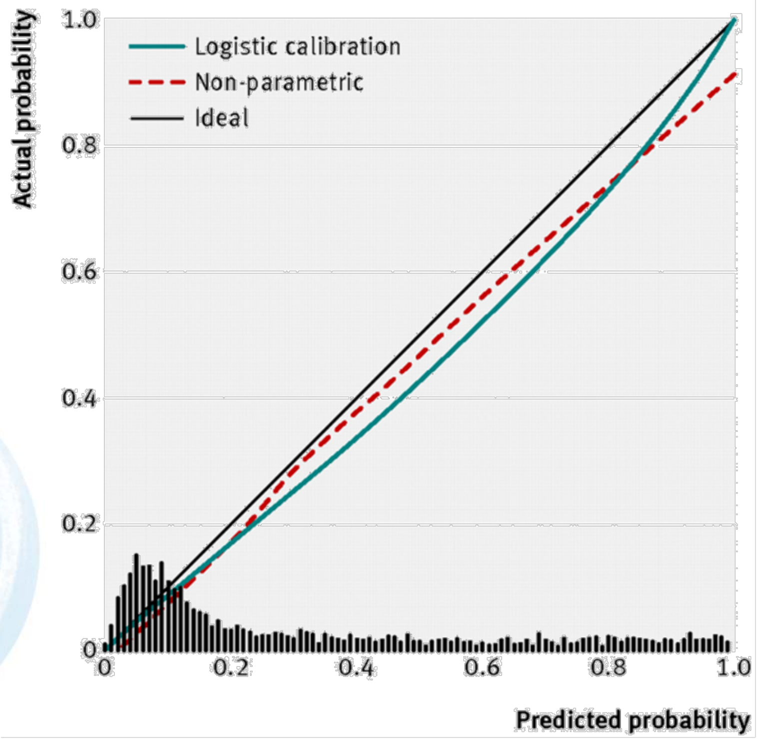
$$b = 1$$
$$a_{b=1} < 0$$

$$b = 1$$
$$a_{b=1} > 0$$



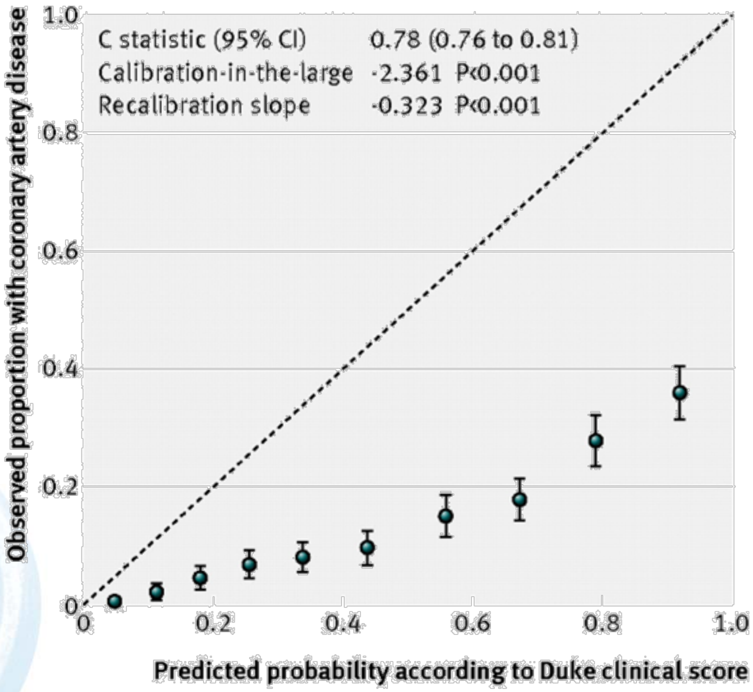
Methods: flexible calibration curves

- $logit(Y) = a + f \times L$, with f based on loess or splines



Methods: grouped calibration curves

- E.g. per decile of predicted risk (flexible)



(Genders et al, BMJ 2012)

1. Mean calibration

- "Calibration in the large"

- Average predicted risk equals event rate

- Assessment
  - Compare average risk with event rate
  - Calibration intercept $a_{b=1}$

- Clearly insufficient (can miss overfitting)

2. Weak calibration

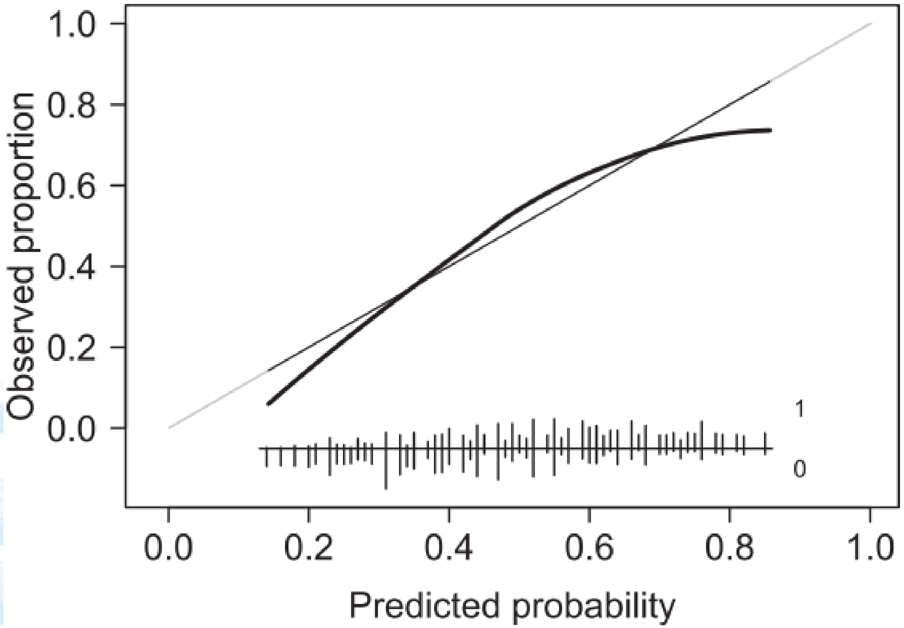
- b = 1, $a_{b=1}$ = 0
  - Logistic calibration curve equals the diagonal

- No overfitting or underfitting, no general over- or underestimation

- Insufficient:
  - By definition satisfied on development data when basic ML is used, independent of how predictors are modeled

KU LEUVEN

Simulated results



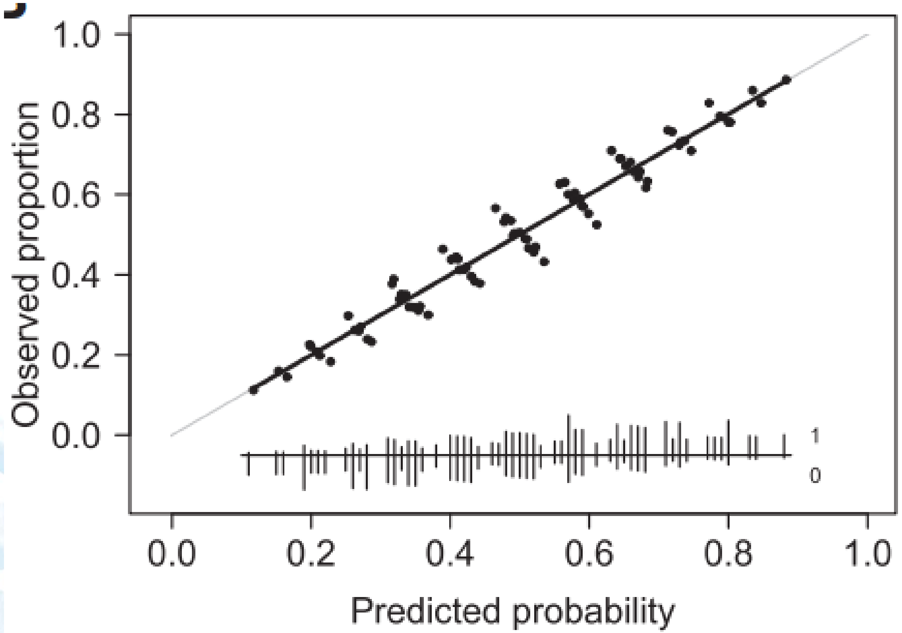
3. Moderate calibration

- Predicted risk equals event rate per level of predicted risk
  - Among patients with x% risk, x out of 100 have the event
  - Flexible calibration curve on diagonal

- Can reveal miscalibration missed by logistic recalibration

- But not perfect yet…

Simulated results



4. Strong calibration

- Predicted risk equals event rate *per covariate pattern*
- Different covariate patterns may have same predicted risk but different event rate
- Clinically desirable: unbiased risk predictions for all patients
- Always assessed relative to predictors in the model!

KU LEUVEN

Assessment of strong calibration

- Usually impossible: too few patients per covariate pattern

- Method that approaches the assessment of strong calibration:
  - Compare average predicted risk and event rate for subgroups of patients determined by values of one or more predictors
  - Still: curse of dimensionality!

Strong calibration: utopia

- Model (given the predictors) is correct for the validation population
- Is it realistic to have the correct model?
  - Correct model specification (e.g. GLM with logit link)
  - ML estimation only gives asymptotically unbiased estimates
  - Overfitting: combination of estimated model coefficients
  - All nonlinear effects and interaction effects are correctly modeled
  - (Systematic) measurement error
- Vach (2013): “the idea to identify the true model by statistical means is just a great wish which cannot be fulfilled”

KU LEUVEN

Clinical decision making

- Assume risk threshold T to decide whether or not to treat

- Odds(T) is harm-to-benefit ratio (Pauker & Kassirer, NEJM 1975)
  - T = 0.25, odds 1:3, one TP is worth up to 3 FP

- Net Benefit (Vickers & Elkin, MDM 2006) quantifies utility of decisions
  - $NB = \frac{TP - odds(T) \times FP}{N}$, net proportion of TP
  - Plot NB by threshold: decision curve

- Compare NB of model at T with NB of treat all or treat none
  - Model worse than treat all or treat none: harmful decisions

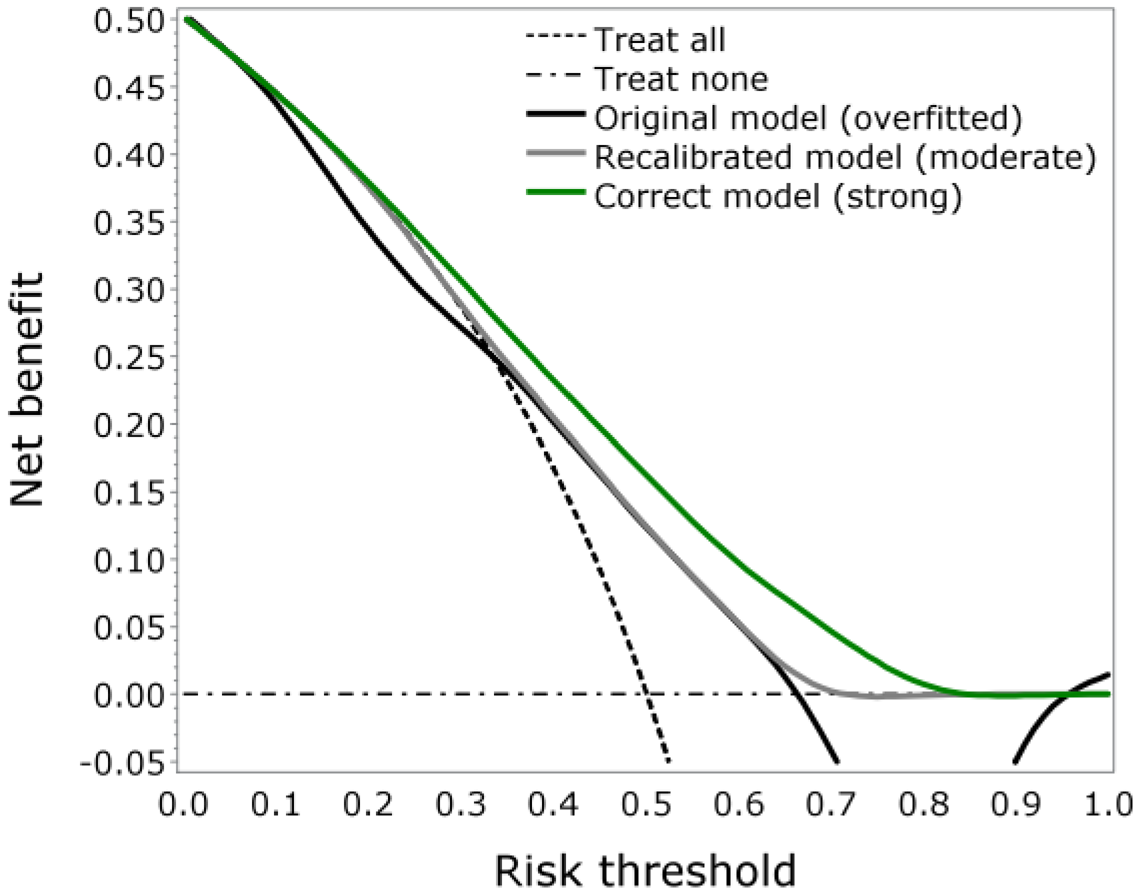
KU LEUVEN

Calibration and clinical decision making

- Strong calibration: utility of decisions (NB) maximized
- Moderate calibration: non-harmful decisions guaranteed (proof in paper)
- Below moderate calibration: harmful decisions at some *T* (Van Calster & Vickers, MDM 2015)

KU LEUVEN

Simulated results



Pragmatic focus

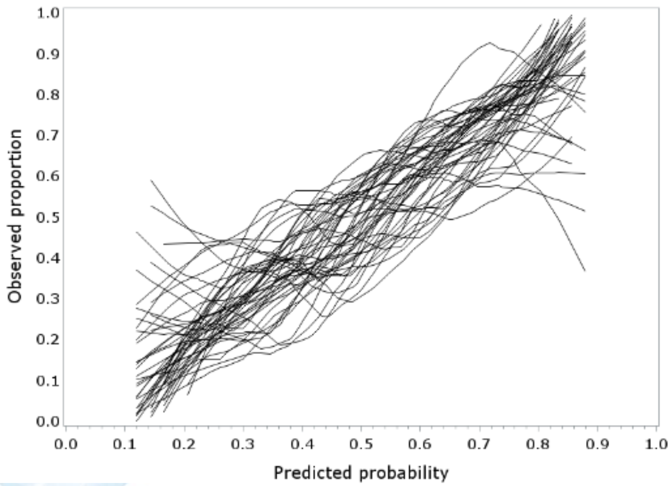
When developing or validating models, focus on moderate calibration

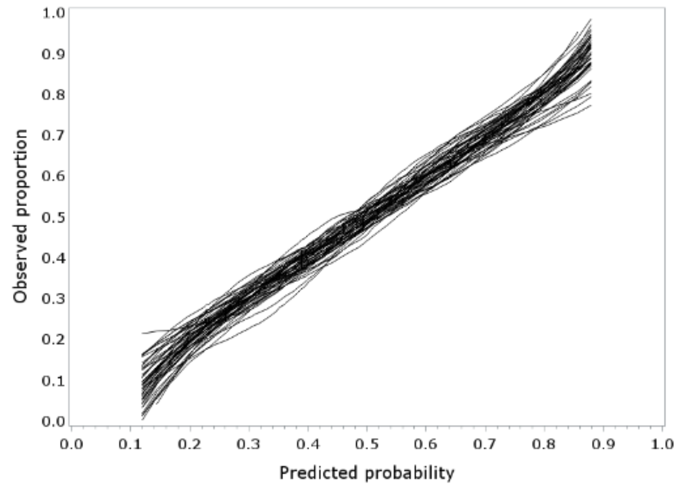
- Guarantees non-harmful clinical decisions
- Strong calibration is utopic and counterproductive
- Weak/moderate calibration is hard enough as it is…

KU LEUVEN

Sample size for validation

- Observed calibration curves will usually not be on diagonal
- Confidence intervals are important
- At least 200 events for flexible curves



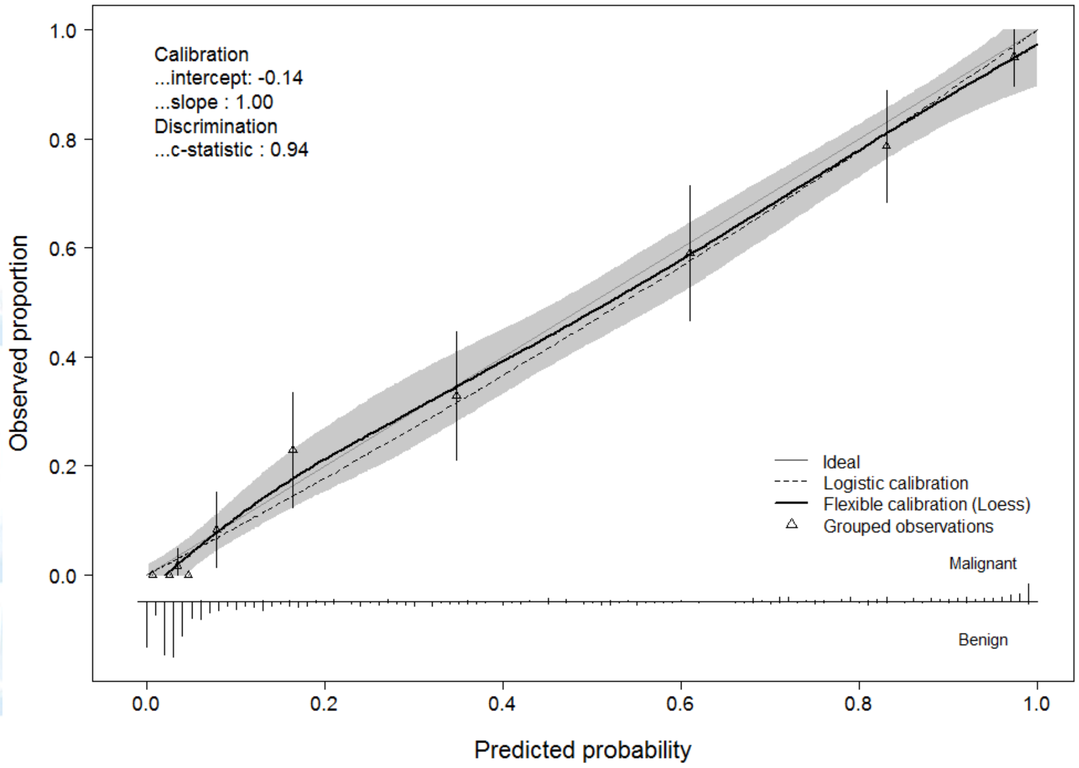


KU LEUVEN

Example

External validation of ADNEX model for ovarian tumor diagnosis

N=610, 182 events (Sayasneh et al, BJC in press)



KU LEUVEN

val.prob.ci.2

https://github.com/BavoDC/CalibrationCurves
www.clinicalpredictionmodels.org