



# Calibration of risk prediction models: decision making with the lights on or off?

Ben Van Calster  
KU Leuven (B), LUMC (NL)



# TG6: Evaluating diagnostic tests and prediction models

- **Chairs:**

- Ewout Steyerberg (Leiden LUMC)
- Ben Van Calster (KU Leuven)

- **Members (alphabetically):**

- Patrick Bossuyt (AMC Amsterdam)
- Tom Boyles (U Witwatersrand, Johannesburg; clinician member)
- Gary Collins (U Oxford)
- Kathleen Kerr (U Washington, Seattle)
- Petra Macaskill (U Sydney)
- David McLernon (Aberdeen)
- Carl Moons (UMC Utrecht)
- Maarten van Smeden (UMC Utrecht)
- Andrew Vickers (MSKCC, New York)
- Laure Wynants (U Maastricht)

# Risk prediction or binary prediction?



**Andrew Vickers**

@VickersBiostats



I've heard it a million times: don't give a doctor a risk prediction for a patient, they can't handle it. You have to give them a cut-point and put patients into hi vs. lo risk-groups so they can make an easy clinical decision. Any reason to believe this argument is true?

12:36 AM · Aug 18, 2020



100



61 people are Tweeting about this



# Risk prediction or binary prediction?



**Raj Mehta, MD** @raj\_mehta · Aug 18

Replying to @VickersBiostats

No. We do risk prediction with probability every day (i.e. ASCVD calculator).

Furthermore, most of this can be integrated and automated in the EHR, so old arguments about keeping things simple to calculate by hand no longer apply.



2



16



**Luke Oakden-Rayner** @DrLukeOR · Aug 18

But you don't. That is a risk \*stratification\* tool. You turn the probability into low 10 yr risk vs high risk.

Unless you could make 100 different decisions, what is the point of a 100 point scale?



2



4



**Raj Mehta, MD** @raj\_mehta · Aug 18

Decision stratification comes after risk prediction.

Once we have the probability, we can create multiple different decision cut-offs with patient preferences in mind.

We don't need a single cut-off; a suggestion of various values (like A1c goals) is a better alternative.

out my thoughts on a common argument: should models produce probabilities or decisions? I.e. 32% chance of cancer vs "do a biopsy".

I favour the latter, because IMO it is both more useful and... more honest. IMO:

Not at all! No human can balance a 30% chance of cancer vs a 32% chance of cancer. This is #TMI.

Even in shared decision making, most patients prefer terms like "rare" and "almost certainly" vs 3% or 95%.

# Risk prediction or binary prediction?

Risk is most interpretable, acknowledges imperfect prediction, can be combined with other information, and allows to vary decision thresholds.

If you predict risk, you can assess the accuracy of the estimates (calibration). Binary predictions easily hide potential miscalibration.



**Jeremy Sussman**  
@JeremySussman

Replying to @DrLukeOR

The assumption here is that there is zero information that would enter into a decision outside what's in the model. This is almost never the case, nor should it be.

3:28 AM · Jul 30, 2020 · Twitter for iPhone



**Jeremy Sussman** @JeremySussman · Jul 30  
Replying to @JeremySussman and @DrLukeOR

Most important are patients values for shared decisions, which is lost with "get a biopsy." But it also misses unencoded variables or ones that aren't part of the risk model. "You didn't ask, but my dad had lung cancer at age 30." Clinically, these are the norm, not exceptions.



**Laure Wynants**  
@laure\_wynants

Replying to @obenfine and @DrLukeOR

I disagree. I understand the cognitive burden and am aware of studies that demonstrate limited numeracy of clinicians. No one argues against thresholds or try recommendations based on them. But if the threshold is 30%, it makes a difference whether the risk is 29% or 1%.

# Level 1-2 TG6 paper on calibration

Van Calster et al. *BMC Medicine* (2019) 17:230  
<https://doi.org/10.1186/s12916-019-1466-7>





BMC Medicine

OPINION

Open Access

## Calibration: the Achilles heel of predictive analytics



Ben Van Calster<sup>1,2,6\*</sup> , David J. McLernon<sup>3,6</sup> , Maarten van Smeden<sup>2,4,6</sup> , Laure Wynants<sup>1,5</sup>, Ewout W. Steyerberg<sup>2,6</sup>   
On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative<sup>6</sup>

### Abstract

**Background:** The assessment of calibration performance of risk prediction models based on regression or more flexible machine learning algorithms receives little attention.

**Main text:** Herein, we argue that this needs to change immediately because poorly calibrated algorithms can be misleading and potentially harmful for clinical decision-making. We summarize how to avoid poor calibration at algorithm development and how to assess calibration at algorithm validation, emphasizing balance between model complexity and the available sample size. At external validation, calibration curves require sufficiently large samples. Algorithm updating should be considered for appropriate support of clinical practice.

**Conclusion:** Efforts are required to avoid poor calibration when developing prediction models, to evaluate calibration when validating models, and to update models when indicated. The ultimate aim is to optimize the utility of predictive analytics for shared decision-making and patient counseling.

**Keywords:** Calibration, Risk prediction models, Predictive analytics, Overfitting, Heterogeneity, Model performance

# The Achilles heel of predictive analytics

Systematically wrong risk estimates can distort decision-making

- Risk overestimated: can lead to many unnecessary interventions
- Risk underestimated: can lead to withholding many important interventions

Calibration often not assessed during model validation.

So for many models, it is not known how accurate the risks are in a specific setting. In that case, you are in fact using a model with the lights off.



# The Achilles heel of predictive analytics

But if AUC is high, the ranking of patients into lower vs higher risk must be very good?

→ Good relative performance does not imply good absolute performance!

Using binary predictions only (e.g. treat vs don't treat), you are not avoiding the problem. I think you aggravate it by pretending to avoid the problem.





# First-Trimester Prognosis When an Early Gestational Sac is Seen on Ultrasound Imaging

Logistic Regression Prediction Model

Peter M. Doubilet, MD, PhD , Catherine H. Phillips, MD, Sara M. Durfee, MD, Carol B. Benson, MD

Published online in  
J Ultrasound Med  
on Aug 11 2020

**Objective: develop risk model for first trimester miscarriage in very early pregnancies**

- Retrospective data, single institution.
- 590 pregnancies, 345 miscarried; 9 parameters studied.
- Most important predictor (hCG rise) missing in 79%.
- No validation at all.

*“It might appear to be a weakness of our study that the first trimester loss rate was considerably higher than the rates found by other investigators (48% vs 10-30%). The rate is high because of the high prevalence of pregnancy risk factors in our population.”*

Web-calculator given that allows risk estimation. I cannot support that.

# How can risks be inaccurate?

- Methodological issues at model development or validation
  - Overfitting, leading to overly extreme risk estimates on new data
    - “in small datasets, it is reasonable for a model not to be developed at all”
  - Heterogeneity of measurement error between settings (Luijken et al, Stat Med 2019)
- Variables and characteristics unrelated to model development
  - Patient characteristics and outcome incidence/prevalence vary greatly between settings
  - Patient populations change over time within setting (“drift”)
  - So there is “Heterogeneity across time and place”

# Levels of calibration

Journal of Clinical Epidemiology 74 (2016) 167–176

A calibration hierarchy for risk models was defined: from utopia to empirical data

Ben Van Calster<sup>a,b,\*</sup>, Daan Nieboer<sup>b</sup>, Yvonne Vergouwe<sup>b</sup>, Bavo De Cock<sup>a</sup>, Michael J. Pencina<sup>c,d</sup>,  
Ewout W. Steyerberg<sup>b</sup>

1. Mean calibration / calibration-in-the-large
2. Weak calibration
3. Moderate calibration
4. Strong calibration

Work motivated by a very nice and thought provoking paper from Werner Vach (JCE 2013;66)

# 1. Mean calibration

The average estimated risk is accurate

Compare average risk with outcome prevalence/incidence



## 2. Weak calibration

On average, the model does not overestimate or underestimate risk, and does not give too extreme or too modest risks

‘Logistic recalibration’ framework:

Evaluate calibration intercept  $a$ :  $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = a + L$

$a < 0$  means overestimation,  $a > 0$  means underestimation

Evaluate calibration slope  $b$ :  $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = a + bL$

$b < 1$  means too extreme risks,  $b > 1$  means too modest risks

# 3. Moderate calibration

Observed proportion of events correspond to estimated risk

Construct a flexible calibration curve based on  $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = a + f(L)$ .

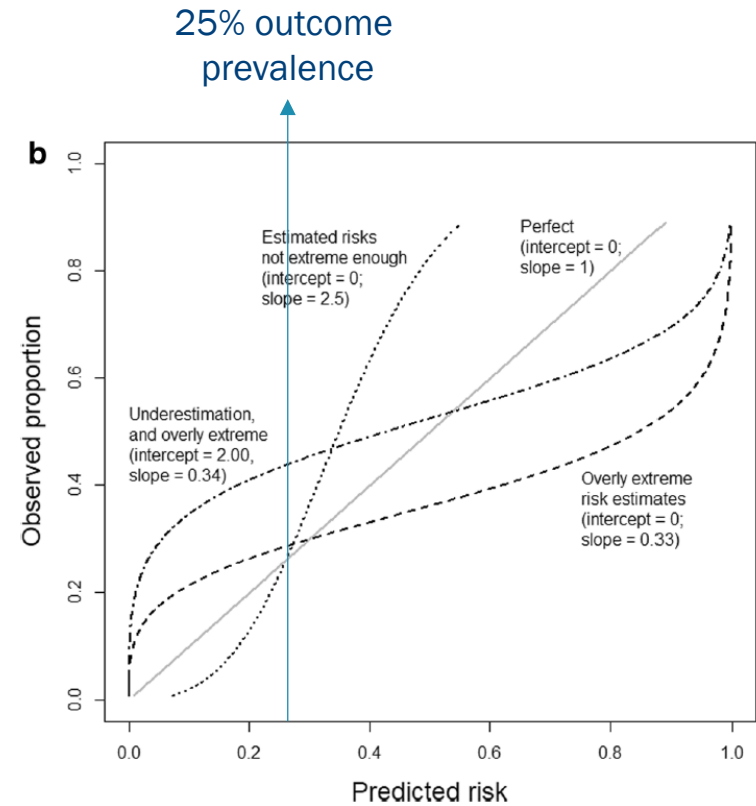
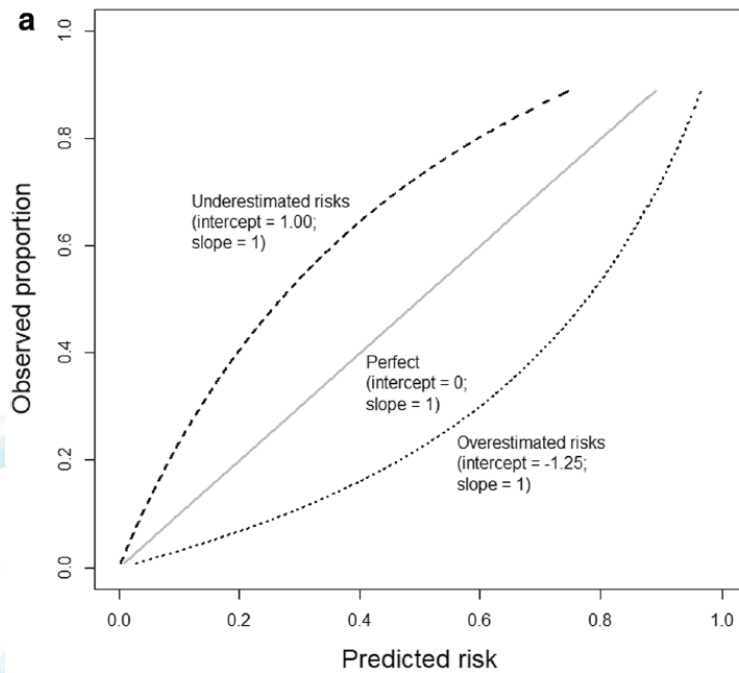
$f(\cdot)$  is usually a loess fit, but can also be based on splines.

This is preferable at external validation, but sufficient N needed.

Intercept and slope are nice summaries, but reduce calibration to 2 numbers (weak).

The slope is usually sufficient for internal validation (using bootstrapping or cross-validation), but the intercept or plotting a curve can sometimes be defended as well.

# Some reference calibration curves



# Example curves with low N

240 cases, 27 events (Caesarean delivery)

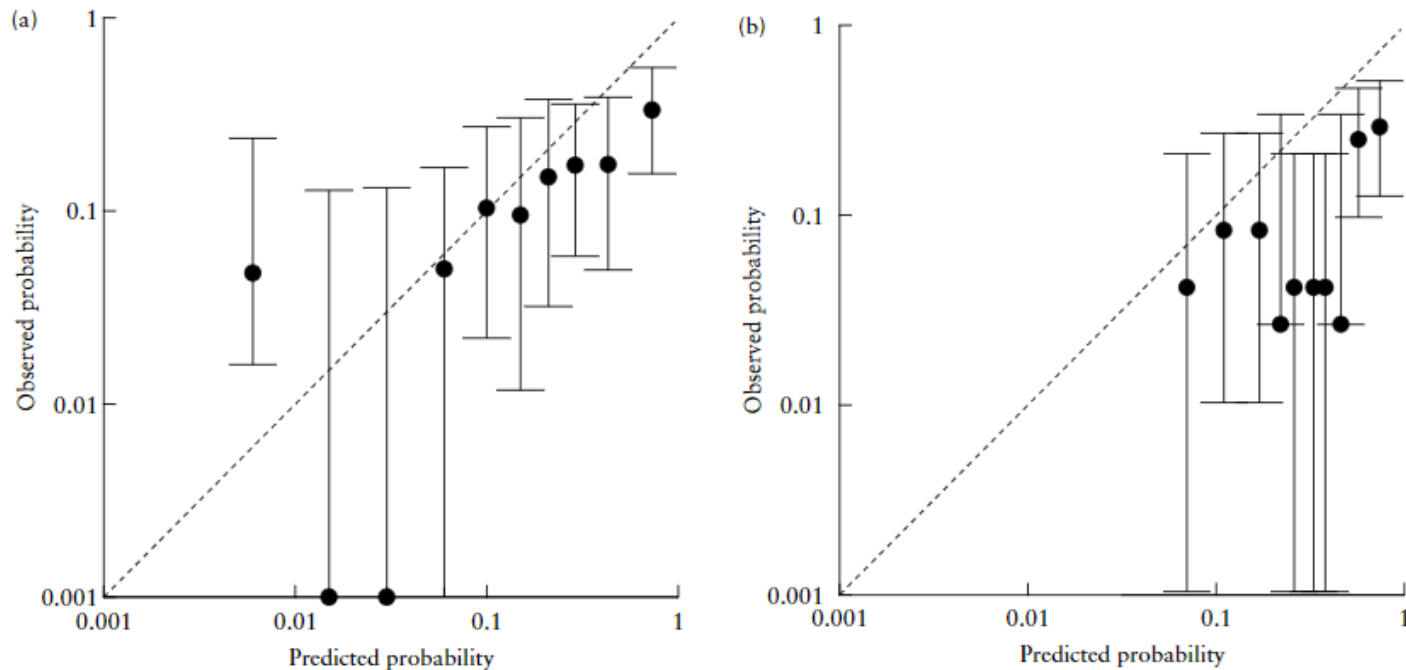


Figure 1 Calibration plots of the Peregrine *et al.* (a) and Rane *et al.* (b) prediction models for Cesarean delivery after induction of labor. Bars indicate 95% CIs of observed probability.

*“Calibration of the model on the right was not as good as the calibration of the model on the left”*



# 4. Strong calibration

Observed proportion of events correspond to estimated risk for each covariate pattern

Hard to assess (unless the model has only a few dichotomous predictors)

This is clinically desirable but **utopic**. The model needs to be fully correct.

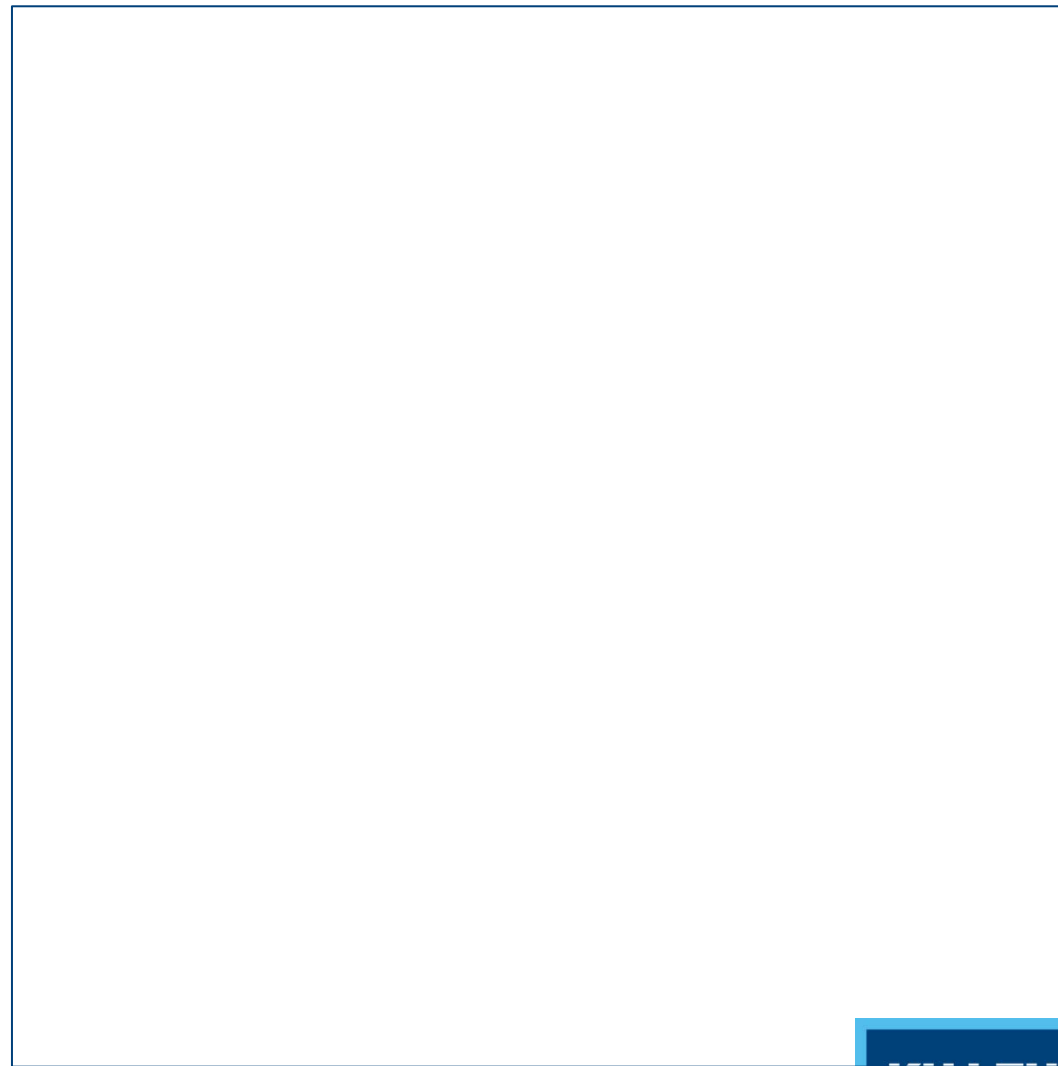
A diagonal calibration curve (i.e. moderate) does not imply strong calibration.

We have shown that moderate calibration cannot lead to harmful decisions (in the framework of decision curve analysis).

# Example external validation

Validation of models to diagnose ovarian cancer in patients managed surgically or conservatively: multicentre cohort study

N=4905, 978 events



# Multinomial outcomes?

1. Calibration intercepts and slopes can be calculated for multinomial logistic regression by extending the approach for binary outcomes to

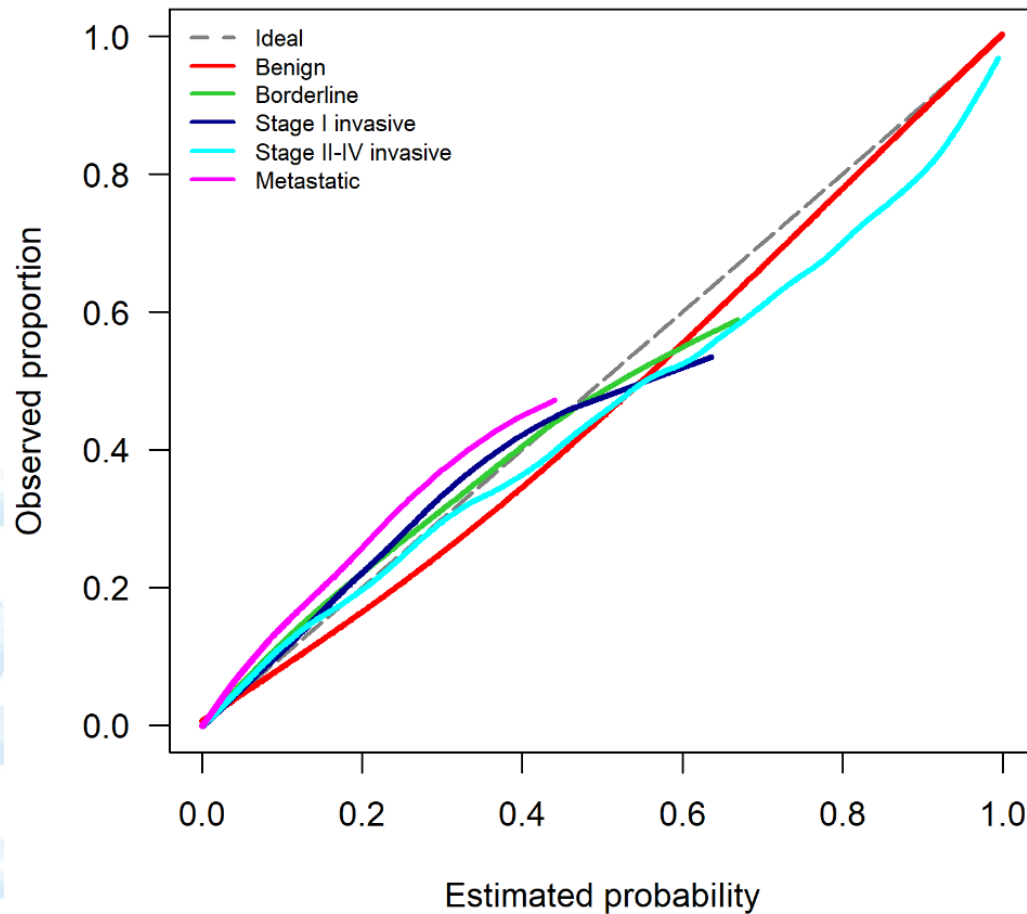
$$\log \left( \frac{P(Y = k)}{P(Y = J)} \right) = a_k + \sum_{i=1}^{K-1} b_{k,i} L_i$$

2. Flexible calibration curves can be obtained by using vector splines  $s(\cdot)$

$$\log \left( \frac{P(Y = k)}{P(Y = J)} \right) = a_k + \sum_{i=1}^{K-1} s_{k,i}(L_i)$$

This can be extended to risk models for ordinal outcomes, and to risk models based on e.g. machine learning algorithms

# Multinomial: example



# Heterogeneity between centers

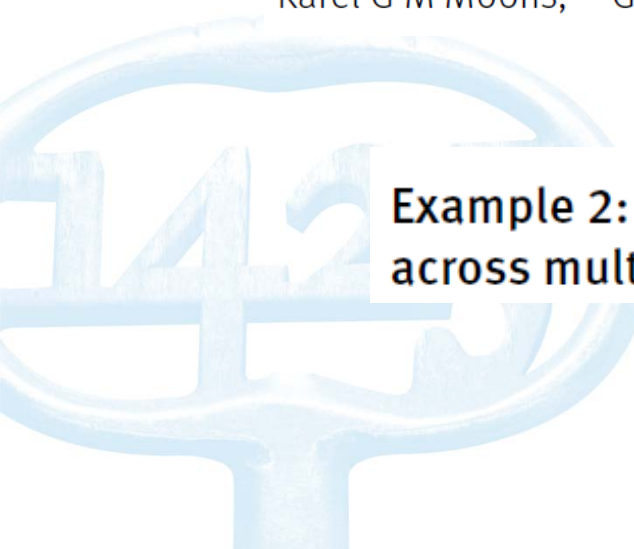
rithm can benefit from updating. Due to local healthcare systems and referral patterns, population differences between centers and regions are expected; it is likely that prediction models do not include all the predictors needed to accommodate these differences. Together with the phenomenon of population drifts, models ideally require continued monitoring in local settings in order to maximize their benefit over time. This argument will become even more vital with the growing popularity of highly flexible algorithms. The ultimate aim is to optimize the utility of predictive analytics for shared decision-making and patient counseling.

Calibration: the Achilles heel of predictive analytics

# Heterogeneity between centers

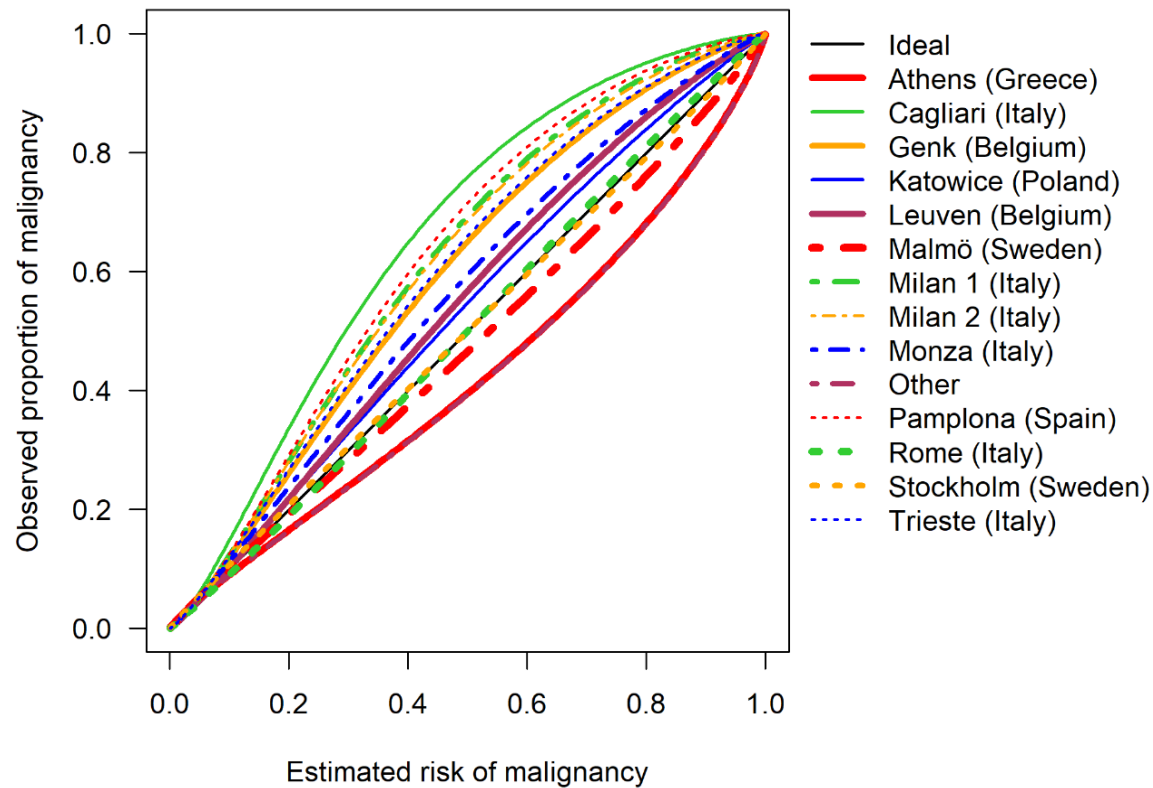
**External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges**

Richard D Riley,<sup>1</sup> Joie Ensor,<sup>1</sup> Kym I E Snell,<sup>2</sup> Thomas P A Debray,<sup>3,4</sup> Doug G Altman,<sup>5</sup> Karel G M Moons,<sup>3,4</sup> Gary S Collins<sup>5</sup>



**Example 2: Examining consistency in performance across multiple practices**

# Heterogeneity: example



# Heterogeneity: example

Centre-specific and overall logistic (i.e. non-flexible) calibration curves:  
logistic recalibration model with random intercept and random slope for the  $J$  centres (Wynants et al, SMMR 2018):

$$\log \left( \frac{P(Y=1)}{P(Y=0)} \right) = \alpha + a_j + \beta L + b_j L,$$

$$\text{where } \begin{bmatrix} a_j \\ b_j \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_a^2 & \tau_{ab} \\ \tau_{ab} & \tau_b^2 \end{bmatrix} \right).$$



# Cox models (TG6 paper in preparation)

**Assessing the performance of prediction models with a time-to-event outcome:  
guidance for validation and updating with a new prognostic factor**

## **Authors, contribution**

David J McLernon; statistical analysis in SAS and draft paper

Daniele Giardiello; to repeat analysis in R

Ben van Calster; input into paper plan and drafts

Laure Wynants; input into paper plan and drafts,

Maarten van Smeden; input into paper plan and drafts

Terry Therneau; comment from TG4 perspective

Ewout W Steyerberg; planned paper from TG6 and commented on structure thus far.  
on behalf of STRATOS

# Cox models

What you can do depends on the information you have (next to the validation dataset)

Level	Available information about the model
Level 1	Only model coefficients (very common)
Level 2	Coefficients + cumulative baseline hazard at $t_1$ , $H_0(t_1)$
Level 3	Original dataset

In my view, level 2 is what is needed for clinical application. It is also what TRIPOD recommends (Moons et al, Ann Intern Med 2005).



# Cox models

If  $H_0(t_1)$  is available, flexible adaptive hazard regression can be used to generate a flexible calibration curve at time  $t_1$

$$\log(h(t)) = g\left(\log\left(-\log(1 - p_{t_1})\right), t\right), \text{ with}$$

$$p_{t_1} = 1 - \left[\exp(-H_0(t_1))\right]^{\exp(\beta^T \mathbf{X})}$$

Can also be used for other time-to-event models.

See Austin, Harrell, van Klaveren (Stat Med 2020).

# 3 myths about risk thresholds (TG6 paper)

Wynants *et al. BMC Medicine* (2019) 17:192  
<https://doi.org/10.1186/s12916-019-1425-3>


BMC Medicine

OPINION

Open Access

## Three myths about risk thresholds for prediction models



Laure Wynants<sup>1,2\*</sup> , Maarten van Smeden<sup>3,4</sup>, David J. McLernon<sup>5</sup>, Dirk Timmerman<sup>1,6</sup>, Ewout W. Steyerberg<sup>4</sup>, Ben Van Calster<sup>1,4</sup> and on behalf of the Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative

KU LEUVEN

# 3 myths about risk thresholds

1. Risk groups are more useful than continuous risk estimates  
→ Clinically actionable groups (that have consensus) can make sense, but this remains rough for decision making at individual level
2. You can ask your statistician to get you the threshold  
→ Depends on clinical context, you need reasonable information on misclassification costs
3. The threshold is a part of the model  
→ Different preferences, different healthcare systems

These 3 issues are obviously related to each other.

# Further plans TG6

Practical guidance on validation of risk models for time-to-event outcomes

Practical guidance on validation of risk models accounting for competing risks

Simple paper (level 1) with advice for prediction model development

Multicenter diagnostic test evaluations: guidance on design and analysis

Hands-on tutorial of tools to assess calibration for different outcomes

“Medicine is a science of uncertainty and an art of probability”

William Osler

