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
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)
What is calibration?
What is calibration?

Non-systematic review of methodological literature

40 papers
What is calibration?

<table>
<thead>
<tr>
<th>Description</th>
<th>Percentage of papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>RISK EQUALS EVENT RATE PER LEVEL OF PREDICTED RISK</td>
<td>53%</td>
</tr>
<tr>
<td>VAGUE (AGREEMENT PREDICTED AND OBSERVED RISKS)</td>
<td>25%</td>
</tr>
<tr>
<td>AVERAGE RISK EQUALS EVENT RATE</td>
<td>13%</td>
</tr>
<tr>
<td>AVERAGE RISK EQUALS EVENT RATE IN SUBSETS</td>
<td>10%</td>
</tr>
<tr>
<td>RISK EQUALS EVENT RATE PER COVARIATE PATTERN</td>
<td>8%</td>
</tr>
<tr>
<td>NO DESCRIPTION PROVIDED</td>
<td>5%</td>
</tr>
</tbody>
</table>
Hierarchy of calibration

- Mean calibration
- Weak calibration
- Moderate calibration
- Strong 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: \{\(x_1\) \ldots \(x_p\)\}
Methods: logistic recalibration (Cox, Biometrika 1958)

- \( \text{logit}(Y) = a + b \times L \)

- \( b \) 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
Methods: logistic calibration curves

\[
\begin{align*}
  b < 1 & \quad a_{b=1} = 0 \\
  b > 1 & \quad a_{b=1} = 0 \\
  b = 1 & \quad a_{b=1} < 0 \\
  b = 1 & \quad a_{b=1} > 0
\end{align*}
\]
Methods: flexible calibration curves

- \( \text{logit}(Y) = a + f \times L \), 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
  o Compare average risk with event rate
  o 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
Simulated results
3. Moderate calibration

• Predicted risk equals event rate per level of predicted risk
  o Among patients with x% risk, x out of 100 have the event
  o 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!
Assessment of strong calibration

• Usually impossible: too few patients per covariate pattern

• Method that approaches the assessment of strong calibration:
  o Compare average predicted risk and event rate for subgroups of patients determined by values of one or more predictors
  o 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?
  
  o Correct model specification (e.g. GLM with logit link)
  o ML estimation only gives asymptotically unbiased estimates
  o Overfitting: combination of estimated model coefficients
  o All nonlinear effects and interaction effects are correctly modeled
  o (Systematic) measurement error

• Vach (2013): “the idea to identify the true model by statistical means is just a great wish which cannot be fulfilled”
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
  - $NB = \frac{TP - \text{odds}(T) \times FP}{N}$, 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
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)
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...
Sample size for validation

• Observed calibration curves will usually not be on diagonal

• Confidence intervals are important

• At least 200 events for flexible curves
Example

External validation of ADNEX model for ovarian tumor diagnosis
N=610, 182 events (Sayasneh et al, BJC in press)
val.prob.ci.2

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