ELSEVIER

Journal of Clinical Epidemiology ■ (2016) ■

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

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

Ben Van Calster[a,b,*], Daan Nieboer[b], Yvonne Vergouwe[b], Bavo De Cock[a], Michael J. Pencina[c,d], Ewout W. Steyerberg[b]

[a]*KU Leuven, Department of Development and Regeneration, Herestraat 49 Box 7003, 3000 Leuven, Belgium*
[b]*Department of Public Health, Erasmus MC, 's-Gravendijkwal 230, 3015 CE Rotterdam, The Netherlands*
[c]*Duke Clinical Research Institute, Duke University, 2400 Pratt Street, Durham, NC 27705, USA*
[d]*Department of Biostatistics and Bioinformatics, Duke University, 2424 Erwin Road, Durham, NC 27719, USA*

Accepted 23 December 2015; Published online xxxx

Abstract

Objective: Calibrated risk models are vital for valid decision support. We define four levels of calibration and describe implications for model development and external validation of predictions.

Study Design and Setting: We present results based on simulated data sets.

Results: A common definition of calibration is "having an event rate of *R%* among patients with a predicted risk of *R%*," which we refer to as "moderate calibration." Weaker forms of calibration only require the average predicted risk (mean calibration) or the average prediction effects (weak calibration) to be correct. "Strong calibration" requires that the event rate equals the predicted risk for every covariate pattern. This implies that the model is fully correct for the validation setting. We argue that this is unrealistic: the model type may be incorrect, the linear predictor is only asymptotically unbiased, and all nonlinear and interaction effects should be correctly modeled. In addition, we prove that moderate calibration guarantees nonharmful decision making. Finally, results indicate that a flexible assessment of calibration in small validation data sets is problematic.

Conclusion: Strong calibration is desirable for individualized decision support but unrealistic and counter productive by stimulating the development of overly complex models. Model development and external validation should focus on moderate calibration.

Keywords: Calibration; Decision curve analysis; External validation; Loess; Overfitting; Risk prediction models

1. Introduction

There is increasing attention for the use of risk prediction models to support medical decision making. Discriminatory performance is commonly the main focus in the evaluation of performance, whereas calibration often receives less attention [1]. A prediction model is calibrated in a given population if the predicted risks are reliable, that is, correspond to observed proportions of the event. Commonly, calibration is defined as "for patients with an predicted risk of *R%*, on average *R* out of 100 should indeed suffer from the disease or event of interest." Calibration is a pivotal aspect of model performance [2–4]: "For informing patients and medical decision making, calibration is the primary requirement" [2], "If the model is not […] well calibrated, it must be regarded as not having been validated […]. To evaluate classification performance […] is inappropriate" [4].

Recently, a stronger definition of calibration has been emphasized in contrast to the definition of calibration given previously [4,5]. Models are considered strongly calibrated if predicted risks are accurate for each and every covariate pattern. In this paper, we aim to define different levels of calibration and describe implications for model development, external validation of predictions, and clinical decision making. We focus on predicting binary end points (event vs. no event) and assume that a logistic regression model is developed in a derivation sample with performance assessment in a validation sample. We expand on examples used in recent work by Vach [5].

Funding: This study was supported in part by the Research Foundation Flanders (FWO) (grants G049312N and G0B4716N) and by Internal Funds KU Leuven (grant C24/15/037).

Conflict of interest: None.

* Corresponding author. Tel.: +32 16377788.
E-mail address: ben.vancalster@med.kuleuven.be (B. Van Calster).

http://dx.doi.org/10.1016/j.jclinepi.2015.12.005
0895-4356/

What is new?

Key findings

- We defined a new hierarchy of four increasingly strict levels of calibration, referred to as mean, weak, moderate, and strong calibration.
- Strong calibration of risk prediction models implies that the model was correct given the included predictors. We argue that this is unrealistic.
- Moderate calibration of risk prediction models guarantees that decision making based on the model does not lead to harm.
- The reliability of calibration assessments, most notably of flexible calibration plots, is highly dependent on the sample size of the validation data set.

What this adds to what was known?

- The evaluation of risk prediction models in terms of calibration is often described as a crucial aspect of model validation. However, a systematic framework for levels of calibration for risk prediction models was lacking, and the characteristics of different levels were unclear.
- We find that strong calibration of risk models occurs only in utopia, whereas moderate calibration does not and is sufficient from a decision-analytic point of view.

What is the implication and what should change now?

- At model development, researchers should not aim to develop the correct model. This is practically impossible and may backfire by developing overly complex models that overfit the available data. Our focus should be on achieving moderate calibration, for example, by controlling model complexity and shrinking predictions toward the average.
- At model validation, sufficiently large data sets should be available to reliably assess moderate calibration. We suggest a minimum of 200 events and 200 nonevents.

2. Assessing calibration at external validation

2.1. Methods

We assume that the predicted risks are obtained from a previously developed prediction model for outcome Y (1 = event, 0 = nonevent), for example, based on logistic regression analysis. The model provides a constant (model intercept) and a set of effects (model coefficients). The linear combination of the coefficients with the covariate values in a validation set defines the linear predictor L: $L=a+b_1\times x_1+b_2\times x_2+\ldots+b_i\times x_i$, where a is the model intercept, b_1 to b_i a set of regression coefficients, and x_1 to x_i the predictor values that define the patient's covariate pattern.

Calibration of risk predictions is often visualized in calibration plots. These plots show the observed proportion of events associated with a model's predicted risk [6]. Ideally the observed proportions in the validation set equal the predicted risks, resulting in a diagonal line in the plot. The observed proportions per level of predicted risk cannot be directly observed. We consider their estimation in three ways. First, the observed event rates can be obtained after categorizing the predicted risks, for example, using deciles. This is commonly done for the Hosmer-Lemeshow test [7]. Then, for each group, the average predicted risk can be plotted vs. the observed event rate to obtain a calibration curve, see [8] for an example. Second, the logistic recalibration framework can be used [9,10], where a logistic model is used for the outcome Y as a function of L. More technically, the logistic recalibration framework fits the following model: $\text{logit}(Y)=a+b_L\times L$. Using the results of this model to estimate the observed proportions results in a logistic calibration curve. If $b_L=1$ and $a=0$, the logistic calibration curve coincides with the diagonal line. The coefficient b_L is the calibration slope that gives an indication of the level of overfitting ($b_L<1$) or underfitting ($b_L>1$). Overfitting is most common, reflected in a linear predictor that gives too extreme values for the validation data: high risks are overestimated, and low risks are underestimated. The intercept a can be interpreted when fixing b_L at 1, that is, $a|b_L=1$. This calibration intercept is obtained by fitting the model $\text{logit}(Y)=a+\text{offset}(L)$, where the slope b_L is set to unity by entering L as an offset term to the model. Predicted risks are on average underestimated if $a|b_L=1>0$, and overestimated if $a|b_L=1<0$.

Third, a flexible, nonlinear, calibration curve can be considered using the model $\text{logit}(Y)=a+f(L)$. Here, f may be a continuous function of the linear predictor L, such as loess or spline transformations [6,11]. We used a loess smoother in this article.

2.2. Illustration: examples 1–5

For illustration, we consider five simulated examples, as previously presented [5]. We randomly generate four independent predictor variables x_1 to x_4. These predictor variables are ordinal with three categories (−1, 0, and 1) that each have 33% prevalence; this allows to visualize calibration by covariate pattern (see below). Let outcome Y be generated by an underlying logistic regression model with the true linear predictor