ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 1296−1301

Journal of Clinical Epidemiology

Calibration of clinical prediction rules does not just assess bias

Werner Vach*

Clinical Epidemiology, Institute of Medical Biometry and Medical Informatics, Freiburg University Medical Center, Freiburg, Germany

Accepted 2 June 2013; Published online 8 September 2013

Abstract

Objectives: Calibration is often thought to assess the bias of a clinical prediction rule. In particular, if the rule is based on a linear logistic model, it is often assumed that an overestimation of all coefficients results in a calibration slope less than 1 and an underestimation in a slope larger than 1.

Study Design and Setting: We investigate the relation of the bias and the residual variation of clinical prediction rules with the typical behavior of calibration plots and calibration slopes, using some artificial examples.

Results: Calibration is not only sensitive to the bias of the clinical prediction rule but also to the residual variation. In some circumstances, the effects may cancel out, resulting in a misleading perfect calibration.

Conclusion: Poor calibration is a clear indication of limited usefulness of a clinical prediction rule. However, a perfect calibration should be interpreted with care as this may happen even for a biased prediction rule.

Keywords: Bias; Calibration; External validation; Prognosis; Prognostic model; Residual variation

1. Introduction

Clinical prediction rules, also called prognostic models, are often the result of many efforts to use an available data set to select appropriate prognostic factors and develop a well-fitting model describing the relation between the factors and the event of interest. Hence, they often tend to be too extreme; that is, they overestimate the risk of high-risk patients and underestimate the risk of low-risk patients [1−3]. Moreover, the prognostic value of factors and their interrelation may vary even between similar patient populations, such that a rule may work well in one population but not necessarily in another [4−6]. For both reasons, clinical prediction rules should be validated in an external data set, which is not related to the data used to develop the prognostic model [7−10].

For the validation in an external data set, two basic principles are usually advocated and used in practice [8,11−13]: calibration and discrimination. Calibration aims to check whether the event probabilities according to the prediction rule coincide with the event rates that we can observe in the external validation data set [14,15]. Discrimination refers to the ability of the prognostic model to separate subjects with an event from subjects without an event. The latter is often approached by receiver operating characteristic (ROC) curves and related statistics like the c-index or the area under the ROC curve [1,16].

In this article, we focus on the first step, calibration. This is typically approached by a calibration plot, that is, some type of nonparametric regression relating the binary outcome Y to the probability values $\widehat{\pi}$ according to the prediction rule in the external validation data set. A popular choice is the division of the subjects into some risk groups according to the probability values from the prediction rule and plotting the relative frequency of $Y = 1$ against the mean of $\widehat{\pi}$ in each risk group. Sometimes, a smoothing method is used in addition. If a prediction rule is perfect, then the resulting points should be on the diagonal and the smoothed line should coincide with the diagonal. Any deviation from the diagonal indicates some imperfectness. Often, it is observed that the frequency of $Y = 1$ is smaller than that suggested by $\widehat{\pi}$ for high-risk patients and larger than that suggested by $\widehat{\pi}$ for low-risk patients, suggesting that the rule is indeed too extreme. This behavior can also be caught by computing a calibration slope $\widehat{\beta}_{\text{calib}}$ by a (logistic) regression of Y against $\widehat{\pi}$, with $\widehat{\beta}_{\text{calib}} < 1$ reflecting the situation of a rule with too extreme values. Janssen et al. [17] describe the typical interpretation of the calibration slope in the following way: ‘‘A calibration slope smaller than 1 indicates optimism; the regression coefficients of the original model were too large, which results in too

Conflict of interest: This material has neither been published nor is under consideration for publication elsewhere. There is no external funding involved in this study, and there are no possible conflicts of interest.

* Corresponding author. Tel.: +49 761 203 6722; fax: +49 761 203 6711.
E-mail address: wv@imbi.uni-freiburg.de

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2013.06.003

What is new?

- Calibration is often applied in the external validation of a clinical prediction rule.
- Poor calibration is often thought to reflect some bias of the prediction rule.
- Poor calibration can be due to variation or bias.
- Perfect calibration can occur even for a biased prediction rule.

extreme predictions in the new patients … A calibration slope that is larger than 1 indicates that the regression coefficients of the original model were too close to zero.''

In this article, we try to check whether calibration plots and calibration slopes are indeed useful to detect the type of bias described previously; that is, whether they correctly reflect a bias in the prediction rule. For this purpose, we assume in the external data set that the true-event probabilities follow a linear logistic model with known regression coefficients and investigate how certain choices of the regression coefficients for a prognostic model translate into patterns in the calibration plot and into certain values of the calibration slope.

2. Methods

We consider the artificial situation of an external validation study for which we know both the distribution of the covariates and the true model relating the binary outcome Y to the covariates. The four covariates X_1, …, X_4 are assumed to be independent and each taking the values -1, 0, and 1 with a probability of 1/3. The true model is assumed to be of a linear logistic type, that is, with $\pi_0(x_1, \ldots, x_4) = P(Y = 1|X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$ it can be written as

$$\text{logit}\pi_0(x_1,\ldots,x_4) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4.$$

The choice of the true values of the regression parameters is shown in the first line of Table 1, with effects of the covariates ranging from small effects of 0.2 to moderate effects in the magnitude of 0.8. Moreover, we consider seven different prediction rules $\widehat{\pi}_1,\ldots,\widehat{\pi}_7$, which have been developed in other studies. All these rules are again based on linear logistic models; that is, they can be expressed as

$$\text{logit}\widehat{\pi}_j(x_1,\ldots,x_4) = \widehat{\beta}_{0\text{j}} + \widehat{\beta}_{1j} x_1 + \widehat{\beta}_{2j} x_2 + \widehat{\beta}_{3j} x_3 + \widehat{\beta}_{4j} x_4.$$

The values of the regression coefficients chosen are shown in Table 1. For the first three choices, two coefficients are underestimated and two are overestimated, but with increasing magnitude of the difference to the true coefficients when moving from $\widehat{\pi}_1$ to $\widehat{\pi}_3$. In the forth choice, all coefficients are overestimated. The final three choices reflect the situation that all coefficients are underestimated. However, they differ in the variation of the extent of the underestimation. In $\widehat{\pi}_5$, two coefficients are estimated nearly correctly and two are estimated even with an incorrect sign, that is, with a bias of greater than 100%. In $\widehat{\pi}_6$, one is estimated correctly, two are underestimated by about 50%, and one is underestimated by 100%. In $\widehat{\pi}_7$, all coefficients are underestimated with a bias in the range between 25% and 75%.

For each of the seven prediction rules $\widehat{\pi}_j$, we consider the joint distribution of $\widehat{\pi}(X_1,\ldots,X_4)$ and $\pi_0(X_1,\ldots,X_4)$; that is, we consider for all $3^4 = 81$ possible values for $x_1,\ldots, x_4$ the pairs $[\widehat{\pi}(x_1,\ldots,x_4), \pi_0(x_1,\ldots,x_4)]$. We start with considering $\widehat{\pi}$ in dependence on π_0 in a corresponding scatter plot, and we fit a regression line to this scatter plot. As both for π_0 and all prediction rules, the average probability is close to 0.5, the bias of $\widehat{\pi}$ is directly described by the slope $\widehat{\beta}_{\text{bias}}$ of this regression line: a bias slope of 1 indicates no bias, a bias slope greater than 1 indicates a rule with too extreme probability values, and a bias slope less than 1 indicates a rule which is too pessimistic: the risk of high-risk patients is underestimated, and the risk of low-risk patients is overestimated. Then, we turn to the perspective of calibration; that is, we consider Y in dependence on $\widehat{\pi}$. For this, we draw a random sample of 1,600 observations following the true model, divide the values of $\widehat{\pi}$ into eight risk groups of equal size, and plot the observed frequency of $Y = 1$ in each risk group vs. the mean value of $\widehat{\pi}$ in each risk group. In addition, the lowess smoother [18] is used to obtain a smooth regression line. As the expectation of Y given $X_1 = x_1$, $X_2 = x_2$, $X_3 = x_3$, $X_4 = x_4$ is just given by $\pi_0(x_1, \ldots, x_4)$, we can study the typical behavior of calibration also by considering π_0 in dependence on $\widehat{\pi}$. Hence, we provide also a scatter plot of π_0 vs. $\widehat{\pi}$ together with the corresponding regression line. The slope $\widehat{\beta}_{\text{calib}}$ of this regression line describes the expectation for the slope of a regression line fitted to a calibration plot, and we refer to it as the calibration slope.

In the literature, calibration plots are typically presented on the probability scale, whereas calibration slopes are considered on the logit scale. As Appendix at www.jclinepi.com, we provide also calibration plots on the logit scale

Table 1. Regression coefficients in the true model and for seven clinical prediction rules

	Regression coefficients				
	β_0	β_1	β_2	β_3	β_4
True model π_0	0.0	0.21	0.37	0.64	0.77
Prediction rule $\widehat{\pi}_1$	0.0	0.25	0.30	0.51	0.92
Prediction rule $\widehat{\pi}_2$	0.0	0.32	0.19	0.32	1.16
Prediction rule $\widehat{\pi}_3$	0.0	0.40	0.04	−0.06	1.62
Prediction rule $\widehat{\pi}_4$	0.0	0.29	0.59	1.20	0.85
Prediction rule $\widehat{\pi}_5$	0.0	0.20	−0.09	−0.16	0.73
Prediction rule $\widehat{\pi}_6$	0.0	0.11	0.19	0.64	0.00
Prediction rule $\widehat{\pi}_7$	0.0	0.08	0.27	0.38	0.19