ELSEVIER

Journal of Clinical Epidemiology 63 (2010) 938–942

Journal of Clinical Epidemiology

LETTERS TO THE EDITOR

Use of Brier score to assess binary predictions

The use of the Brier score [1] in medical research to assess and compare the accuracy of binary predictions or prediction models is increasingly popular; see, for example Refs. [2–5]. In [6, Box 1] an overview of a variety of measures of model performance is offered and [7] propose cutoffs for appraising the value of a computed score. How Brier scores can be formally compared is detailed in Ref. [8].

Because of a growing number of applications and in light of the description in [9, p. 1253], we would like to briefly discuss the Brier score and its connection to Spiegelhalter's calibration test [10].

For n predictive probabilities $\boldsymbol{p} = (p_1,\ldots,p_n)$ with $0 \leq p_i \leq 1$ and n realizations $\boldsymbol{x} = (x_1,\ldots, x_n)$ of Bernoulli random variables $X_i \sim \text{Ber}(\pi_i)$ with $0 \leq \pi_i \leq 1$, $\pi = (\pi_1, \ldots, \pi_n)$ and $x_i \in \{0, 1\}$, the Brier score, defined as

$$B(\boldsymbol{p},\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n}(x_i - p_i)^2 = n^{-1}\sum_{i=1}^{n}(x_i - p_i)(1-2p_i) + n^{-1}\sum_{i=1}^{n}p_i(1-p_i), \quad (1)$$

equals the mean squared error of prediction. As a proper scoring rule, the Brier score simultaneously addresses calibration, the statistical consistency between the predicted probability and the observations as well as sharpness, which refers to the concentration of the predictive distribution, see Ref. [11]. This feature is also nicely illustrated by Murphy's decomposition of the Brier score, see Ref. [12]. When assessing the predictive accuracy of different binary predictions, for example, several logistic regression models, the Brier score can be used to compare model performances, see [6, Box 1]. Being mainly a relative measure, a lower score points to a superior model; the actual value of the score seems of limited value.

In the decomposition (1) the first summand has expectation 0 under perfect calibration, that is, if $\boldsymbol{p} = \pi$. This is exploited in the construction of Spiegelhalter's z-statistic [10,13] that enables formal assessment of calibration of binary predictions. The z-statistic is defined as

$$Z(\boldsymbol{p},\boldsymbol{x}) = \frac{\sum_{i=1}^{n}(x_i - p_i)(1-2p_i)}{\sqrt{\sum_{i=1}^{n}(1-2p_i)^2 p_i(1-p_i)}}. \quad (2)$$

The null hypothesis of calibration, that is, $\boldsymbol{p} = \pi$ is rejected at the significance level α if $|Z(\boldsymbol{p}, \boldsymbol{x})| > q_{1-\alpha/2}$, where q_α is the α-quantile of the standard normal distribution. This short summary makes it clear that

- calibration is neither *equal* to prediction error
- nor does a lower Brier score necessarily indicate better calibration,

as suggested by [9, p. 1253]. Indeed, suppose two Bernoulli experiments are performed and two forecasters issue predictions $\boldsymbol{p}_1 = (0.2,\ 0.2)$ and $\boldsymbol{p}_2 = (0.4,\ 0.5)$, and $\boldsymbol{x} = (1,\ 0)$ materializes. The resulting Brier scores and values of Spiegelhalter's z-statistic for these two competing models are provided in Table 1.

According to the value of the Brier score, the second model is to be preferred over the first; however, because $|Z(\boldsymbol{p}_2, \boldsymbol{x})| > |Z(\boldsymbol{p}_1, \boldsymbol{x})|$ this second model is *less* calibrated than the first. This simple example reveals that it is not generally true that a lower Brier score implies better model calibration. The reason is that the Brier score simultaneously addresses calibration *and* sharpness, see the discussion above. To exclusively address calibration, Spiegelhalter's z-test should be used. To us, it is therefore unclear what the purpose is of testing the Brier score on the value of 0, as indicated by [9, Table 2]. Instead, we hazard a guess that the authors actually intended to say that their models marked with a "*" in [9, Table 2] are not well *calibrated*. However, this does not correspond to have a Brier score value that is significantly different from 0 but Spiegelhalter's z-statistic different from 0.

To conclude, we advocate the use of the Brier score to assess predictive accuracy of binary prediction models, and we also agree that calibration of such models is an important issue and should be addressed when comparing models via the Brier score. With this short note, we intended to clarify some aspects when using these tools.

Table 1
Value of Brier score and Spiegelhalter's z-statistic for the two models and the realization $\boldsymbol{x} = (1, 0)$

Model	$\boldsymbol{p}$	$B(\boldsymbol{p}, \boldsymbol{x})$	$Z(\boldsymbol{p}, \boldsymbol{x})$
1	(0.2, 0.2)	0.34	1.06
2	(0.4, 0.5)	0.30	1.22

Kaspar Rufibach
Biostatistics Unit
Institute of Social and Preventive Medicine University of Zurich
Zurich, Switzerland
E-mail address: kaspar.rufibach@ifspm.uzh.ch (K. Rufibach)

References

[1] Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;78:1–3.
[2] Itoh S, Ikeda M, Mori Y, Suzuki K, Sawaki A, Iwano S, et al. Lung: feasibility of a method for changing tube current during low-dose helical CT. Radiology 2002;224:905–12.
[3] Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol 2005;58:475–83.
[4] Huo D, Senie RT, Daly M, Buys SS, Cummings S, Ogutha J, et al. Prediction of BRCA mutations using the BRCAPRO model in clinic-based African American, Hispanic, and other minority families in the United States. J Clin Oncol 2009;27:1184–90.
[5] Steyerberg EW. Clinical prediction models. New York, NY: Springer; 2009.
[6] Harrison DA, Brady AR, Parry GJ, Carpenter JR, Rowan K. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. Cri Care Med 2006;34:1378–88.
[7] Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774–81.
[8] Redelmeier DA, Bloch DA, Hickam DH. Assessing predictive accuracy: how to compare Brier scores. J Clin Epidemiol 1991;44:1141–6.
[9] Lix LM, Yogendran MS, Leslie WD, Shaw SY, Baumgartner R, Bowman C, et al. Using multiple data features improved the validity of osteoporosis case ascertainment from administrative databases. J Clin Epidemiol 2008;61:1250–60.
[10] Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986;5:421–33.
[11] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 2007;102:359–78.
[12] Murphy AH. Scalar and vector partitions of the probability score: Part I. Two-state situation. J Appl Meteorol 1972;11:273–82.
[13] StataCorp. STATA, reference A-F. Stata Corporation; 2003.

doi: 10.1016/j.jclinepi.2009.11.009

Brier score summarizes model calibration and discrimination - Reply

Thank you for the opportunity to respond to the comments of Dr Rufibach. In the article by Lix et al. [1], the *c*-statistic (equal to the area under the receiver operator characteristic curve for a binary outcome variable) and Brier score were used to evaluate algorithms for classifying osteoporosis cases and noncases identified from a bone mineral density database. The algorithms were constructed using a number of variables defined from hospital, physician, and prescription administrative databases. The Brier score provided a measure of the agreement between the observed binary outcome (i.e., case vs. noncase) and the predicted probability of that outcome. It is a sum of both a calibration component and a discrimination (or refinement) component [2,3], with lower scores indicating improved model accuracy.

Spiegelhalter's *z*-test [4] is used to evaluate the calibration component of the Brier score. This was not clearly described by Lix et al. [1]. The note to Table 2 should have indicated that values of the Brier score distinguished by a * were associated with a statistically significant value of Spiegelhalter's *z*-test (evaluated at $\alpha = 0.05$), indicating poor calibration. I appreciate Dr Rufibach's clarification of the interpretation of the study results.

Lisa M. Lix
School of Public Health
University of Saskatchewan
Saskatoon SK, Canada
E-mail address: lisa.lix@usask.ca (L.M. Lix)

References

[1] Lix LM, Yogendran MS, Leslie WD, Shaw SY, Baumgartner R, Bowman C, et al. Using multiple data features improved the validity of osteoporosis case ascertainment from administrative data. J Clin Epidemiol 2008;61:1250–60.
[2] Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;78:1–3.
[3] Blattenberger G, Lad F. Separating the Brier score into calibration and refinement components: a graphical exposition. Am Stat 1985;39:26–32.
[4] Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986;5:421–33.

doi: 10.1016/j.jclinepi.2009.11.008

Testing for baseline balance: Can we finally get it right?

To the Editor:

Austin et al. [1] followed in the footsteps of other authors in condemning hypothesis testing of baseline covariates as inappropriate and illogical. Unfortunately, the pied piper does not always lead us in the right direction; hence we need diligence to determine which advice to follow. Citing 161 references would appear to cover all the bases, but the key publications pointing out the flaw in the logic of the aforementioned argument—including one by Swingler and Zwarenstein [2] in this very journal—were curiously missing. The conclusion that baseline testing is illogical is predicated on the notion that all outcomes of the randomization are equally likely or, if restrictions are used, then with 1:1 allocation (equal group sizes) any outcome is exactly as likely as its mirror image. This condition, if true, would ensure balance among the population of randomization outcomes and would render formal baseline testing illogical.