ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII^th SISMEC Congress - Vol. 20 Suppl. 1

Improving Calibration Assessment near Clinical Thresholds: The Bayesian Calibration Error

Zamagni Giulia[1,2], Barbati Giulia[1]

(1) Unità di Biostatistica, Dipartimento di Scienze Mediche e Chirurgiche, Università di Trieste, Trieste, Italia
(2) SCR Epidemiologia Clinica e Ricerca sui Servizi Sanitari – IRCCS Burlo Garofolo, Trieste, Italy

CORRESPONDING AUTHOR: *Zamagni Giulia, giulia.zamagni@burlo.trieste.it*

INTRODUCTION

Calibration of predictive models is essential to ensure the clinical reliability of risk estimates, particularly when decisions are based on well-defined probability thresholds. However, especially in machine learning (ML) applications, calibration is often overlooked, and model performance is typically evaluated using discrimination metrics alone [1,2]. Several calibration metrics have been proposed, including the Brier Score, Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and Integrated Calibration Index (ICI). Each of these has limitations: for example, the Brier Score reflects a global average and may mask local errors; ECE and MCE are highly sensitive to binning strategies and become unstable with limited data; the ICI, while more robust, does not focus specifically on clinically relevant thresholds [3–5]. As a result, these metrics may fail to detect or emphasize calibration errors in the areas most critical for clinical decision-making.

OBJECTIVES

To introduce the Bayesian Calibration Error (BCE), a metric that quantifies both the magnitude and concentration of miscalibration around a clinically relevant threshold, and to evaluate its use alongside the Absolute Calibration Error (ACE).

METHODS

BCE integrates three components: (i) quantile-based adaptive binning, (ii) a Bayesian formulation to estimate local calibration error (LCE), which accounts for the number of events in each bin rather than relying solely on observed proportions, and (iii) a Gaussian weighting function centered around the decision threshold t. For each bin i, the mean predicted probability p_{pred_i} is compared with the expected value of the observed frequency, modeled using a non-informative Beta(1,1) prior. The posterior distribution becomes Beta(α_{post} = 1 + k, β_{post} = 1 + n – k), where k is the number of events and n is the number of observations in the bin. The local calibration error (LCE) is then defined as:

$$LCE_i = \left| p_{pred_i} - E[Beta(1+k, 1+n-k)] \right|.$$

After defining a decision threshold t (i.e., a predicted probability associated with a clinical "action"), derived through decision curve analysis and/or clinician input, a Gaussian weight is assigned to each bin:

$$w_i = \exp\left(-\frac{(p_{pred_i} - t)^2}{2\sigma^2}\right),$$

where σ (e.g., 0.1) controls the concentration around the threshold Weights are normalized to have unit mean. BCE is then computed as the weighted average of the LCEs. A high BCE indicates that miscalibration is particularly concentrated around the threshold.

We applied this approach to a dataset of 3,672 pregnant women carrying small-for-gestational-age (SGA) fetuses, enrolled in the TRUFFLE 2 multicenter study. Three predictive models were developed—Logistic Regression (LR), Random Forest (RF), and XGBoost—using 11 routine clinical variables to predict adverse perinatal outcomes. The decision threshold was set at t = 0.3 based on prior decision analyses.

DOI: 10.54103/2282-0930/29534

RESULTS

The incidence of adverse outcomes was 13%. ACE confirmed the same performance ranking across models (LR: 0.0198, RF: 0.1126, XGBoost: 0.2290). However, BCE imposed a stricter penalty on RF (BCE = 0.1916) and an even higher one on XGBoost (BCE = 0.2633), indicating that miscalibration was concentrated around the decision threshold. Although the RF model showed a more pronounced local peak of miscalibration, XGBoost had a broader spread of error in bins adjacent to the threshold, resulting in a higher overall BCE. Conversely, the LR model maintained a low BCE (0.0216), suggesting good local calibration.

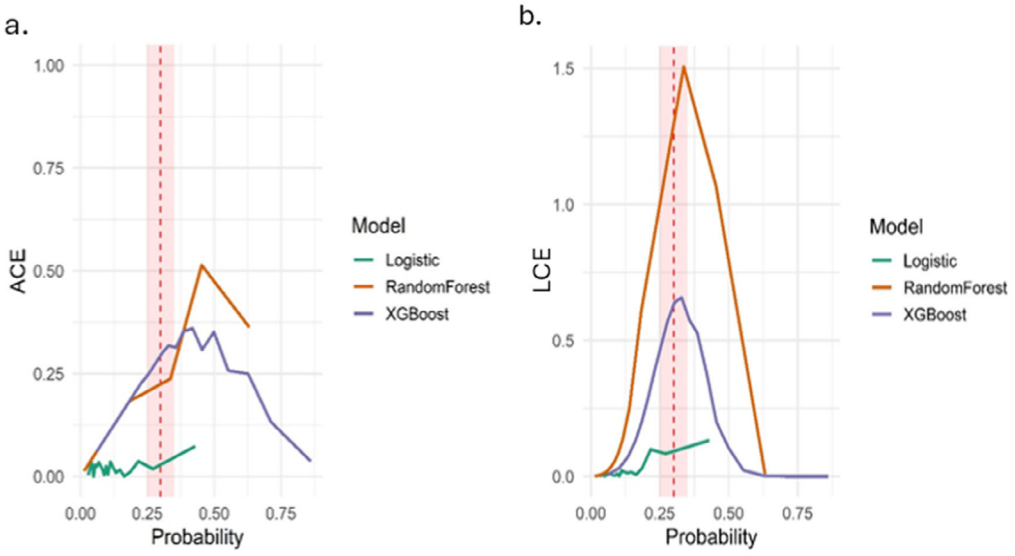
CONCLUSIONS

BCE complements global calibration metrics by quantifying whether miscalibration is concentrated around the clinical decision threshold. While ACE reflects the average accuracy of risk estimates across the entire prediction range, BCE captures local consistency near the threshold, offering a more nuanced evaluation. This distinction is particularly important in clinical contexts, where decisions hinge on specific risk cutoffs. When a clinical "action" threshold is defined, we recommend reporting both ACE and BCE to support informed model assessment. Moreover, BCE enables identification of models that, despite satisfactory global calibration, underperform near the decision threshold—and conversely, models with less favorable global performance that maintain adequate reliability in clinically critical regions.

REFERENCES

1. Kleinrouweler CE, Cheong-See FM, Collins G.S. et al. Prognostic models in obstetrics: available, but far from applicable. Am J Obstet Gynecol. 2016;214:79–90.
2. Christodoulou E., Ma J., Collins G.S. et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22.
3. Barreñada, L., Dhiman, P., Timmerman, D. et al. Understanding overfitting in random forest for probability estimation: a visualization and simulation study. Diagn Progn Res 8, 14 (2024). https://doi.org/10.1186/s41512-024-00177-1.
4. Nixon J., Dusenberry M.W., Zhang L. et al.; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019, pp. 38-41.
5. Austin P.C., Steyerberg E.W. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in Medicine. 2019; 38: 4051–4065. https://doi.org/10.1002/sim.8281.

Figure 1. Absolute Calibration Error (ACE) and (b) Local Calibration Error (LCE), weighted by proximity to the decision threshold. The red dashed line indicates the t=0.3 decision threshold, with the shaded red area representing the critical region defined by σ=0.1



DOI: 10.54103/2282-0930/29534