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

Integrating Calibration into the Evaluation of Clinical Utility: A Proposal for a Weighted Net Benefit

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

Decision Curve Analysis (DCA) is a widely used framework for evaluating diagnostic and prognostic strategies, as it explicitly incorporates the clinical consequences associated with decision-making [1]. Within this framework, Net Benefit (NB) is a key indicator of the clinical utility of a predictive model. However, the original NB formulation does not account for model calibration, despite strong evidence that poor calibration systematically reduces clinical utility [2]. In real-world settings, many predictive models, particularly those based on machine learning (ML), often show suboptimal global calibration compared to traditional statistical models, due to their greater complexity and susceptibility to overfitting.

Therefore, when focusing on a specific probability threshold as a decision point, it becomes crucial to evaluate calibration in the vicinity of that threshold. This targeted approach may lead to a different comparative assessment of the clinical potential of various predictive algorithms.

OBJECTIVES

To propose a pragmatic extension of DCA based on the concept of Weighted Net Benefit (WNB), in which a model's utility is penalized more heavily for calibration errors in the decision-making region. This avoids discarding models that, while globally less calibrated, are reliable near the clinical threshold.

METHODS

The framework involves four steps:

i) the decision threshold is defined a priori through DCA and/or clinical consultation;

ii) calibration is assessed around the threshold using the Bayesian Calibration Error (BCE);

iii) the weighted posterior standard deviation (wSD) is computed to quantify the statistical uncertainty associated with local calibration error (LCE) estimates;

iv) NB at the predetermined threshold is adjusted according to the following formula:

$$WNB = \frac{1}{1+BCE+wSD} NB.$$

Specifically, the BCE is defined as the weighted mean of local calibration errors (LCE), calculated as the absolute difference between the average predicted probability in each bin and the Bayesian estimate of the observed event rate, computed as the posterior mean of a Beta(1 + k, 1 + n − k) distribution, where k is the number of events and n the total number of observations in the bin. Each LCE is weighted using a Gaussian function centered on the decision threshold, thus emphasizing calibration errors in clinically critical regions.

The wSD is calculated as the weighted mean of the standard deviations of the posterior Beta distributions in each bin, using the same weighting function. This penalizes models not only for local miscalibration but also for greater

DOI: 10.54103/2282-0930/29535

statistical uncertainty in the estimation of calibration error near the decision threshold [3]. In this way, the WNB provides a more cautious and context-aware estimate of clinical utility, allowing recognition of models that, despite suboptimal global calibration, provide reliable predictions around the decision threshold.

We applied this framework to a clinical dataset of 3,672 pregnancies with small-for-gestational-age fetuses collected in the multicenter TRUFFLE 2 study. Three predictive models—logistic regression (LR), Random Forest (RF), and XGBoost—were developed using 11 clinical variables to predict adverse perinatal outcomes. NB and WNB were calculated at the decision threshold t = 0.3.

Additionally, to assess prediction instability, NB and WNB were computed as the mean and standard deviation over 500 bootstrap replicates at the clinical threshold.

RESULTS

The incidence of adverse outcomes was 13%.

RF achieved the best discrimination after bootstrap optimism correction (AUROC = 0.91, 95% CI: 0.85–0.94), while LR showed the poorest performance (AUROC = 0.71, 95% CI: 0.67–0.75). XGBoost had intermediate performance (AUROC = 0.83, 95% CI: 0.74–0.88).

At the threshold t = 0.3, LR showed a low NB (0.02 ± 0.001) and an even lower WNB (0.01 ± 0.001), reflecting limited clinical utility at the threshold despite near-optimal global calibration.

RF yielded the highest NB (0.08 ± 0.002), though it received the strongest penalty (WNB = 0.04 ± 0.001), while XGBoost displayed intermediate behavior (NB = 0.05 ± 0.003; WNB = 0.03 ± 0.002).

These results suggest that, when discrimination is high, even a model with suboptimal calibration may retain clinical utility—if the expected decision quality near the critical region compensates for calibration penalties.

Low variability across bootstrap samples (SD ≤ 0.003) indicates that both NB and WNB are highly stable around the clinical threshold t = 0.3 for all models (Figure 1).

CONCLUSIONS

The WNB offers a tailored evaluation of the clinical utility of predictive models by incorporating local calibration information near the decision threshold.

This approach is not intended to replace traditional DCA but rather to serve as a methodologically coherent extension—particularly relevant when comparing predictive algorithms of varying complexity in terms of calibration performance.

In clinical contexts where decisions are based on well-defined thresholds, WNB can support more informed, reliable, and decision-centered evaluations.

REFERENCES

1. Vickers A.J., Elkin E.B. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006 Nov-Dec;26(6):565-74. doi: 10.1177/0272989X06295361. PMID: 17099194; PMCID: PMC2577036.
2. Van Calster B, Vickers AJ. Calibration of Risk Prediction Models: Impact on Decision-Analytic Performance. Medical Decision Making. 2014;35(2):162-169. doi:10.1177/0272989X14547233.
3. Zamagni G, Barbati G. Improving calibration assessment near clinical thresholds: the bayesian calibration error. Abstract SISMEC 2025

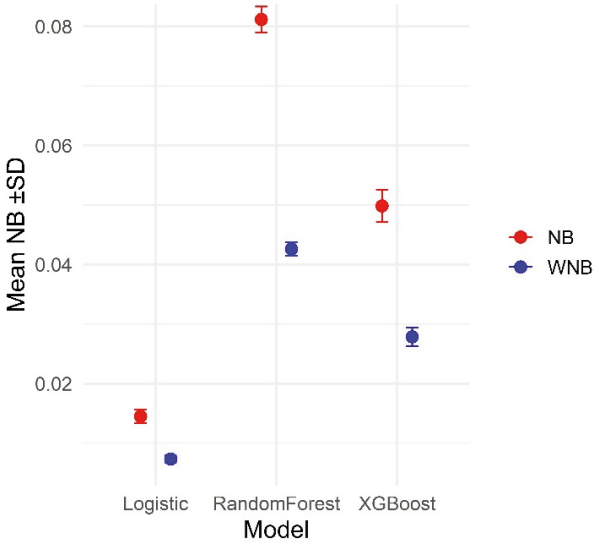

Figura 1. Media e deviazione standard di NB e WNB nei 500 campioni di bootstrap per modello

DOI: 10.54103/2282-0930/29535