Journal of Biomedical Informatics 54 (2015) 283–293

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# A spline-based tool to assess and visualize the calibration of multiclass risk predictions

K. Van Hoorde [a,b], S. Van Huffel [a,b], D. Timmerman [c,d], T. Bourne [c,d,e], B. Van Calster [d,*]

[a] KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, Leuven, Belgium
[b] KU Leuven, iMinds Medical Information Technologies, Leuven, Belgium
[c] Department of Obstetrics & Gynecology, University Hospitals Leuven, Leuven, Belgium
[d] KU Leuven, Department of Development & Regeneration, Leuven, Belgium
[e] Queen Charlotte's & Chelsea Hospital, Imperial College, Du Cane Road, London W12 0HS, UK

ARTICLE INFO

*Article history:*
Received 21 October 2014
Accepted 30 December 2014
Available online 9 January 2015

*Keywords:*
Risk models
Probability estimation
Machine learning
Logistic regression
Calibration
Multiclass

ABSTRACT

When validating risk models (or probabilistic classifiers), calibration is often overlooked. Calibration refers to the reliability of the predicted risks, i.e. whether the predicted risks correspond to observed probabilities. In medical applications this is important because treatment decisions often rely on the estimated risk of disease. The aim of this paper is to present generic tools to assess the calibration of multiclass risk models.

We describe a calibration framework based on a vector spline multinomial logistic regression model. This framework can be used to generate calibration plots and calculate the estimated calibration index (ECI) to quantify lack of calibration. We illustrate these tools in relation to risk models used to characterize ovarian tumors. The outcome of the study is the surgical stage of the tumor when relevant and the final histological outcome, which is divided into five classes: benign, borderline malignant, stage I, stage II–IV, and secondary metastatic cancer. The 5909 patients included in the study are randomly split into equally large training and test sets. We developed and tested models using the following algorithms: logistic regression, support vector machines, k nearest neighbors, random forest, naive Bayes and nearest shrunken centroids.

Multiclass calibration plots are interesting as an approach to visualizing the reliability of predicted risks. The ECI is a convenient tool for comparing models, but is less informative and interpretable than calibration plots. In our case study, logistic regression and random forest showed the highest degree of calibration, and the naive Bayes the lowest.



## 1. Introduction

For medical applications, prediction models that provide probabilistic (risk) estimates of an event of interest are useful for clinical decision support, personalized healthcare, and shared decision making. Prior to the implementation of such tools in clinical practice, validation with respect to discrimination and calibration is required [1–6]. A model should be able to distinguish between different possible outcome categories (discrimination). This can be evaluated using the area under the receiver operating characteristic curve (AUC) or multiclass extensions of this approach. Calibration assessment is often overlooked, but is of importance for several applications where risk models may be used. Such applications include decisions whether or not to treat a patient [7], start preventive action, or to inform the choice of treatment [8]. Calibration is also relevant when informing patients about risk [9], when comparing hospitals with respect to quality of care (e.g. benchmarking based on mortality risk) [10], and when identifying high risk patients for inclusion in clinical trials [11]. The optimal use of risk models in these situations

*Abbreviations:* AUC, Area under the ROC curve; CI, Confidence Interval; ECI, estimated calibration index; IOTA, International Ovarian Tumor Analysis; IQR, InterQuartile Range; kNN, k nearest neighbor; LR-BT, Logistic Regression Binary Tree - sequential dichotomous model to Binary Tree sequential dichotomous Logistic Regression; LR-PC, Pairwise Coupling Logistic Regression; MLR, multinomial logistic regression; NB, Naive Bayes; NSC, nearest shrunken centroids; PDI, Polytomous Discrimination Index; RF, Random Forest; ROC, Receiver Operating Characteristic; SVM, Support Vector Machines; SVM-BT, Support Vector Machines Binary Tree - sequential dichotomous model to Binary Tree sequential dichotomous Support Vector Machines; SVM-PC, Pairwise Coupling Support Vector Machines.

* Corresponding author at: KU Leuven, Department of Development & Regeneration, Herestraat 49 Box 7003, B3000 Leuven, Belgium.
*E-mail address:* ben.vancalster@med.kuleuven.be (B. Van Calster).

http://dx.doi.org/10.1016/j.jbi.2014.12.016
1532-0464/

relies on reliable risk estimation. For example, a classic result from decision analysis states that the adopted risk threshold to decide whether or not to take further action implies specific misclassification costs [12]: the odds of the risk threshold equals the ratio of the harm of a false positive test result to the benefit of a true positive result. For example if a risk threshold of 10% is adopted, the assumption is that 1 true positive is worth 9 false positives. If a poorly calibrated risk model is then used to assess whether patients exceed the planned threshold, inappropriate decisions may be taken.

For binary outcomes, the relationship between predicted and observed probabilities can be visualized by means of a calibration plot [1,13,14]. Observed probabilities are sometimes obtained by computing event rates within groups of patients with similar predicted probabilities (e.g. decile split). However, often flexible smoothing methods such as local regression (loess) or splines are used to link predicted probabilities to estimated observed probabilities [1].

Recently, our group extended binary calibration plots to multiclass models based on multinomial logistic regression (MLR) [15]. We proposed two frameworks, one parametric and one non-parametric. Logistic regression is a common algorithm to build binary and multiclass clinical prediction models, and naturally works with risk estimates. However, machine learning algorithms are also used for clinical risk prediction [16–21], and are very frequently used in high dimensional and/or "large p, small n" prediction studies (i.e. a large number of predictors and a small number of patients) [22–24]. Moreover, although using machine-learning approaches for classification problems is often less suited to probability estimation, methods do exist to facilitate this [25–30]. The calibration performance of risk models is an issue that is often neglected, and it is not surprising that with a few exceptions this is frequently the case for models based on machine learning algorithms [13,26,27,30–32].

The aim of this paper is to introduce a non-parametric framework to evaluate the calibration of multiclass risk models irrespective of the modeling technique used. Based on this framework we also derive a calibration measure to quantify and compare calibration performance between models. We illustrate these methods with a case study looking at the classification of ovarian tumors. We develop and validate risk models to diagnose tumor pathology based on logistic regression, support vector machines, k-nearest neighbors, random forest, naive Bayes and nearest shrunken centroids.

## 2. Non-parametric recalibration framework

Our group developed calibration tools for risk models based on multinomial logistic regression (MLR) [15]. Assume an MLR or 'baseline-category logit' model [33] with $m$ predictors $x_1$ to $x_m$ for an outcome with $J$ ($j = 1, \ldots, J$) categories. If category 1 is chosen as the reference category, the model is written as

$$\begin{cases} \log\left[\frac{P(Y=2)}{P(Y=1)}\right] = \alpha_2 + \sum_{l=1}^{m}\beta_{2l}x_l = \mathrm{lp}_{21} \\ \log\left[\frac{P(Y=3)}{P(Y=1)}\right] = \alpha_3 + \sum_{l=1}^{m}\beta_{3l}x_l = \mathrm{lp}_{31} \\ \ldots \\ \log\left[\frac{P(Y=J)}{P(Y=1)}\right] = \alpha_J + \sum_{l=1}^{m}\beta_{Jl}x_l = \mathrm{lp}_{J1} \end{cases} \quad (1)$$

and the multiclass risks are obtained as

$$\begin{cases} P(Y=1) = p_1 = \frac{1}{1+\exp(\mathrm{lp}_{21})+\exp(\mathrm{lp}_{31})+\cdots+\exp(\mathrm{lp}_{J1})} = \frac{1}{1+\sum_{j=2}^{J}\exp(\mathrm{lp}_{j1})} \\ P(Y=2) = p_2 = \frac{\exp(\mathrm{lp}_{21})}{1+\exp(\mathrm{lp}_{21})+\exp(\mathrm{lp}_{31})+\cdots+\exp(\mathrm{lp}_{J1})} = \frac{\exp(\mathrm{lp}_{21})}{1+\sum_{j=2}^{J}\exp(\mathrm{lp}_{j1})} \\ \ldots \\ P(Y=J) = p_J = \frac{\exp(\mathrm{lp}_{J1})}{1+\exp(\mathrm{lp}_{21})+\exp(\mathrm{lp}_{31})+\cdots+\exp(\mathrm{lp}_{J1})} = \frac{\exp(\mathrm{lp}_{J1})}{1+\sum_{j=2}^{J}\exp(\mathrm{lp}_{j1})} \end{cases} \quad (2)$$

Let $\left\{\hat{\mathrm{lp}}_{21}, \ldots, \hat{\mathrm{lp}}_{J1}\right\}$ denote the estimated linear predictors and $\{\hat{p}_1, \ldots, \hat{p}_J\}$ the estimated multiclass risks. The non-parametric recalibration framework for such models relates the multiclass outcome $Y$ on the estimated $J-1$ linear predictors $\left\{\hat{\mathrm{lp}}_{21}, \ldots, \hat{\mathrm{lp}}_{J1}\right\}$ from the MLR risk model through a vector spline [34] MLR analysis [15]:

$$\begin{cases} \log[P(Y=2)/P(Y=1)] = a_2 + \sum_{j=2}^{J}\left(b_{2j}\cdot s_2\left(\hat{\mathrm{lp}}_{j1}\right)\right) \\ \log[P(Y=3)/P(Y=1)] = a_3 + \sum_{j=2}^{J}\left(b_{3j}\cdot s_3\left(\hat{\mathrm{lp}}_{j1}\right)\right) \\ \ldots \\ \log[P(Y=J)/P(Y=1)] = a_J + \sum_{j=2}^{J}\left(b_{Jj}\cdot s_J\left(\hat{\mathrm{lp}}_{j1}\right)\right) \end{cases} \quad (3)$$

with $s(\cdot) = [s_2(\cdot), s_3(\cdot), \ldots, s_J(\cdot)]$ a vector spline smoother applied to each linear predictor [15,34]. This vector spline smoother $s(\cdot)$ is a natural extension of the cubic spline smoother to vector responses and

**Table 1**
Descriptive statistics of the ovarian tumor case study.

| | Benign | Borderline | Stage I | Stage II–IV | Metastatic |
|---|---|---|---|---|---|
| *Outcome, N* | 3980 | 339 | 356 | 988 | 246 |
| *Variable, N (%) or median (IQR)* | | | | | |
| Age (years) | 42 (32–54) | 49 (36–62) | 54 (44–64) | 59 (50–67) | 57 (47–68) |
| Serum CA125 (U/mL)[a] | 19 (11–39) | 31 (16–100) | 52 (21–190) | 447 (147–1215) | 81 (30–271) |
| Family history of ovarian cancer | 79 (2.0) | 10 (3.0) | 13 (3.7) | 57 (5.8) | 5 (2.0) |
| Maximal diameter of lesion (mm) | 63 (45–87) | 86 (51–150) | 106 (71–153) | 85 (56–123) | 86 (56–124) |
| Solid tissue | | | | | |
| Presence of solid tissue | 1322 (33.2) | 267 (78.8) | 328 (92.1) | 968 (98.0) | 234 (95.1) |
| Proportion solid tissue if present (%) | 42 (20–100) | 37 (24–59) | 61 (38–100) | 100 (56–100) | 100 (64–100) |
| Number of papillary projections | | | | | |
| None | 3424 (86.0) | 135 (39.8) | 227 (63.8) | 772 (78.1) | 213 (86.6) |
| 1 | 333 (8.4) | 69 (20.4) | 25 (7.0) | 56 (5.7) | 12 (4.9) |
| 2 | 80 (2.0) | 21 (6.2) | 17 (4.8) | 30 (3.0) | 0 (0) |
| 3 | 66 (1.7) | 24 (7.1) | 17 (4.8) | 28 (2.8) | 2 (0.8) |
| >3 | 77 (1.9) | 90 (26.5) | 70 (19.7) | 102 (10.3) | 19 (7.7) |
| More than 10 cyst locules | 199 (5.0) | 74 (21.8) | 69 (19.4) | 93 (9.4) | 36 (14.6) |
| Acoustic shadows | 676 (17.0) | 8 (2.4) | 18 (5.1) | 30 (3.0) | 10 (4.1) |
| Ascites | 64 (1.6) | 28 (8.3) | 65 (18.3) | 473 (47.9) | 90 (36.6) |
| *Missing values for CA125, N (%)* | 1447 (36.4) | 62 (18.3) | 71 (19.9) | 163 (16.5) | 62 (25.2) |

Abbreviations: IQR; InterQuartile Range.
[a] Results for Serum CA125 are based on single imputation of missing values.