Received: 20 September 2021 | Revised: 30 August 2023 | Accepted: 18 September 2023

DOI: 10.1002/sim.9921

RESEARCH ARTICLE

Statistics in Medicine WILEY

Calibrating machine learning approaches for probability estimation: A comprehensive comparison

Francisco M. Ojeda[1,2] | **Max L. Jansen**[3] | **Alexandre Thiéry**[3] | **Stefan Blankenberg**[1,2,4] | **Christian Weimar**[5,6] | **Matthias Schmid**[7] | **Andreas Ziegler**[1,2,3,8,9]

[1]Department of Cardiology, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[2]Centre for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[3]Cardio-CARE, Medizincampus Davos, Davos, Switzerland

[4]German Center for Cardiovascular Research (DZHK), Partner Site Hamburg/Kiel/Lübeck, Hamburg, Germany

[5]BDH-Klinik Elzach, Baden-Wuerttemberg, Germany

[6]Institute for Medical Informatics, Biometry and Epidemiology, University of Duisburg-Essen, Essen, North Rhine-Westphalia, Germany

[7]Institute of Medical Biometry, Informatics and Epidemiology, Faculty of Medicine, University of Bonn, Bonn, North Rhine-Westphalia, Germany

[8]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

[9]Swiss Institute of Bioinformatics, Lausanne, Waadt, Switzerland

Correspondence
Andreas Ziegler, Cardio-CARE, Medizincampus Davos, Herman-Burchard-Str. 1, 7265 Davos Wolfgang, Switzerland.
Email: ziegler.lit@mailbox.org

Statistical prediction models have gained popularity in applied research. One challenge is the transfer of the prediction model to a different population which may be structurally different from the model for which it has been developed. An adaptation to the new population can be achieved by calibrating the model to the characteristics of the target population, for which numerous calibration techniques exist. In view of this diversity, we performed a systematic evaluation of various popular calibration approaches used by the statistical and the machine learning communities for estimating two-class probabilities. In this work, we first provide a review of the literature and, second, present the results of a comprehensive simulation study. The calibration approaches are compared with respect to their empirical properties and relationships, their ability to generalize precise probability estimates to external populations and their availability in terms of easy-to-use software implementations. Third, we provide code from real data analysis allowing its application by researchers. Logistic calibration and beta calibration, which estimate an intercept plus one and two slope parameters, respectively, consistently showed the best results in the simulation studies. Calibration on logit transformed probability estimates generally outperformed calibration methods on nontransformed estimates. In case of structural differences between training and validation data, re-estimation of the entire prediction model should be outweighted against sample size of the validation data. We recommend regression-based calibration approaches using transformed probability estimates, where at least one slope is estimated in addition to an intercept for updating probability estimates in validation studies.

KEYWORDS

calibration, logistic regression, machine learning, probability estimation, probability machine, updating

1 | INTRODUCTION

The estimation of probabilities is important in many life science applications. Examples in medical practice include diagnosis,[1] prognosis.[2] and decision on therapy.[3] The most frequently used approach for probability estimation in classical biostatistics is logistic regression (LogReg). Assumptions underlying LogReg are relatively strict and include that all the important predictors and supposed interactions must be entered correctly in the model to avoid the problem of model misspecification. Machine learning approaches for probability estimation are an alternative. They are generally able to deal with many independent variables, nonlinearity of independent variables and even unknown interactions between predictors. Machine learning algorithms are, however, black boxes, in general. This is considered to be their main disadvantage because interpretability may be an important criterion for the acceptance of a model by clinicians. In the literature, learning machines that provide probabilistic estimates of an outcome are called probability machines, a term coined by Uttley in 1956.[4]

Once a predictive model has been derived, the aim is its application to a sample of new patients at different time points or in other centers.[5,6] It is therefore important to assess its generalizability, that is, its ability to provide accurate predictions in temporal or external validation data.[5] Since population characteristics may differ between the training data, that is, the data which are used for model development, and the validation data, the model might require an update so that it has an improved predictive performance in the new sample. One such approach is termed *calibration*, which is generally used in the context of LogReg models.[7] In the simplest approach, all coefficients from the LogReg model obtained in the training data are kept unchanged except for the intercept, which is re-estimated on the validation data. This calibration is termed *intercept calibration*, and it is an approach to correct for too low or too high predicted probabilities. Alternative approaches based on LogReg include logistic calibration[7] and beta calibration.[8] Some authors use the terms calibration and re-calibration interchangeably. In our opinion, re-calibration refers to a calibration that is done again, and we thus prefer the term calibration.

Learning machines for probability estimation also require updating for application to new patients, and calibration methods have been developed independently by the machine learning community. For example, logistic calibration is termed sigmoid calibration or Platt scaling by machine learners.[9] The latter term is used because this approach is attributed to Platt[10] who introduced it for support vector machines (SVM). Isotonic calibration is also termed receiver operating characteristic (ROC) convex hull method or pool-adjacent violators.[8,11,12] A general approach for adjusting probability estimates, which uses only the outcome prevalences from the derivation and validation datasets, was derived by Elkan,[13] and we developed a LogReg-based calibration approach for random forests (RF).[14]

In previous works, some of the calibration methods were compared analytically,[9] by topic-specific simulation studies[14] or using data available at the University of California in Irvine (UCI) machine learning data repository.[8,15] However, a comprehensive comparison of the various calibration approaches on different learning machines for probability estimation using Monte-Carlo simulation studies and real data sets with training and temporal and/or external validation data is lacking.

This work therefore has three aims. First, we provide an overview of the various approaches proposed in the literature for calibrating machine learning approaches for probability estimation. Second, we report the results of a comprehensive simulation studies in which 10 different calibration approaches were compared and coupled with five approaches for probability estimation. Sample sizes were varied in the simulation study as were outcome probabilities, the number of noise variables, and the effects between model building (training) and validation (calibration) data. The basic concept of the simulation study follows a design described elsewhere.[14]

The third aim is to apply the different calibration approaches and learning machines for probability estimation to two real datasets. One of the datasets is freely available from a data repository. All analysis code for estimating the learning machines for this dataset plus the code for the calibration methods is provided with this manuscript. Our work therefore may thus provide some practical guidance for applied statisticians. All package and function names in this paper refer to the R software for statistical computing.[16] The R code for all the simulations is available as Appendix S2. The code for one of the real data illustrations is available as Appendix S4.

2 | METHODS

The different calibration approaches are discussed first. Next, we describe the simulation study, which is a major expansion of previous work.[14] After that the two real datasets are introduced. The learning machines for probability estimation

elastic net (EN), gradient boosting (GB), logistic regression (LogReg), RF, and SVMs are standard machines, and we therefore refer the interested reader to the literature; references are given below. Where relevant we describe the choice of design parameters or hyper parameters that we used in the simulations and the real data analyses.

2.1 | Calibration methods

2.1.1 | The general concept for calibration

The basic calibration problem is described in Box 1. The solution in terms of the general calibration approach is provided in Algorithm 1. We assume that a learning machine for probability estimation, such as a LogReg or an RF has been developed using a training dataset of size n_t for a dichotomous outcome y and a q-dimensional covariate vector, that is, feature vector $\mathbf{x}$ (Algorithm 1, Step 1). The conditional probability is denoted by $\pi_j = \mathbb{P}(y_j = 1|\mathbf{x}_j)$ for subject $j, j = 1, \ldots, n_t$ in the n_t training data, and its estimate is $\hat{\pi}_j = \hat{\mathbb{P}}(y_j = 1|\mathbf{x}_j)$.

Box 1. The calibration problem

Problem	**Example**
• A learning machine for probability estimation, that is, a prediction model for probabilities has been developed on training data. • The aim is to apply the learning machine for probability estimation to new data from a validation population.	• A model to estimate the probability of a myocardial infarction at the emergency department has been developed using cohort data from clinic A. • The model is to be applied to new patients from clinic B. Cohort data is available for patients at clinic B.

Algorithm 1. The general calibration approach

Step 1 A learning machine for probability estimation 1 has been developed on training data and is available.

Step 2 Apply the learning machine for probability estimation 1 to the validation data. Store the estimated linear predictors/the estimated probabilities.

Step 3 Develop the learning machine for probability estimation 2, whose purpose is to calibrate probabilities, using the dichotomous outcome of the validation data as dependent variable and the linear predictor/the estimated probabilities from Step 2 as independent variable.

Step 4 To estimate a calibrated probability for a new observation from the validation population

a apply the learning machine for probability estimation 1 and store the estimated linear predictor/the estimated probability,

b apply the learning machine for probability estimation 2 using the the estimated linear predictor/the estimated probability.

The calibrated probability is the probability estimate from the learning machine for probability estimation 2.

It is assumed that a validation dataset with n_v new observations is available. The theoretical conditional outcome probabilities in these validation data are $p_i = \mathbb{P}'(y_i = 1|\mathbf{x}_i)$, for $i = 1, \ldots, n_v$. They may thus differ from the conditional probabilities in the training data. When the learning machine for probability estimation, which has been developed on the training data, is applied to the validation data, estimates $\hat{\pi}_i$ are obtained.

The estimation problem is to generate improved probability estimates $\hat{p}_i$, when data $(y_i, \mathbf{x}_i)$ from a validation dataset is available together with a learning machine for probability estimation developed on training data. To this end, the learning machine for probability estimation derived on the training data is applied to all $\mathbf{x}_i$, $i = 1, \ldots, n_v$, from the validation data, yielding estimates $\hat{\pi}_i$ (Algorithm 1, Step 2).

In Step 3 of Algorithm 1, a learning machine for probability estimation is developed on the validation data by using the estimates $\hat{\pi}_i$ or a transformation of these probability estimates, such as the linear predictor $\hat{\eta}_i = \text{logit}(\hat{\pi}_i)$, as independent variable and the outcome of the validation data y_i as dependent variable.

The calibrated probability for a new observation with covariates $\mathbf{x}^*$ is obtained by applying the first learning machine for probability estimation to $\mathbf{x}^*$, which yields $\hat{\pi}^*$ or equivalently $\hat{\eta}^* = \text{logit}(\hat{\pi}^*)$ (Algorithm 1, Step 4a). This estimate is then applied to the second learning machine for probability estimation (Algorithm 1, Step 4b), yielding the estimate $\hat{p}^*$.

It is stressed that, in principle, any learning machine for probability estimation could be used for the estimation problem in Step 3 of Algorithm 1. For example, one could use a LogReg on the linear predictor $\hat{\eta}_i$ with outcome y_i. In fact, many different approaches have been proposed in the literature, and the next section provides an overview of the literature for approaches tackling the calibration problem. A brief summary of the different calibration methods is compiled in Table 1.

TABLE 1 Summary of calibration approaches.

Method	R package/functions	Parameterization	Assumptions	Comment
Logistic	stats/glm	Parametric	Linear predictor linearly associated with outcome on log odds scale	Estimation of two parameters: intercept and slope
Beta	stats/glm	Parametric	$\log(\pi)$ and $\log(1-\pi)$ linearly associated with outcome on log odds scale	More flexible than logistic. Estimation of three parameters
Intercept	stats/glm	Parametric	Linear predictor linearly associated with outcome on log odds scale with slope equal to one	Single parameter estimated: intercept
Slope	stats/glm	Parametric	Linear predictor linearly associated with outcome on log odds scale and an intercept equal to one	Single parameter estimated: slope
RCS[a]	rms/lrm, rcs	Parametric	Degree of flexibility depends on number of knots	More flexible than logistic. More parameters estimated
Platt	stats/glm	Parametric	Probabilities linearly associated with outcome on log odds scale	Similar to logistic but acts on probabilities rather than linear predictor
Isotonic	stats/isoreg	Non-parametric	Monotone increasing relationship	Large datasets required. May perform poorly on small datasets
Elkan	NA[b]		Linear predictor distribution conditional on the outcome identical for training and calibration data	Closed formula. No regression model required
Chen	stats/density	Semi-parametric	Sample sizes for both affected and unaffected subjects should be sufficiently large for stable estimation	Requires estimation of densities in affected and unaffected subjects
SBA[c]	kknn/kknn	Nonparametric	Does not use outcome information y_i from validation data for calibrating outcome probability for subject i	Complete validation dataset needs to be available to which k-nearest neighbor calibration algorithm is applied
SLR[d]	segmented/ segmented	Parametric	Linear predictor has piecewise linear association with outcome	More flexible than logistic. More parameters estimated

Note: The listed R packages and functions may be used for estimation, and alternative implementations may be available.

[a] Restricted cubic splines.

[b] Not available to the best of our knowledge, but easily implemented. R code is provided in the Appendices S1 to S5.

[c] Similarity-binning averaging.

[d] Segmented logistic regression.

2.1.2 | Regression-based calibration

One approach to better adjust the probability estimates $\hat{\pi}_i$ to characteristics of the validation dataset is by the use of standard parametric regression models. To this end, a regression can be fitted that regresses the dichotomous outcome variable y_i on the estimated probability $\hat{\pi}_i$ or its linear predictor $\hat{\eta}_i = \text{logit}(\hat{\pi}_i)$. In biostatistics, the most popular approach for this is to use a simple generalized linear model on the logit of the estimated probability, that is,

$$p_i = F(\alpha + \beta\ \hat{\eta}_i) = F\big(\alpha + \beta\ \text{logit}\ (\hat{\pi}_i)\big), \tag{1}$$

where F is a cumulative distribution function. The calibrated probability is $\hat{p}_i$, that is, the predicted value from Equation (1). Below, we introduce common special cases of Equation (1).

Logistic calibration

In logistic calibration,[17,18] updated probabilities are obtained by using the LogReg model so that $F = \text{expit}$ with $\text{expit}(x) = e^x/(1 + e^x)$. In the logistic calibration model, a LogReg is thus estimated with an intercept and a slope, and the linear predictor from the initial learning machine for probability estimation is serving as covariate. Logistic calibration can be optimal for different probability distributions, and conditions have been formulated that lead to this optimality.[9]

Platt scaling

The terms logistic calibration and Platt scaling[10] are often used synonymously in the literature. However, there are two small but important differences between the two approaches: First, Platt scaling was initially introduced for SVMs, where the coefficients $\boldsymbol{\beta}^*$ were estimated for the separating hyperplane. The predictions for a new observation $\mathbf{x}_i$ from the SVM are $\hat{\eta}_i^* = h(\mathbf{x}_i)'\hat{\boldsymbol{\beta}}^*$, where h is the hyperspace function. Platt scaling is therefore based on a non-transformed estimate $\hat{\eta}_i^*$, and a transformation was not considered because η_i^* is inherent to the idea of Platt scaling. This means that the regression model to be considered is $p_i = F(\alpha + \beta\ \hat{\pi}_i)$ if the learning machine for probability estimation outputs a probability estimate and not a linear predictor. Second, to fit the regression model, Platt[10] applied an additional regularization to the target function, employing smoothed versions of the class labels 0 and 1. Thus, only if the linear predictor is considered in Platt scaling and if the original class labels are used for model fitting, logistic calibration and Platt scaling are equivalent.[9] In this work, we do not consider any smoothed versions of the class labels, and Platt scaling is directly applied to probability estimates.

Sigmoid calibration

Logistic calibration is also sometimes synonymously called sigmoid calibration.[9] However, any sigmoid cumulative distribution function could be used for estimation, such as the cumulative distribution function from the standard normal distribution.

Intercept calibration and slope calibration

Intercept calibration and slope calibration are natural modifications of Equation (1). Specifically, slope calibration was used by Cox[18] in his seminal paper to introduce the concept of calibration. In slope calibration, the intercept α is dropped such that the model $p_i = \text{expit}(\beta\ \hat{\eta}_i)$ is estimated in case of a LogReg, while the slope is dropped in intercept calibration. This way, $\hat{\eta}_i$ is used as offset in an intercept-only LogReg model,[19] and logistic calibration is the combination of intercept and slope calibration.

Beta calibration

Beta calibration is a relatively new approach, and it was derived by Kull et al[8] using the beta distribution. In beta calibration, the model of Equation (1) is modified, and the two components from the logit model are modeled separately by

$$p_i = F\big(\alpha + \beta_1 \log(\hat{\pi}_i) - \beta_2 \log(1 - \hat{\pi}_i)\big). \tag{2}$$

Kull et al[8] specifically used the logit link, that is, $F = \text{expit}$ and called their approach "beta calibration via logistic regression." Beta calibration is equivalent to logistic calibration if $\beta_1 = \beta_2$.[8,9] Otherwise, it is more flexible than logistic calibration.

Splines calibration

The use of splines leads to an even more flexible regression-based calibration approach.[20] A spline calibration approach which is closely related to logistic calibration is the use of logistic calibration with RCS,[21] to which we refer as RCS calibration. In this approach, logistic calibration is used at the tail of the distribution, and a more flexible polynomial type of calibration in the inner part of the probability distribution.

Other regression-based calibration approaches

As outlined above, any reasonable regression approach can be used that relates the dichotomous outcome y_i from the validation data set to its predicted probabilities from the validation dataset, but estimated using the learning machine for probability estimation applied to the training data. These include local regression approaches, such as segmented logistic regression, also termed piecewise logistic regression,[22] or generalized additive models.[23] Unfortunately, some of these approaches are not readily available in standard software.

Regression-based calibration can also be combined with ensembles by calibrating each element of the ensemble separately. For example, Dankowski and Ziegler[14] proposed an updating method tailored to RF. In a similar direction one could use logistic model trees[24] for the construction of probability trees in an RF. These could then be used for tree-based calibration as proposed by Leathart et al for a single tree.[25]

2.1.3 | Isotonic calibration

Instead of using parametric regression methods for calibration, non-parametric approaches may be employed. Isotonic regression is such a nonparametric calibration method. It uses a step function with monotonically increasing values to obtain calibrated probability estimates $\hat{p}_i$, $i = 1, \dots, n_v$.[26-28] The theory has been described in detail by de Leeuw et al.[26] Several algorithms are available to find the stepwise function. The best-known algorithm is the PAV algorithm, whose solution has been shown to be optimal under a broad class of loss functions.[29] An implementation of the PAV algorithm is, for example, available in the function `isoreg` within the `stats` package, and the basic form of the PAV algorithm is provided in Algorithm 2.

Algorithm 2. The pool-adjacent violators (PAV) algorithm

Step 1	A learning machine for probability estimation has been developed on training data and is available.
Step 2	Apply the learning machine for probability estimation to the validation data and store the linear predictor values $\hat{\eta}_i$.
Step 3	Sort the linear predictor values $\hat{\eta}_{(i)}$ from the validation sample descendingly.
Step 4	Start with the lowest linear predictor value $\hat{\eta}_{(i)}$ and check whether the corresponding outcome data $y_{(i)}$ are ordered.
Step 5	Adjacent violators of this order are pooled into one group, and the estimated probabilities in the group are averaged.
Step 6	Go back to step 4 and iterate until all violations have been disentangled.

In Table 2 we illustrate the calculation of the PAV algorithm, and this reveals one specific property of this procedure: the idea is that the calibrated probabilities form a monotone increasing sequence $\hat{p}_{(1)} \leq \cdots \leq \hat{p}_{(n_v)}$, where the order refers to the order of the predictors $\hat{\eta}_i$. If in iteration *iter* two estimates do not satisfy the inequality property $\hat{p}_{(i)}^{[\text{iter}]} \leq \hat{p}_{(i+1)}^{[\text{iter}]}$, the PAV algorithm replaces both of them by their average $\hat{p}_{(i)}^{[\text{iter}+1]} = \hat{p}_{(i+1)}^{[\text{iter}+1]} = (\hat{p}_{(i)}^{[\text{iter}]} + \hat{p}_{(i+1)}^{[\text{iter}]})/2$. Thus, probability estimates are binned to guarantee the monotonic relationship, and this results in a step function. In the artificial example of Table 2, there are just four different calibrated probabilities after Step 5 of the algorithm, although all 10 linear predictors are different.

Several other binning methods can be found in the literature, such as quantile binning.[30] In this approach, observations from the n_v validation data are ordered in the first step according to their estimated probabilities $\hat{\pi}_{(i)}$. The number of bins b is prespecified, and a small number of bins is used. Specifically, Zadrozny and Elkan[31] used 10 bins in their numerical evaluation of the quantile binning method. The calibrated probability for a new subject is the average over the probability estimates from the bin in which the uncalibrated probability estimate for this subject falls. The clear disadvantage of the quantile binning approach is similar to the disadvantage of isotonic regression. It is limited by the number

TABLE 2 Illustration of the pool-adjacent violators (PAV) algorithm.

Subject	$\hat{\eta}_{(i)}$	$y_{(i)}$	Step 1	Step 2	Step 3	Step 4	Step 5
3	0.99	1	1	1	1	1	1
2	0.95	1	1	1	1	1	1
8	0.80	0	0	0	0	0	1/2
10	0.75	0	0	0	1/2		1/2
9	0.50	1	1	1		2/3	1/2
4	0.40	1	1	1	1		1/2
1	0.35	0	0	1/3	1/3	1/3	1/3
7	0.30	0	1/2	1/3	1/3	1/3	1/3
5	0.25	1		1/3	1/3	1/3	1/3
6	0.05	0	0	0	0	0	0

of bins, and probability estimates are constant within bins. The quantile binning approach has been expanded to an ensemble method, termed Bayesian binning quantiles (BBQ).[30] Essentially, BBQ computes a Bayesian score to perform model averaging over the space of quantile binnings with varying bin numbers. Other binning approaches have been reviewed by Böken.[9]

2.1.4 | General updating approaches

Most approaches introduced above are based on regression models. Recently, Chen et al[32] described a semi-parametric updating approach, which is based on a latent continuous decision variable paired with maximum likelihood estimation. It can be described as a general updating approach (GUA) using the ratio of normal class densities. Here, we consider the following modification of Chen's GUA: Let $b' = \mathbb{P}'(y = 1)$ denote the base rate, that is, the unconditional event probability in the validation population. Furthermore, assume that $\hat{\pi}$ has been estimated for a subject in the validation data using the learning machine for probability estimation derived on the training data. The estimated probability $\hat{\pi}$ is logit-transformed, that is, $\hat{\eta} = \text{logit}(\hat{\pi})$. Then the posterior probability can be obtained as follows[32]

$$\mathbb{P}'(y = 1|\hat{\eta}) = \frac{f'(\hat{\eta}|y = 1)\ \mathbb{P}'(y = 1)}{f'(\hat{\eta}|y = 1)\ \mathbb{P}'(y = 1) + f'(\hat{\eta}|y = 0)\ \mathbb{P}'(y = 0)} = \frac{\text{LR}\ b'}{\text{LR}\ b' + (1 - b')}, \tag{3}$$

where the likelihood ratio (LR) is $\text{LR} = f'(\hat{\eta}|y = 1)/f'(\hat{\eta}|y = 0) = f_1(\hat{\eta})/f_0(\hat{\eta})$. Estimates of the densities f_0 and f_1 for the distribution of the linear predictor $\hat{\eta}$ can be obtained from kernel density estimates, which are separately estimated for the affected and unaffected subjects, respectively, from the validation dataset. Note that this approach slightly differs from the original method by Chen et al, in that it performs kernel density estimation instead of transforming $\hat{\eta}$ to a latent probability space before estimating the LR assuming normal class distributions. Nevertheless, we will refer to it as Chen calibration in the following.

Elkan's[13] GUA avoids the re-estimation from a regression model. Instead, it captures the difference in base rates between the population in which the learning machine for probability estimation was developed, and the population to which it should be applied after calibration. To this end, let $b = \mathbb{P}(y = 1)$ and $b' = \mathbb{P}'(y = 1)$ denote the base rates, that is, the unconditional event probabilities in the two populations, in which the machine has been developed and to which it is to be applied, respectively. Using the training data, the probabilities $\mathbb{P}(y = 1|\mathbf{x})$ can be estimated from a learning machine for probability estimation for observations with features $\mathbf{x}$. The validity of Elkan's GUA relies on the assumption that the change in the base rate is the only difference between the two populations. In particular this means that $\mathbb{P}(\mathbf{x}|y = 0) = \mathbb{P}'(\mathbf{x}|y = 0)$ and $\mathbb{P}(\mathbf{x}|y = 1) = \mathbb{P}'(\mathbf{x}|y = 1)$ are assumed to hold. The probability $\mathbb{P}'(y = 1|\mathbf{x})$ can then be expressed as a function of $\mathbb{P}(y = 1|\mathbf{x})$, b and b':

$$\mathbb{P}'(y = 1|\mathbf{x}) = \frac{b'\ \mathbb{P}(y = 1|\mathbf{x})\ (1 - b)}{b'\ \mathbb{P}(y = 1|\mathbf{x}) + b - b\ \mathbb{P}(y = 1|\mathbf{x}) - bb'}. \tag{4}$$

If both base rates are known or if estimates are available for the base rates, the formula may be easily applied. While Elkan's GUA can be used for updating probability estimates from any method, its applicability is limited by the assumption of equal covariate distributions in both datasets. Otherwise, calibrated probability estimates might be biased.[14]

Other GUAs are related to tree-based methods, such as the Laplace method.[27] To this end, assume that n subjects from the validation data reside in a terminal node with k of them being cases. The uncalibrated probability estimate is k/n, and the Laplace calibrated is $(k+1)/(n+2)$. This estimate adjusts probability estimates to be closer to $1/2$. As pointed out by Zadrozny and Elkan,[31] this is not reasonable when the two groups are far from being balanced.

2.1.5 | Similarity-binning averaging

Bella et al[33] have described an entirely different updating strategy. It requires all the observations plus the estimated probabilities from the training data set or, alternatively, from another independent dataset. Algorithm 3 provides a description of the similarity-binning averaging (SBA) procedure.

Algorithm 3. SBA algorithm

Step 1 A learning machine for probability estimation 1 has been developed on training data and is available.
Step 2 Estimate $\hat{\pi}_j, j = 1, \ldots, n_v$, using the learning machine for probability estimation for all n_v observations from the validation data.
Step 3 For an observation i in the validation data, estimate its k nearest neighbors (kNN) in the validation data from the features $(\mathbf{x}_i, \hat{\pi}_i)$.
Step 4 The k probability estimates from observation i's kNN are averaged, that is, $\hat{p}_i = \frac{1}{k}\sum_{j\in\{k \text{ nearest neighbors of } i\}} \hat{\pi}_j$, where the neighborhood of i is defined using a metric on the pairs $(\mathbf{x}_j, \hat{\pi}_j)$. The average yields the calibrated probability.

An important property of this approach is that it does not utilize outcome information y_i from the validation data. Observations with similar covariates and similar estimated probabilities are grouped. Furthermore, the metric for determining similarity has not been reported by Bella et al,[33] and it remains to investigate the dependency of SBA on the choice of the metric. Additionally, this binning procedure inherits several disadvantages of the kNN approaches.[9] The most important shortcoming is that the rule cannot be simply transferred to a new data point. In general, the entire dataset needs to be available which is used to define the kNN. A final relevant shortcoming is the tuning of k. Here, we refer to the assumptions that need to be met for kNN to yield consistency, results on the convergence rate and asymptotic normality.[34]

2.1.6 | Other updating methods

Various alternative updating methods have been proposed in the literature; for reviews, see, for example, References 9,27, and 28; for a taxonomy, see Reference 35. In principle, there is no limitation in the variability of updating methods. Indeed, calibration is a straightforward probability estimation problem, and any learning machine for probability estimation may be used for calibration. Instead of considering calibration as a prediction problem, it may be interpreted as a problem in which the posterior probability for an event needs to be estimated by combining specific information from the structure of the validation data, covariate information from individuals and the learning machine for probability estimation estimated on the training data.

A different path for calibrating LogReg has been followed by Jiang et al.[36] Their approach requires a confidence interval together with the predicted probability for calibration.

2.1.7 | Choice of calibration methods for the simulation study

Standard approaches in biostatistics are logistic calibration, intercept calibration, and slope calibration. The best-known calibration method in machine learning is Platt scaling. From these standard approaches we selected logistic

calibration, slope calibration, and Platt scaling for our simulation study. We also included beta calibration, which includes the standard logistic calibration method as special case. For even greater flexibility, we chose RCS calibration. For the choice of knots, we followed the recommendations of Harrell.[21] For example, for a limiting sample size of ≥ 100, the choice of knots was 5, and knots were placed following Harrell as well. In case of nonconvergence, the number of knots was reduced by one. In case of nonconvergence with three knots, the standard logistic calibration was used. We also selected the PAV algorithm from the isotonic calibration approaches and chose to employ both Chen calibration and Elkan calibration. Elkan calibration was already demonstrated to be inferior to logistic calibration.[14] However, we included it because of its simplicity and good performance, when specific model assumptions are met. We did not include intercept calibration in the simulation study because it may lead to biased estimates if there are more structural differences between the model-building data and the calibration data but differences in baseline probabilities. Intercept calibration is in this sense similar to Elkan calibration. Finally, we refitted the original model in the validation data. The selected calibration approaches are readily available for applications because they are included in standard software packages or are easily implemented.

2.2 | Performance measures and comparison of calibration approaches

2.2.1 | Performance measures

Several measures are well suited for measuring the performance of machine learning approaches for probability estimation. We stress that the area under the ROC curve (AUC) and the ROC curve are adequate measures in classification problems, but they are not deemed suitable in probability estimation problems because the AUC is invariant under monotonic transformations of probabilities.[9] In biostatistics, a natural performance measure is the mean squared error (MSE), which is estimated by

$$\widehat{\mathrm{MSE}} = \frac{1}{n_v}\sum_{i=1}^{n_v}\left(p_i - \hat{p}_i\right)^2, \tag{5}$$

where p_i is the true probability and $\hat{p}_i$ is its calibrated estimator. Since the theoretical probabilities are only available in simulations, the Brier score (BS) is its pendant for real data analysis.[37,38] The BS estimator is given by

$$\widehat{\mathrm{BS}} = \frac{1}{n_v}\sum_{i=1}^{n_v}\left(y_i - \hat{p}_i\right)^2, \tag{6}$$

and it is thus the estimated observed quadratic loss.

Another popular measure is the log loss[39] (LogLoss), which is also known as probabilistic cost, binary cross-entropy or entropy. It is related to the Kullback-Leibler distance between the true and the estimated probability model and likely leads to good prediction algorithms. Specifically, a prediction algorithm that is optimal under the LogLoss function will also be optimal under the BS loss.[39] The reverse is not necessarily true. The expected and observed log loss estimators are given by

$$\widehat{\mathrm{LogLoss}}_{\mathrm{exp}} = -\frac{1}{n_v}\sum_{i=1}^{n_v}\left(p_i \ln \hat{p}_i + (1-p_i)\ln(1-\hat{p}_i)\right) \quad \text{and} \quad \widehat{\mathrm{LogLoss}} = -\frac{1}{n_v}\sum_{i=1}^{n_v}\left(y_i \ln \hat{p}_i + (1-y_i)\ln(1-\hat{p}_i)\right), \tag{7}$$

respectively. The BS and LogLoss scores are strictly proper,[39,40] which means that the scores are only minimal if the true probabilities are inserted as estimates. Additional performance measures are described in Appendix S1.

2.2.2 | Calibration plot

The relationship between the observed and predicted probabilities of a binary outcome can be graphically depicted using the calibration plot.[19] This plot can be obtained by regressing the binary outcome on the machine predicted probabilities using smoothing techniques and plotting the resulting curve. A line showing the angle bisector is usually

displayed for orientation. In the context of a simulation study, where the true probabilities are known, the binary outcome is replaced by the true probabilities. In this case, the underlying pairs of predicted and true probabilities can also be presented in the calibration plot. In this work, we used generalized additive model (GAM) estimates as smoother for the calibration plots in the simulation study and locally weighted scatterplot smoothing (LOESS) for the real data. For the simulation study, the results of all replications considered are displayed using a single calibration plot.

2.2.3 | Comparison of two calibration approaches

For the comparison of two calibration approaches or learning machines for probabilities, the LogLoss of the two machines can be compared. This can be done by a resampling procedure not only for the comparison of two LogLosses but also two other performance measures, such as the BS.[38,41]

2.2.4 | Comparison of multiple calibration approaches—critical difference

Demšar[42] showed how classifiers may be compared over multiple datasets. His approach can be used for comparing different learning machines for probability estimates and/or calibration approaches[9] by evaluating homogeneous subgroups nonparametrically based on the Nemenyi post-hoc test approach.[42] In the following, we assume that C calibration approaches have been used on M datasets. The datasets may correspond to the simulated datasets for a specific simulation scenario (see Section 2.3). Let r_c^m denote the rank of calibration approach c on the mth dataset. In the first step, the average rank of calibration approach c is calculated: $R_c = \frac{1}{M}\sum_{m=1}^{M} r_c^m$. Demšar[42] proposed to proceed with the pairwise post-hoc test. Two classifiers perform significantly differently if the corresponding average ranks differ by at least the critical difference (CD)

$$\mathrm{CD} = q_\alpha \sqrt{\frac{C(C+1)}{6M}} \Big/ \sqrt{2}, \tag{8}$$

where q_α is the upper $1-\alpha$ quantile of the studentized range distribution. Because the sample sizes of the datasets are sufficiently large, the quantile is calculated for an infinite number of degrees of freedom. The quantile function q_α is implemented in `R` and may be called by `qtukey(.95, 10, df = Inf)` for the 5% type I error level and 10 calibration methods to be compared. A nice illustration on the calculation of the post-hoc CD has been provided by Demšar.[42] Results may be visualized in critical difference plots (CD plots) by displaying the average ranks of the machines and/or calibration approaches, with lower average ranks showing better performance. Homogeneous groups of algorithms are displayed by bars. These bars are shorter than the CD.

2.3 | Simulation study

The aim of the simulation study is to compare the performance of several often-used calibration approaches. The code for the simulation study is provided in Appendix S2. The different simulation settings are summarized in Table 3. An extended version of this table is presented in Appendix S3a, Table 2.1. Details of the data generation are described below. Here, we provide a brief summary. Scenario 1 is a basic scenario with two continuous predictors and without noise variables. Scenarios 2 to 4 are similar to scenario 1 but with an increasing number of noise variables (10, 50, 100). Scenarios 5 to 8 are comparable to scenarios 1 to 4, except that the disease prevalence in the model building and in the calibration data differs between scenarios 1 to 4 and scenarios 5 to 8. Scenarios 9 to 12 are again similar to scenarios 1 to 4, but they have 10 predictors, five being continuous and five being categorical, with the number of noise variables varying from 0 to 100. Scenario 13 is a model with substantial nonlinearity, and scenario 14 has a smaller dataset for estimating the calibration model. Scenario 15 combines scenarios 13 and 14. The model thus is nonlinear and has only a small dataset available for estimating the calibration model. Models 16 to 23 all investigate the effect of differences in the covariate distribution or differences in the regression coefficients between training and calibration data. All 23 simulation scenarios were generated using a sample size of 1000 for model building and calibration testing. Two different sample sizes were

TABLE 3 Main settings in the simulation scenarios: Number of continuous covariates (Cont), number of categorical covariates (Cat), number of noise variables (Noise), outcome prevalence in model building data (Prev MB), outcome prevalence in cal-training data (Prev Cal) and, if applicable, additional distinctive feature (Distinctive feature).

Scenario	Cont	Cat	Noise	Prev MB	Prev Cal	Distinctive feature
1	2	0	0	0.5	0.69	—
2	2	0	10	0.5	0.69	10 noise variables
3	2	0	50	0.5	0.69	50 noise variables
4	2	0	100	0.5	0.69	100 noise variables
5	2	0	0	0.2	0.41	Lower disease prevalence
6	2	0	10	0.2	0.41	Lower disease prevalence
7	2	0	50	0.2	0.41	Lower disease prevalence
8	2	0	100	0.2	0.41	Lower disease prevalence
9	5	5	0	0.57	0.77	Additional covariates
10	5	5	10	0.57	0.77	Additional covariates plus noise variables
11	5	5	50	0.57	0.77	Additional covariates plus noise variables
12	5	5	100	0.57	0.77	Additional covariates plus noise variables
13	2	0	0	0.44	0.64	Data generation using Mease model
14	2	0	0	0.5	0.69	Smaller calibration data set for training
15	2	0	0	0.44	0.65	Smaller calibration data set for training, data generation using Mease model
16	1	0	0	0.7	0.84	Covariate distribution: MB $\mathcal{N}(1,1)$, Cal $\mathcal{N}(1,1)$
17	1	0	0	0.7	0.81	Unequal covariate distribution: MB $\mathcal{N}(1,1)$, Cal $\mathcal{N}(1,2)$
18	1	0	0	0.69	0.5	Unequal covariate distribution: MB $\mathcal{N}(1,1)$, Cal $\mathcal{N}(-1,1)$
19	1	0	0	0.7	0.5	Unequal covariate distribution: MB $\mathcal{N}(1,1)$, Cal $\mathcal{N}(-1,2)$
20	1	0	0	0.5	0.61	Unequal coefficients: MB $\beta = 1$, Cal $\beta = 3$
21	1	0	0	0.5	0.7	Unequal coefficients: MB $\beta = 1$, Cal $\beta = -1$
22	2	0	10	0.77	0.4	Unequal covariate distribution for both continuous variables: MB $\mathcal{N}(1,1)$, Cal $\mathcal{N}(-1,1)$
23	2	0	10	0.77	0.77	Unequal covariate distribution for one continuous variable: MB $\mathcal{N}(1,1)$, Cal $\mathcal{N}(3,1)$, for the other continuous variable: MB $\mathcal{N}(1,1)$, Cal $\mathcal{N}(-1,1)$

Note: The sample size was 1000 for model building, 1000 for cal-training and 1000 for cal-test datasets, except for scenarios 14 and 15, where the cal-training data had a sample size of 100.

Abbreviations: Cal, Calibration data; MB, Model building data.

used to estimate the calibration models, 1000 and 2000 samples, with the exception of the smaller calibration datasets in scenarios 13 and 15.

2.3.1 | Data generation

The design of the simulation study is related to Dankowski and Ziegler.[14] To this end, we generated data from two populations with different disease prevalences. The data from the first population were used for model building. Data from the second population were split into training data and test data. They are named cal-training data and cal-test data, respectively.

The logistic regression model was used to generate most datasets for the simulation scenarios considered. For all models but scenarios 13 and 15, features $\mathbf{x} = (x_1, \ldots, x_q)$ were generated according to predefined probability distributions, and regression coefficients $\beta_1, \ldots, \beta_q$ and the intercept α were set to specific values. The linear predictor $\eta = \alpha + \beta_1 x_1 + \ldots + \beta_q x_q$ was expit-transformed $p = \mathbb{P}(y = 1|\mathbf{x}) = \text{expit}(\eta)$. The class of the outcome y was then obtained from a Bernoulli distribution with event probability p. For scenarios 5 to 8, p^3 was used as event probability instead of p, when sampling was done from the Bernoulli distribution.

To investigate the effect of nonlinearities in the data generation process, data were created following the Mease model[43] in simulation scenarios 13 and 15. The Mease model is based on a simple two-dimensional circle model, where $\mathbf{x}$ is a two-dimensional random vector uniformly distributed on the square $[0, 50]^2$. The dependent variable y is dichotomous with values 0 or 1 and conditional probabilities defined as

$$\mathbb{P}(y = 1|\mathbf{x}) = \begin{cases} 1 & r(\mathbf{x}) < 8, \\ \frac{28 - r(\mathbf{x})}{20} & 8 \leq r(\mathbf{x}) \leq 28, \\ 0 & r(\mathbf{x}) > 28, \end{cases} \tag{9}$$

where $r(\mathbf{x})$ is the Euclidean distance of $\mathbf{x}$ from $(25, 25)$. Continuous noise variables were generated in this example according to normal distributions with mean 0 and variance 210, which is close to $50^2/12$.

In total, we considered 23 different simulation scenarios. In all of them, with the exception of scenarios 13 and 15 created using the Mease model, the intercept was set to 0 for the model building data, and it was set to 1 for the calibration datasets. In scenarios 5 to 8 the probabilities from the underlying logistic models were cubed, that is, transformed by the power of 3, before drawing the outcome to reduce the outcome prevalence. For scenarios 13 and 15, the probabilities in the calibration datasets were changed to a base rate of 1.7 × the base rate of the training data using Elkan's formula.

Continuous variables were generated according to a standard normal distribution, except for simulation scenarios 13 and 15, where uniform random variables were used, scenario 16 where the $\mathcal{N}(1, 1)$ distribution was used for the model building and calibration data and scenarios 17 to 19, 22 and 23, where the distribution of the continuous variables differed between the model building data and the calibration data. The five additional categorical variables in simulation scenarios 9 to 12 had, respectively, 2, 2, 3, 4, and 5 equally probable categories. The regression coefficients were identical in the model building data and the calibration data sets for all covariates in those scenarios generated using LogReg, except for scenarios 20 and 21. In simulation scenario 20, the coefficient was 1 in the model building data and set to 3 in the calibration data. In all simulation scenarios but simulation scenarios 14 and 15, the number of observations in the model building data, the cal-training data and the cal-test data was 1000 each. In scenarios 14 and 15, the cal-training data set was smaller and comprised only 100 observations. One-hundred replications were generated for each simulation scenario. Additionally, an alternative cal-training data was considered where the sample size was doubled, that is, to 2000 except for scenarios 14 and 15, where the sample size was 200.

2.4 | Real datasets

Two applications were selected for this work. Application 1 has the strength that the covariate distribution is similar between training and validation datasets. In addition, two different models are of interest. While the outcome is well

balanced in one of the two models, it is unbalanced in the other. Application 2 was selected because the data are freely available and covariates show substantial imbalances between training and validation data.

2.4.1 | Application 1: prediction of functional outcome after stroke—German Stroke Study Collaboration data

This study investigated prognosis 90 days after acute ischemic stroke. The study has been described in detail elsewhere.[44,45] For important information to comply with the Methods part of the TRIPOD statement,[46] we refer to Appendix S1. In brief, the training data comprised 1754 patients with ischemic stroke. Data were prospectively collected from 23 neurology departments with acute stroke unit in 1998 and 1999. A systematic literature search was conducted prior to model development to identify possible relevant covariates. The literature search is available upon request from the corresponding author. Patients for validation were enrolled during 2001 and 2002 after sample size calculations for the validation study.[47] Nine hospitals participated in the validation study only, allowing for a combined temporal and external validation in a sample of 874 patients. Four hospitals participated in both the initial and the validation study, providing data for temporal validation from 596 patients. Patients were informed about study participation, and all patients provided informed consent if their personal data were to be transferred to the data management center. The study was approved by the Ethics Committee of the Essen University Hospital, Germany.

The aim was to construct a model for complete restitution vs incomplete restitution or mortality. The Barthel Index (BI)[48] is a widely used rating scale and measures a person's daily functioning, particularly the activities of daily living (ADL) and mobility on a scale from 0 to 100 in steps of five points. Functional independence was assumed for individuals with BI $\geq$ 95. For logistic regression, we used the variables that were previously reported[44,49] (Appendix S1). For the other machine learning approaches, we used all 34 variables for prediction that were available for the training data and the validation data, as before.[49] Table 4 provides descriptive statistics, and they indicate that the distribution differs between the training data and the temporal and external validation data for some covariates, such as fever. Missing values were imputed prior to analysis for pragmatic reasons. In detail, for each continuous variable imputed values were drawn from a normal distribution with the mean and standard deviation calculated from non-missing values. For each categorical variable, imputed values were drawn from a random sample with the probability of each category calculated from non-missing values. In the modeling process for logistic regression, we built on our previous work.[49] More information on the regression models and regression coefficients for the final logistic regression models are provided in Appendix S1.

TABLE 4 Patient characteristics in the stroke data. *n* (%) are displayed for dichotomous variables, mean and standard deviation (SD) for continuous variables.

Variable	Training data	Temporal validation	External validation
Barthel Index after 90 days: $\geq$ 95	1025 (58.4%)	337 (56.5%)	494 (56.5%)
Survival after 90 days	1588 (90.5%)	546 (91.6%)	811 (92.8%)
Prior stroke	353 (20.1%)	137 (23.0%)	172 (19.7%)
Diabetes mellitus	436 (24.9%)	165 (27.7%)	229 (26.2%)
Lenticulostriate arteries infarction	188 (10.7%)	31 (5.2%)	64 (7.3%)
Fever	220 (12.5%)	51 (8.6%)	54 (6.2%)
Neurological complications	73 (4.2%)	26 (4.4%)	37 (4.2%)
Female sex	716 (40.8%)	270 (45.3%)	357 (40.8%)
Age at event (in years)	68.1 (12.7)	68.0 (12.4)	67.8 (12.4)
NIHSS left arm	0.6 (1.2)	0.6 (1.1)	0.6 (1.2)
NIHSS right arm	0.7 (1.2)	0.5 (1.1)	0.6 (1.1)
NIHSS total score	6.9 (6.2)	6.1 (5.7)	6.6 (6.1)
Rankin Scale[a]	3.1 (1.4)	2.4 (1.6)	2.5 (1.6)

Abbreviation: NIHSS, National Institutes of Health Stroke Scale.
[a]Modified Rankin Scale rated 48 to 72 hours after admission.

TABLE 5 Patient characteristics in the coronary artery disease data. *n* (%) are displayed for dichotomous variables, mean and standard deviation (SD) for continuous variables.

	Training data	Validation data		
Variable	**Cleveland**	**Hungarian**	**VA**	**Swiss**
CAD	139 (45.9%)	106 (36.1%)	149 (74.5%)	115 (93.5%)
Female sex	97 (32.0%)	81 (27.6%)	6 (3.00%)	10 (8.13%)
Typical chest pain	23 (7.6%)	11 (3.7%)	8 (4.0%)	4 (3.3%)
Atypical chest pain	50 (16.5%)	106 (36.1%)	14 (7.0%)	4 (3.3%)
Non-anginal chest pain	86 (28.4%)	54 (18.4%)	47 (23.5%)	17 (13.8%)
High fast blood sugar	45 (14.9%)	20 (6.80%)	71 (35.5%)	10 (8.1%)
Normal resting ECG	151 (49.8%)	236 (80.3%)	80 (40.0%)	85 (69.1%)
Abnormal ST-T wave	4 (1.3%)	52 (17.1%)	93 (46.5%)	31 (25.2%)
Exercise induced angina	99 (32.7%)	90 (30.6%)	131 (65.5%)	55 (44.7%)
Age at investigation (in years)	54.4 (9.0)	47.8 (7.8)	59.4 (7.8)	55.3 (9.0)
Resting BP (mm Hg)	131.7 (17.6)	132.6 (17.6)	133.8 (18.3)	130.2 (22.4)
Cholesterol (mg/dL)	246.7 (51.8)	250.8 (64.9)	178.7 (112.0)	—
Max heart rate (beats/min)	149.6 (22.9)	139.1 (23.5)	122.8 (18.8)	121.6 (25.9)
ST depression (mm)	1.03 (1.16)	0.59 (0.91)	1.32 (0.94)	0.65 (1.03)

Abbreviations: BP, Blood pressure; CAD, coronary artery disease; ECG, Electrocardiogram; Hungarian: the Hungarian Institute of Cardiology, Budapest (Hungarian); Swiss, the University Medical Centers Zurich and Basel, Switzerland; VA, Veterans Administration Medical Center in Long Beach, California.

Conditional predictive impact (CPI) and two-sided 95% confidence intervals were estimated for RF with 10,000 trees and the LogLoss as performance measure.[50] The CPI is an importance measure which quantifies the contribution of one or several covariates to the predictive performance of an algorithm, conditional on a complementary set of other covariates. It is a very general importance measure and can be applied to any machine learning approach. Five-fold cross-validation was used when estimating the LogLoss for the CPI.

2.4.2 | Application 2: diagnosis of coronary artery disease—Cleveland Clinic data

In the second application, the aim was to predict coronary artery disease (CAD) using data from the Cleveland Clinic project. The reference standard for diagnosing CAD is coronary angiography, which is expensive, involves radiation exposure and a small risk of complications and death.[51] Therefore, noninvasive testing was of interest to select patients benefiting from coronary angiography. Clinical and test characteristics of 303 consecutively recruited patients were used to predict CAD from patients referred for coronary angiography to the Cleveland Clinic between May 1981 and September 1984. These patients additionally underwent exercise electrocardiogram, thallium scintigraphy and cardiac fluoroscopy, which are all noninvasive. Details of the data collection can be found elsewhere.[52]

Data from three other centers (Veterans Administration Medical Center in Long Beach, California (VA); the Hungarian Institute of Cardiology, Budapest (Hunagarian); the University Medical Centers Zurich and Basel, Switzerland (Swiss)) with a total of 617 patients were used for external validation.[38] In these groups, non-invasive test results were not withheld from the treating physician and might have influenced the decision to perform coronary angiography. Descriptive statistics of these patients are provided in Table 5. The dependent variable was the diagnosis of CAD as determined by the presence or absence of a $> 50\%$ diameter in an angiography,[52] and independent variables are described in Supplementary Material 1. Missing data were imputed for pragmatic reasons prior to analysis in the same way as for Application 1. For sake of simplicity, backward variable selection was used for logistic regression with $P < 0.001$; for further details, see the R code in Appendix S4. Regression coefficients for the logistic regression model are provided in the Appendix S1. CPI was estimated as it was done for Application 1.

The R code for the analysis of the Cleveland Clinic data is provided as Appendix S4, and the results from this analysis are available as Appendix S5.

2.5 | Software packages for estimation, tuning, and calibration

LogReg was fitted using the `glm` function in the `stats` package. When fitting LogReg in the simulation study, no variable selection technique was applied. EN has been proposed by Zou and Hastie,[53] and it was fitted using the `glmnet`[54] package with the `binomial` family. The hyperparameters tuned were the penalty parameter, defining the amount of regularization and the mixing parameter. For GB, reviews are available.[55-57] In this work, we used the `xgboost` package to fit boosted trees using the `logloss` loss function. Tuning was done for the following hyperparameters: depth of the trees, number of trees, learning rate, reduction in the loss function required to split further, proportion of candidate variables sampled at each split, proportion of observations used to fit each tree, and minimum number of observations in a terminal node. RFs have been reviewed, for example, in Reference 58. The `ranger`[59] package was used to grow probability forests. The number of trees was set to 500, and the hyperparameters tuned were the number of candidate predictors sampled at each split and the minimum number of observations in a terminal node. SVMs are described in many articles and textbooks, such as References 60 and 61, and the package `kernlab`[62] was used with a radial kernel. The hyperparameters tuned were the cost of constraint violations and the inverse kernel width for the radial basis function. In the simulation study, five times repeated five-fold cross-validation was used to select the combination of hyperparameters that minimized the observed LogLoss.

In the real data analyses no replicates are available. Bootstrapping with 500 bootstrap replicates was done in all modelling steps for both the training and the validation data. These modelling steps included tuning and calibration. Within each bootstrap step, we reduced cross-validation to two times repeated two-fold cross-validation for selecting the hyperparameters because of high computational demands. For the estimation of the performance measures of the different calibration approaches in the validation data, 10-fold cross-validation was used to obtain honest performance estimates, and this was within each bootstrap step.

For all learning machines for probability estimation we used `tidymodels` as common interface in conjunction with `parsnip`, `dials`, and `tune`. EN, GB, RF, and SVM were tuned using a space-filling grid of hyperparameters using the function `grid_latin_hypercube` of size 50. For the simulation study, distribution estimates were obtained from the replicates within each simulation scenario. For the real data, replicates were not available.

The `glm` function from the `stats` package was used to implement the LogReg-based calibration approaches (Beta, Logistic, Platt, Slope), also see Table 1. RCS calibration was fitted using the functions `lrm` and `rcs` from the `rms` package.[21] Isotonic calibration was done with help of the `isoreg` function from the `stats` package. The function `density` from the `stats` package was used to obtain the kernel density estimates with a Gaussian kernel for the Chen calibration approach.

Critical differences and related plots for comparing the calibration approaches were generated with the help of the `performanceEstimation` package.[63]

3 | RESULTS

3.1 | Simulation study

The complete results of the simulation study are provided in Appendices S3a and S3b. Figure 1 illustrates calibration plots (panel A), boxplots for the LogLoss (panel B), and critical difference plots (panel C) from the logistic regression model together with simulation scenario 11. Simulation scenario 11 is characterized by 50 noise variables in addition to five continuous and five categorical predictors (Table 3). Panel A of Figure 1 shows the clear deviation from the expected angle bisector, when calibration is not done (NoCal). Subsequently, NoCal performed worst on average (panel B) with an average rank just below 10 among the 10 used calibration approaches (panel C). In contrast, logistic calibration, beta calibration, and RCS performed well in simulation scenario 11. With the CD being 1.4, logistic calibration and beta calibration formed a homogeneous subgroup, while Platt calibration performed significantly worse than logistic calibration. However, Platt calibration formed a homogeneous subgroup with beta calibration (panel C).

The average over all machine learning approaches is displayed for all simulation scenarios in the critical difference plots in Figure 2 for the sample size of 1000. The results for the sample size of 2000 are provided in Appendix S3b.

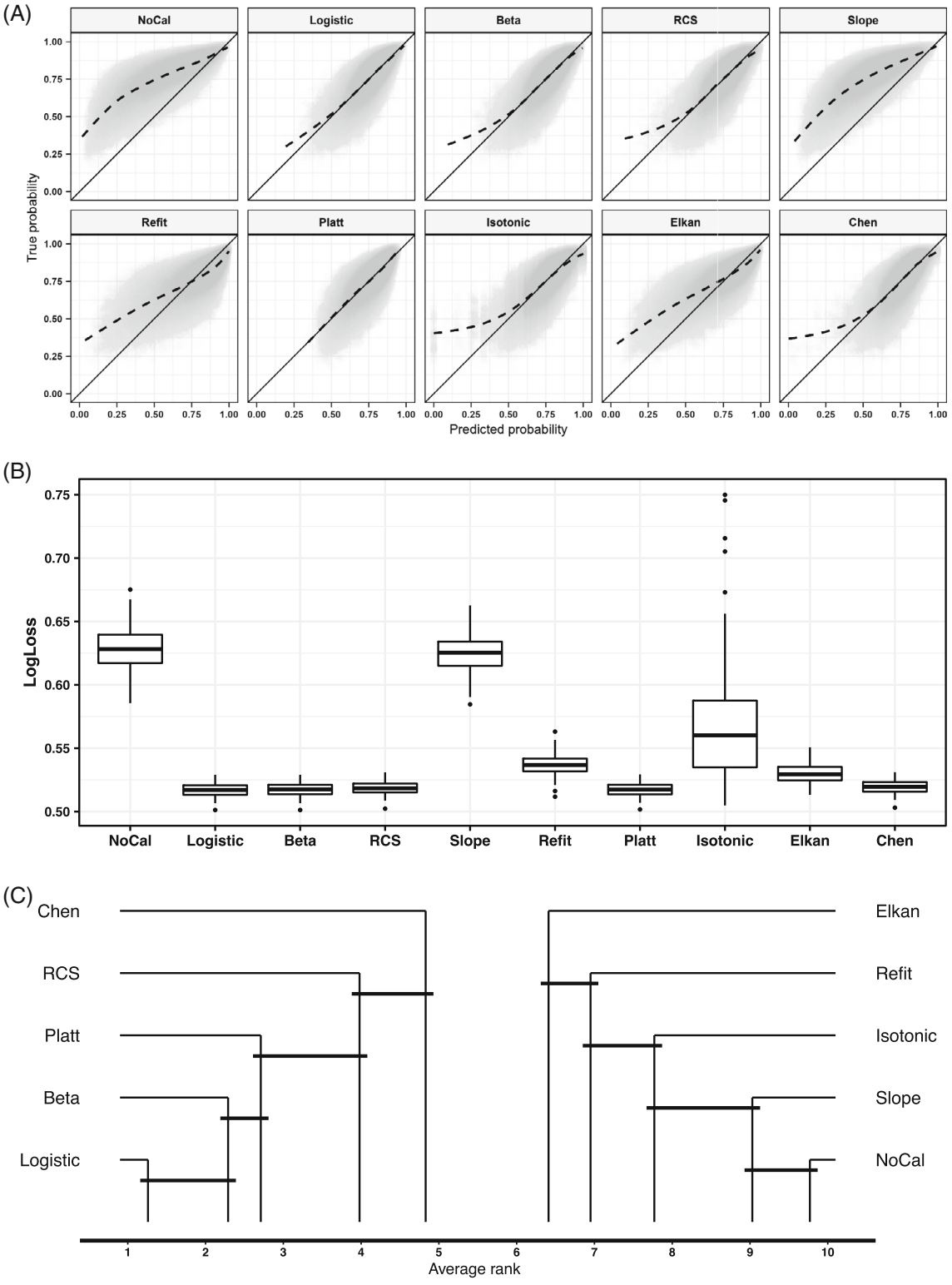

FIGURE 1 Calibration plots (panel A), boxplots for the LogLoss (panel B) and critical difference plot (panel C) for logistic regression in simulation scenario 11. The boxplots and critical difference plot are based on the LogLoss. The calibration plots shows predicted versus true probabilities. A solid line showing the angle bisector is displayed for orientation, and a dashed line provides the generalized additive model (GAM) estimates as smoother in the scatterplot. The lower the average rank of the calibration approach in panel C, the better the performance of the calibration approach. Groups of calibration approaches connected by a horizontal line segment have a lower average rank difference than the critical difference (CD), which is 1.4 in this case, and they form a homogeneous subgroup. They could thus not be shown to perform significantly differently.

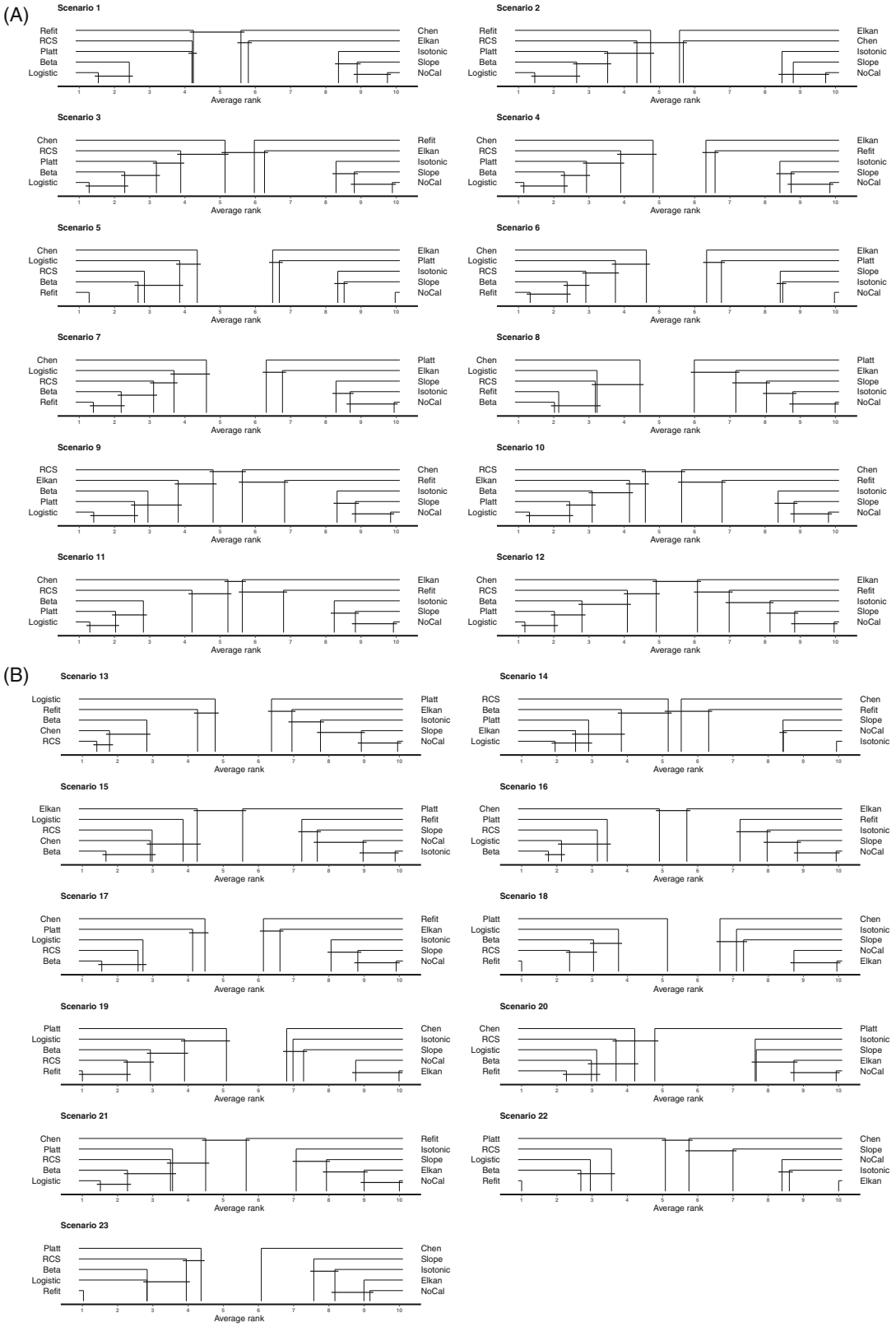

FIGURE 2 (A) Critical difference plots based on the LogLoss for simulation scenarios 1 to 12. Results are calculated over all learning machines for probability estimation. The critical difference equals 1.4. Lower average ranks are considered better. Groups of calibration approaches connected by a horizontal line segment could not be shown to have a significantly different performance. (B) Critical difference plots based on the LogLoss for simulation scenarios 13 to 23. Results are calculated over all learning machines for probability estimation. The critical difference equals 1.4. Lower average ranks are considered better. Groups of calibration approaches connected by a horizontal line segment could not be shown to have a significantly different performance.

As described before for simulation scenario 11, lower average ranks correspond to better performance. As expected, NoCal performed worst. Figure 2 shows that NoCal often formed a homogeneous subgroup with slope calibration. In general, the regression-based calibration approaches logistic calibration, beta calibration, Platt calibration, and RCS calibration performed well. Specifically, logistic calibration and beta calibration outperformed all other calibration methods in simulation scenarios 1 to 4 (two continuous covariates, varying number of noise variables). Specifically, they also outperformed refitting the model in a calibration training dataset before testing it on a calibration test dataset.

Scenarios 5 to 8 were analogous to scenarios 1 to 4, but had a low disease prevalence in the model building data. In these scenarios, refitting the data performed best in case of no or just few noise variables with beta calibration forming a homogeneous subgroup, when there were 10 noise variables or more. RCS calibration and logistic calibration came in next in these scenarios. With an increasing number of noise variables the performance of the data refit decreased, and for 100 noise variables (scenario 8) beta calibration had a slightly lower average rank than the refit. Concurrently, there was a tendency of improved performance of logistic calibration.

In simulation scenarios 9 to 12 with five continuous and five categorical covariates, logistic calibration performed best, formed a homogeneous subgroup with Platt calibration and slightly outperformed beta calibration, which had the third best performance. Refitting the model did not perform well in these models.

Simulation scenario 13 is characterized by strong nonlinearity, and RCS calibration performed best on these data, closely followed by Chen calibration. Beta calibration, refit and logistic calibration were next. When in addition to the strong nonlinearity the dataset available for training the calibration model was small (scenario 15), beta calibration performed well, and RCS calibration performed similarly to logistic calibration. As expected, refit performed badly in case of small datasets available for estimating the calibration model (scenarios 14 and 15). Scenario 14 was designed to demonstrate the strength of Elkan calibration because it only requires information on base disease probabilities and the use of a Bayes formula for application. Unexpectedly, logistic calibration tended to perform better, forming a homogeneous subgroup with Elkan calibration and Platt calibration.

Simulation scenarios 16 to 23 were created to investigate differences in the covariate distribution or differences in the effect of the covariate(s). The assumption of equal covariate distribution underlying Elkan calibration is, by design, violated in all these scenarios. Elkan calibration thus performed poorly in these scenarios. It is important to note that isotonic calibration, which is often used in machine learning, completely failed when the regression coefficients for the model building data and the calibration data had different signs, specifically in scenarios 18, 19, and 21, and its rank was not better than seven in simulation scenarios 16 to 23. The best-performing calibration approaches in these situations were beta calibration, logistic calibration, RCS calibration, and refitting the model. The refit specifically outperformed the calibration approaches, when the distribution of the covariate changed its sign from positive to negative, see simulation scenarios 18, 22, and 23.

The CD plots of Figure 2 provide an overview of the performance of calibration approaches even across different machine learning approaches. Specific learning machines are worth mentioning in conjunction with specific simulation scenarios, and the corresponding calibration plots are displayed in Figure 3A. The top row of Figure 3A shows the calibration plots from simulation scenario 1, that is, no noise variables, on the left and from simulation scenario 4, that is, 100 noise variables, on the right for the LogReg model as learning machine for probability estimation. The figures demonstrate that the variability of the estimates increase with the number of noise variables. Furthermore, while logistic calibration, beta calibration, RCS calibration and refit performed very well in simulation scenario 1, the dashed GAM line clearly deviated from the angle bisector in the refit for this model. The GAM curve for RCS calibration also showed some minor deviation from the angle bisector. The calibration plots are also displayed for EN and RF in simulation scenario 4 (Figure 3A, row 2). EN performed substantially better in this simulation scenario than LogReg, which has only two continuous variables influencing the dichotomous outcome, than logistic regression because of the penalization introduced in EN. In contrast, RF performed worse than LogReg, and a deviation from the angle bisector can be noticed for all calibration approaches (Figure 3A, row 2, column 2). However, EN does not perform well in all simulation scenarios. For example, the nonlinear Mease model of simulation scenario 13 (Figure 3A, row 3, column 2) shows that no calibration approach was able to reliably estimate probabilities. The lowest average rank in the corresponding CD plot therefore was even higher than 4 (Figure 3B, row 3, column 2).

The good performance of Elkan calibration in case of a small calibration dataset (simulation scenario 14), when the model assumptions are met, is shown in Figure 3A, row 3, column 2, for EN. It was only outperformed by logistic calibration in this setting (Figure 3B). The smaller calibration dataset reveals bad performance for the model refit as well as for Chen calibration, isotonic calibration, RCS calibration, and even beta calibration. In fact, isotonic calibration shows a

better performance when sample sizes are large Appendix S3b). Smaller datasets show that the forced monotonicity may lead to fringing along the *x*-axis. A similar pattern can be observed for Chen calibration. This may be explained the necessity to estimate the densities separately for both groups of subjects. A smaller amount of fringing can be noted for beta calibration and RCS calibration for EN, simulation scenario 14. While the model assumptions for Elkan calibration were met in simulation scenario 14, they were violated in simulation scenario 20. In consequence, no calibration, Elkan calibration, and slope calibration performed poorly in this simulation scenario for all machines (SVM displayed in Figure 3A, bottom row, left). This figure also shows the generally larger variability of isotonic calibration and to a lesser extent also for Chen calibration compared to regression-based calibration approaches. For this simulation model 20 and SVM as machine learning approach some minor fringing can be observed in case of low true probabilities. Finally, the bottom right plots in Figure 3.a demonstrate that isotonic calibration and Elkan calibration do not provide reliable probability estimates for the calibration data in case of covariate distributions with an altered sign. For the RF model in the displayed simulation scenario 21, the most reliable machine learning approaches for calibration were the regression-based approaches logistic calibration, beta calibration, and RCS calibration.

3.2 | Application 1: prediction of functional outcome after stroke

Radar plots from the final LogReg and the RF models are displayed in Appendix S1 for both the complete restitution model and the mortality model. Odds ratios are displayed for LogReg and CPI for RF. In brief, radar plots in the complete restitution model appeared to be similar within LogReg and RF, respectively, while CPIs differed between datasets in the mortality model. Since the outcome data are unbalanced in the mortality model with an approximate 1:9 ratio, these findings indicate the larger variability in the mortality model, which may result in lower model stability.

Boxplots for the calibrated learning machines using the LogLoss as performance measure are shown in panel A of Figure 4 for the complete restitution model of the stroke data. The refit of RF, SVM, and GB yielded by far the best performing machines in both the temporal and the external validation data. Furthermore, no calibration, intercept

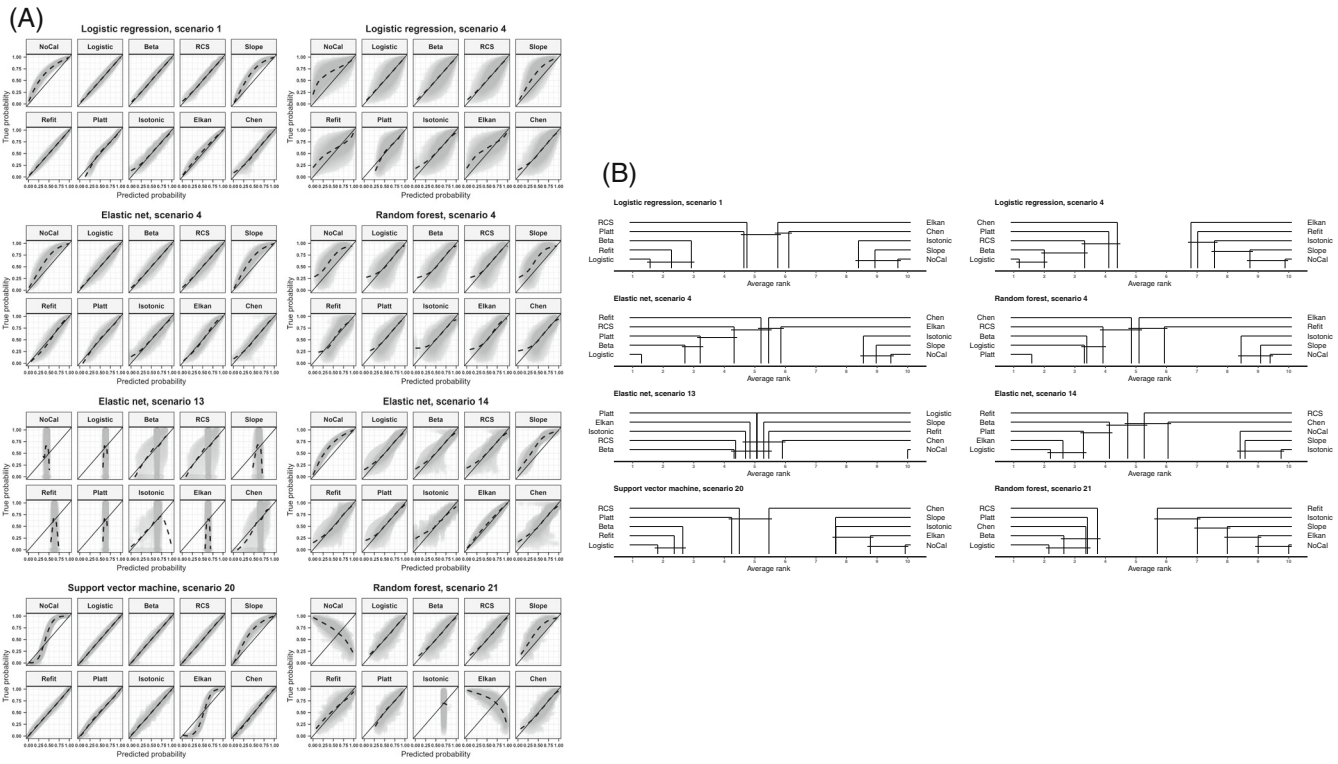

FIGURE 3 (A) Calibration plots for selected simulation scenarios and learning machines for probability estimation. (B) Critical difference plots for selected simulation scenarios and learning machines for probability estimation. The critical difference equals 1.4. Lower average ranks are considered better. Groups of calibration approaches connected by a horizontal line segment could not be shown to have a significantly different performance.

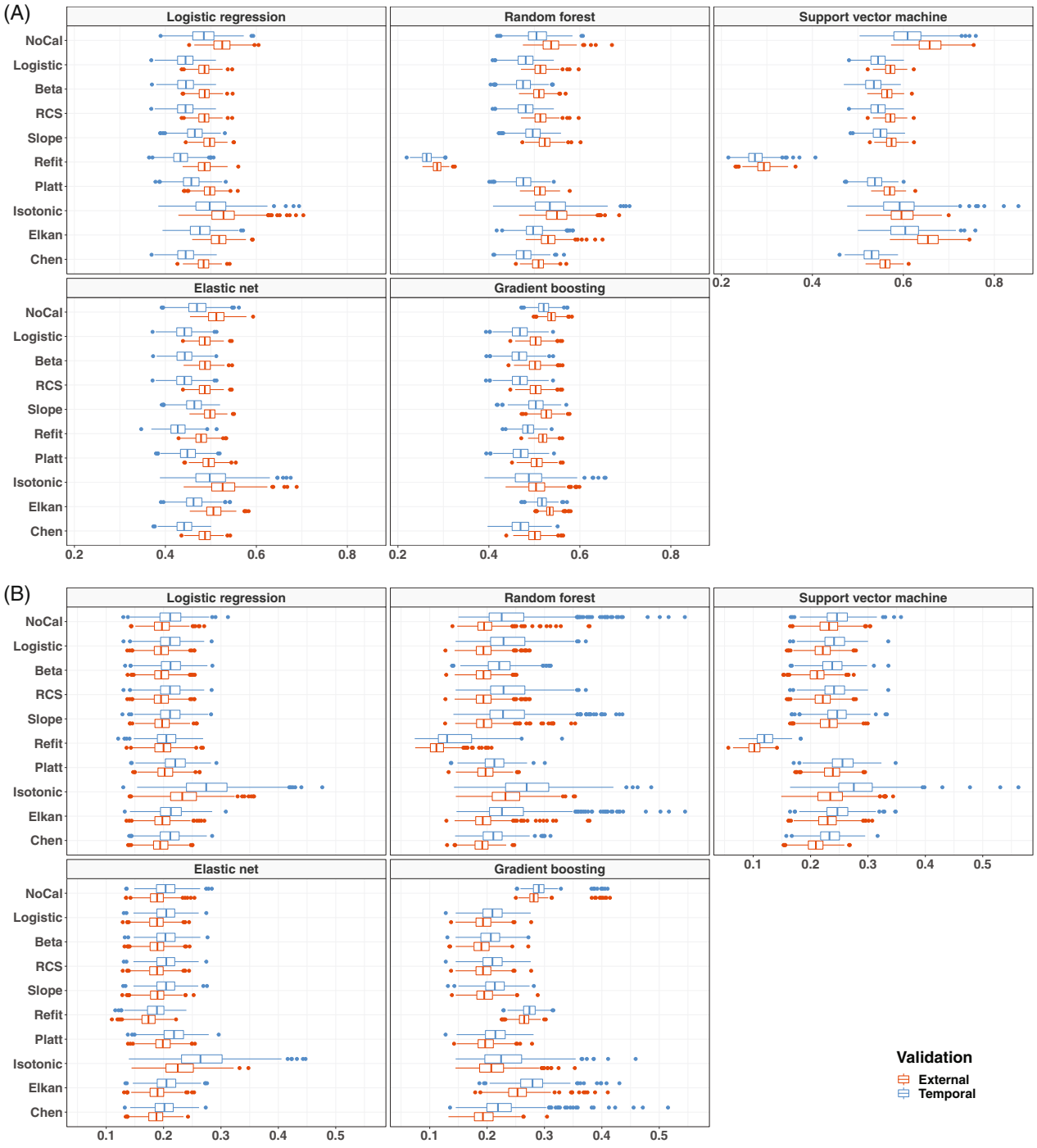

FIGURE 4 Boxplots for calibrated learning machines for probability estimation in the complete restitution and mortality model of the stroke data using the LogLoss (x-axis) as performance measure. Panel A: complete restitution model; panel B: mortality model.

calibration, isotonic calibration, and Elkan calibration performed badly on several machines on the stroke data. This is confirmed by the calibration plots for the complete restitution model, which are displayed for SVM and RF in the first and second panels of Figure 5, respectively.

Patterns differ for the mortality model. Figure 4, panel B shows that the refit performed badly on GB, and Chen, Elkan, and isotonic calibration had many outliers for GB. The refit performed best for RF and SVM. Overall, isotonic calibration and Elkan calibration showed larger LogLoss values than other machines. Calibration plots for the calibrated RFs and the calibrated SVMs (panels C and D of Figure 5) provide additional information. They show that predicted

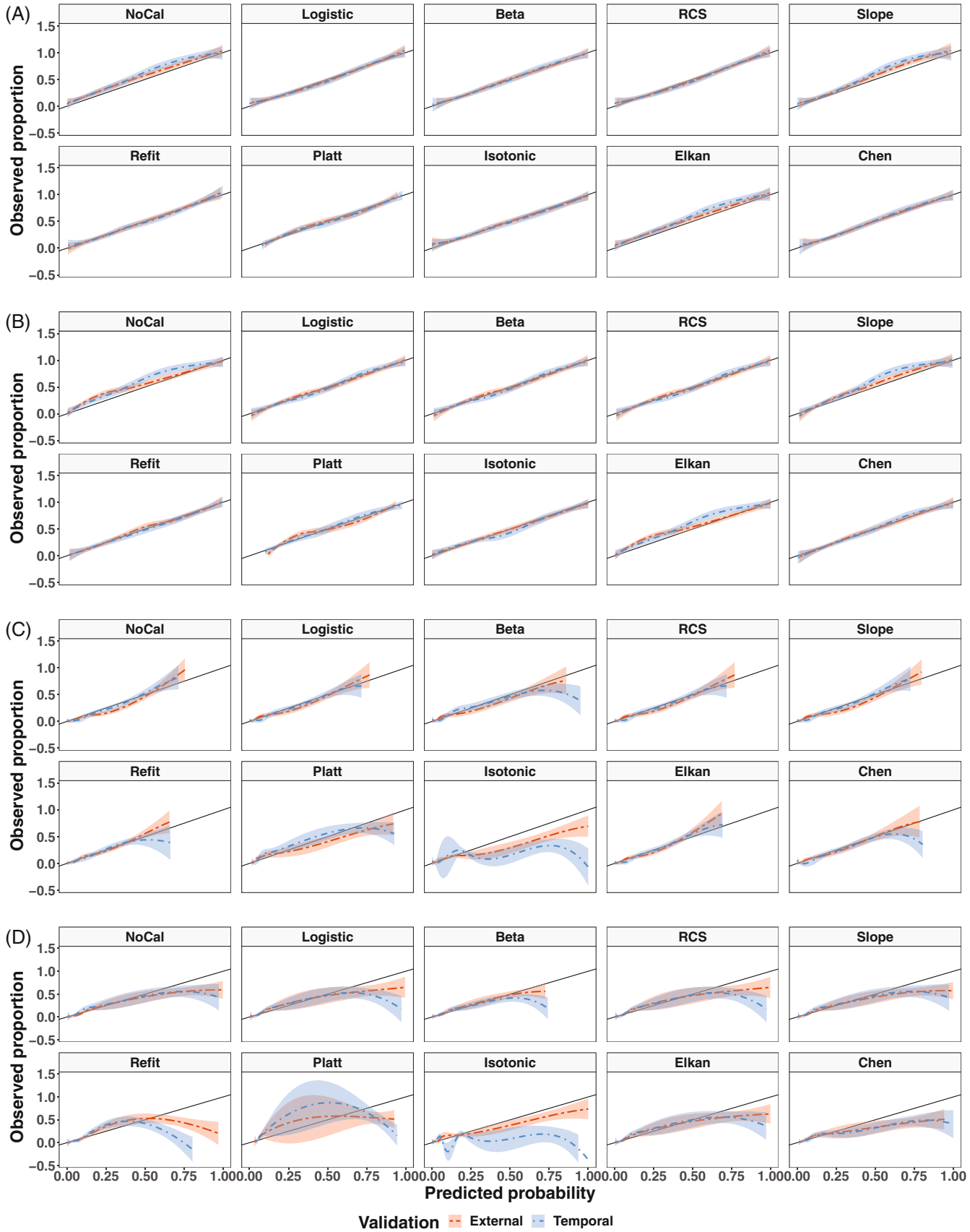

FIGURE 5 Calibration plot for calibrated random forests (RF) and support vector machines (SVM) in the complete restitution model (CRM) and mortality model (MM) of the stroke data.Panel A: CRM for RF; panel B: CRM for SVM; panel C: MM for RF; panel D: MM for SVM.

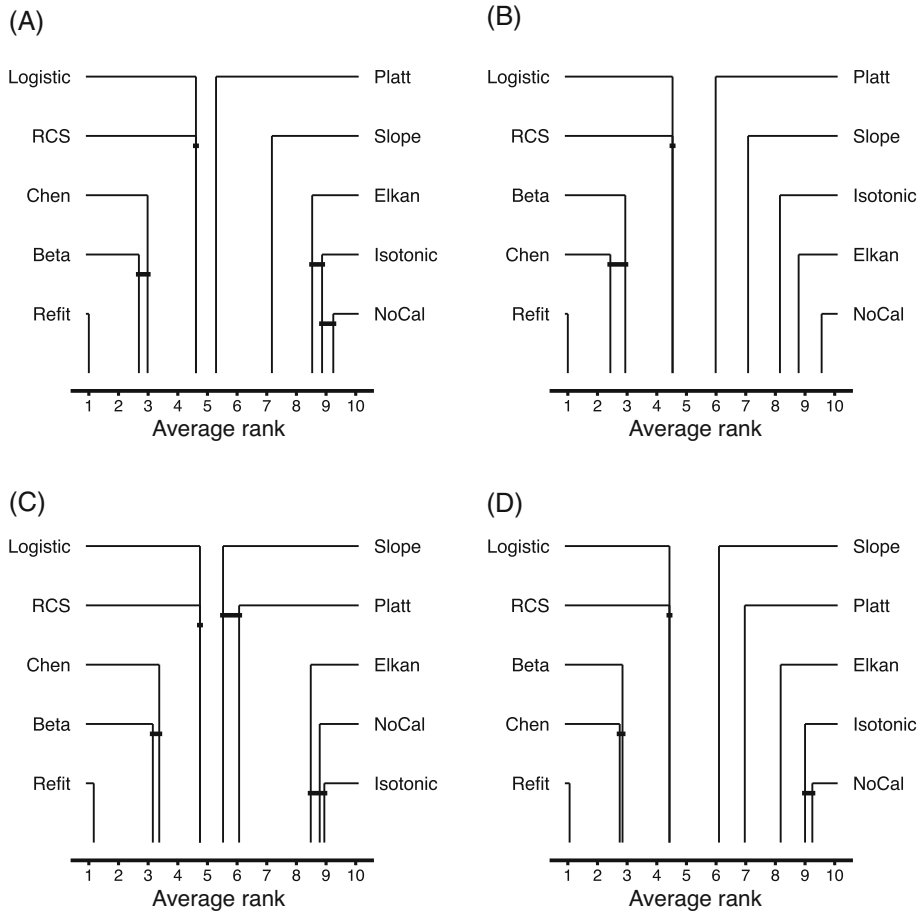

FIGURE 6 Critical difference plot for the LogLoss in the German Stroke Study Collaboration averaged over all learning machines for probability estimation. Lower average ranks are considered better. Groups of calibration approaches connected by a horizontal line segment could not be shown to have a significantly different performance. The critical difference was 0.6.Panel A: complete restitution model (CRM) in the temporal validation data (TVD); panel B: CRM in the external validation data (EVD); panel C: mortality model (MM) in TVD; panel D: MM in EVD.

probabilities deviate from observed proportions for several calibration approaches for the mortality model. While none of the calibration approaches for the SVM showed the expected pattern for the temporal validation data (panel D), logistic calibration, RCS calibration, Elkan calibration, and the refit revealed a good agreement between observed proportions and predicted probabilities in both the temporal and the external validation data for RF (panel C).

Figure 6 provides the CDs for the LogLoss for both models in both the temporal and the external validation averaged over all five machines. The CDs for each machine are presented in Appendix S1 for the two models and the two types of validation data. The refit showed the best performance across all models. Beta calibration and Chen calibration performed second and third best on the stroke data, followed by logistic calibration.

3.3 | Application 2: diagnosis of coronary artery disease

Radar plots from the final LogReg and the RF models are displayed in Appendix S1 for both the training and the validation data. Odds ratios are shown for LogReg and CPI for RF. Results were homogeneous within a modeling approach, but effect sizes varied between training and validation data. Furthermore, the variables with the greatest impact varied between LogReg and RF.

The boxplots of the LogLoss show that the refit performed best in this dataset for all five machine learning approaches (Figure 7). The expectation is confirmed that Elkan calibration does not perform well on this dataset because of differences in covariate distributions between training and external validation data (Figure 7). Since isotonic calibration requires

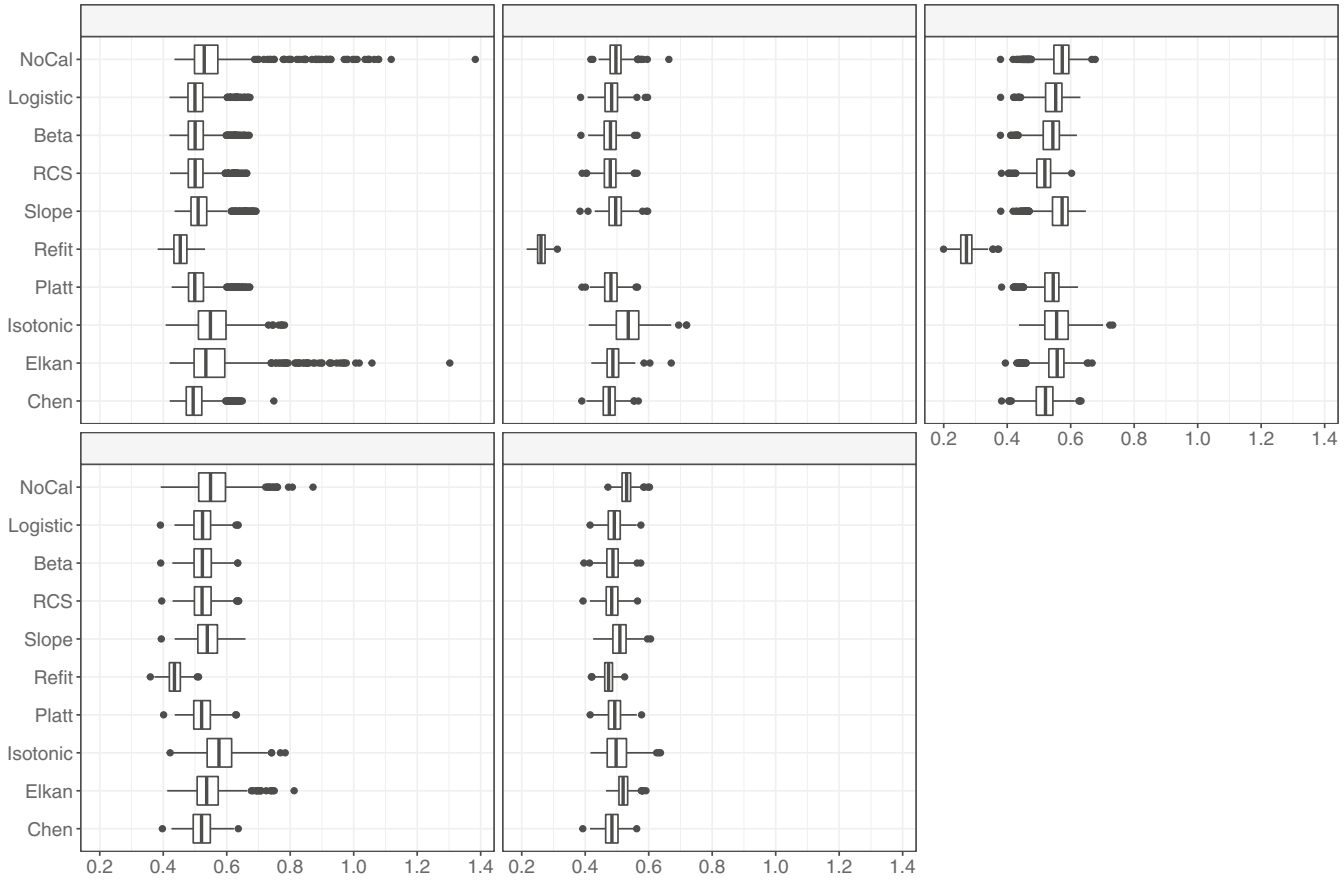

FIGURE 7 Boxplots for calibrated learning machines for probability estimation in the Cleveland Clinic validation data using the LogLoss (x-axis) as performance measure.

large datasets for good performance, but only about 600 subjects were available for validation, the bad performance of isotonic calibration was also expected (Figure 7). The largest homogeneity over the replicates was observed for GB. The CDs for the LogLoss of each machine and the average over all 5 machines are presented in Figure 8. They confirm that the refit performed best on the Cleveland Clinic data. Chen calibration and RCS calibration were second and third best and formed a homogeneous subgroup for most of the models. For RF and GB, beta calibration also performed well. Isotonic calibration, Elkan calibration, and slope calibration performed poorly.

4 | DISCUSSION

The key message is that calibration of any learning machine for probability estimation can be easily done with many different calibration approaches, such as logistic calibration. In the first part of the manuscript, we have reviewed various parametric and nonparametric calibration methods. In a comprehensive simulation study, we have compared 10 different calibration methods, including NoCal and model refit. NoCal turned out to perform poorly in all simulation scenarios and on all machine learning approaches for probability estimation. Similarly, slope calibration failed in almost all simulation scenarios and for almost all learning machines for probability estimation used in this work. The refit performed well, except when the calibration dataset was small. The latter was expected because refitting a model means to build it again.

Both logistic calibration and beta calibration performed well in many simulation scenarios and on real datasets. Beta calibration showed similar but slightly superior performance to logistic calibration in a wide range of simulation scenarios. Beta calibration and logistic calibration also performed well in case of a small dataset available for estimating calibration parameters. This may be explained by the complexity of the calibration model, that is, the number of parameters to be estimated. For logistic calibration only two parameters are estimated, while three parameters need to be estimated

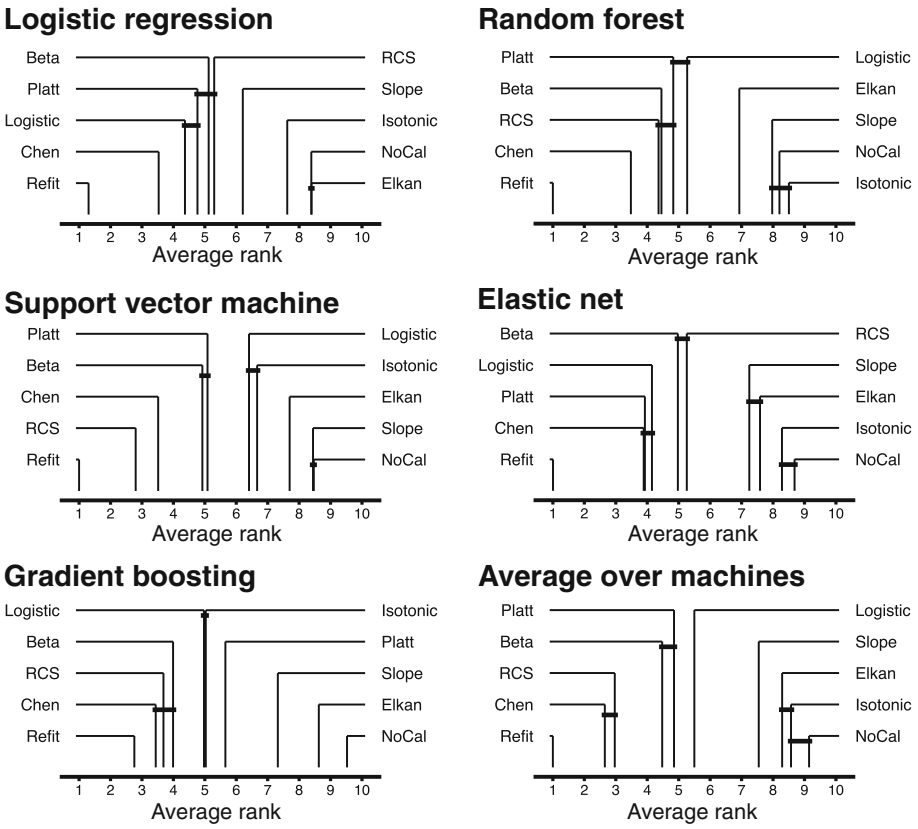

FIGURE 8 Critical difference plot for the LogLoss in the Cleveland Clinic validation data for each machine and averaged over all learning machines for probability estimation. Lower average ranks are considered better. Groups of calibration approaches connected by a horizontal line segment could not be shown to have a significantly different performance. The critical difference was 0.6.

for beta calibration. And the most complex regression approach for calibration is RCS calibration. In fact, Figure 2B (simulation scenario 13) confirms this finding, and the performance of the regression-based calibration methods increases with decreasing model complexity.

The theoretical relationship between logistic calibration and beta calibration has been discussed elsewhere.[8,9] The simulation study also demonstrated the close agreement between these two calibration methods. Böken[9] has additionally demonstrated the optimality of logistic calibration in various scenarios. However, we stress that logistic calibration failed in some simulation scenarios and that other methods outperformed logistic calibration in these settings.

In case of small calibration datasets, Elkan calibration can outperform the other calibration approaches when the underlying model assumptions are met. However, the crucial assumption for Elkan calibration is that the covariate distribution of the (un)affected subjects are identical in both the training and the validation data.[13] When this assumption is not met, Elkan calibration may lead to substantial biases in probability estimates.

In this work, we did not consider intercept calibration because it can adjust for changes in the baseline probabilities only. It may thus lead to biased probability estimates in case of structural differences between model building and calibration data, in analogy to Elkan's calibration. This means that intercept calibration has the weakness of Elkan calibration in the model assumptions, and it also requires a regression model for estimating the calibration parameter.

To distinguish between the use of estimated probabilities for calibration and the estimated linear predictor, we used the terms Platt calibration and logistic calibration. Platt calibration performed slightly worse compared to logistic calibration. We thus generally recommend the use of the linear predictor in logistic regression-based calibration over Platt calibration.

Isotonic calibration[8,12] is one of the most commonly used calibration approaches by the machine learning community.[64] In the simulation study, it did not perform well, and it performed worst on both real datasets. It appeared that the performance of isotonic calibration is sample size dependent. In our opinion, this is caused by the monotonicity assumption for the probabilities. It is best seen in the simulation scenarios, where signs of regression coefficients changed, such as simulation scenario 21 for RF (Figure 3A, bottom right), where calibrated probability estimates were almost identical for all subjects. Sample size for estimating calibration depends not on the total sample size but the size of the smaller

group, which determines the events per predictor (EPP).[65] This may, in turn, be related to the question when is a sample sufficiently large to do a refit of the model on the validation data rather than calibrating the model. We thus conclude that further studies are needed to investigate the required sample size for a good performance of isotonic calibration in general settings.

Chen calibration performed well in their original publication.[32] A limitation of this work is that we explicitly formulated the Chen calibration as a general LR approach and selected a standard method available in R for estimating the LR. Chen calibration generally performed worse than other calibration approaches in our simulation study. Similar to isotonic calibration, Chen calibration requires larger datasets to reliably estimate the likelihood function separately for affected and unaffected subjects. We cannot rule out that a different implementation of this GUA for calibration might perform better.

Another limitation concerns the simulation study itself. As pointed out by one expert reviewer, there are many factors that can affect calibration performance of a machine learning model in validation data. We have therefore included scenarios with different combinations of number of covariates and noise variables and differences between the original development and external datasets. There are other factors that affect model performance, such as how continuous predictors are modelled and tested, whether nonlinearity is present, how high outcome proportions are, and how large the total sample size is. In our simulation study, we have tried to address all these aspects. To make the simulation study fully transparent and easily expandable, we have made available the code for the entire simulation study as Appendix S2.

Generally, the performance of a classifier can be characterized by two fundamental properties, namely discrimination and calibration.[19] While discrimination describes the ability of a classifier to discriminate between cases and controls, measured, for example, by the ROC curve and the area under the ROC curve, calibration describes the degree to which the predicted probability associated with a predictor value $\hat{\eta}_i$ agrees with the relative frequency of cases given $\hat{\eta}_i$. As noted by Chen et al,[32] the calibration function is required to be monotonically increasing in order to calibrate predictors without changing their ability to discriminate. It is therefore reasonable to assume that the predictor—or the initial probability estimate $\hat{\pi}_i$—is monotonically related to the calibrated probability estimate $\hat{p}_i$. Chen et al termed this assumption, which is made, for example, in logistic calibration, "rationality assumption."

The main aim of our study was not to evaluate discrimination but to compare different calibration approaches. With regard to this aspect, we observed substantial differences in the performance of the five different machine learning approaches for probability estimation used in the simulation study and in the real data analysis. In general, calibration based on RF performed poorly in most simulation scenarios. This is in contrast to the real data analysis, where RF combined with logistic calibration or the refit showed good performance. Previously, we demonstrated in a simulation study that there is no single best learning machine for probability estimation,[60] but that learning machines for probability estimation may fail depending on the simulation model. Indeed, we showed using five different simulation settings and 8 different learning machines for probability estimation that each machine failed in at least one simulation scenario completely.[60] In consequence, we cannot derive a general recommendation when to use which combination of a specific machine learning approach for probability estimation with a specific calibration method.

In summary, calibration of learning machines for probability estimation can be done using a plethora of calibration methods. Beta calibration performed well in many simulation scenarios and on real datasets. It showed similar but slightly superior performance to logistic calibration in a wide range of simulation scenarios. Beta calibration would therefore be our calibration method of choice in a wide range of settings. However, the real data analysis and several simulation scenarios have shown excellent performance of a model refit because the dataset for estimating the calibration data was sufficiently large. When the calibration data were small, estimates were better for the simple logistic calibration regression model compared to the calibration models with more parameters. Elkan calibration might be the method of choice if the strong model assumptions of equal covariate distribution between model building and calibration data can be assumed to be met. We recommend to perform a sensitivity analysis for the selected calibration approach by comparing different calibration methods. For future research it would be important that calibration approaches become available for the case of small calibration datasets.

ACKNOWLEDGEMENTS

The German Stroke Study Collaboration was financed by the German Ministry of Education and Research (BMBF) as part of the Competence Net Stroke. We are grateful to the members of the German Stroke Study Collaboration for data collection, including the following departments and responsible study investigators (in alphabetical order): St. Katharinen-Hospital Frechen (R. Adams), Charité Berlin (N. Amberger), Städtisches Krankenhaus München-Harlaching

(K. Aulich, M. J. L. Wimmer), Klinikum Minden (J. Glahn), University of Magdeburg (M. Goertler), Krankenanstalten Gilead Bielefeld (C. Hagemeister), Klinikum München-Großhadern (G. F. Hamann, A. Müllner), Rheinische Kliniken Bonn (C. Kley), University of Rostock (A. Kloth), Benjamin Franklin University of Berlin (C. Koennecke), University of Saarland (P. Kostopoulos), Bürgerhospital Stuttgart (T. Mieck), Universities of Essen (G. Mörger-Kiefer, C. Weimar), Ulm (M. Riepe), Leipzig (D. S. Schneider), and Jena (V. Willig). We thank K. Kraywinkel, MD, MSc, and P. Dommes, PhD, for central data collection and management.

DATA AVAILABILITY STATEMENT

The code for and the results from all simulations is available as Appendices S1 to S5. The Cleveland Clinic data used for illustration are freely available from the machine learning data repository from the University of California in Irvine; details are provided in the Appendices S1 to S5.

ORCID

Francisco M. Ojeda https://orcid.org/0000-0003-4037-144X
Max L. Jansen https://orcid.org/0000-0003-4373-5579
Matthias Schmid https://orcid.org/0000-0002-0788-0317
Andreas Ziegler https://orcid.org/0000-0002-8386-5397

REFERENCES

1. Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. *N Engl J Med*. 1979;300:1350-1358. doi:10.1056/NEJM197906143002402
2. Xie G, Wang R, Shang L, et al. Calculating the overall survival probability in patients with cervical cancer: a nomogram and decision curve analysis-based study. *BMC Cancer*. 2020;20:833. doi:10.1186/s12885-020-07349-4
3. Boyer B, Cazorla C. Methods and probability of success after early revision of prosthetic joint infections with debridement, antibiotics and implant retention. *Orthop Traumatol Surg Res*. 2021;107:102774. doi:10.1016/j.otsr.2020.102774
4. Uttley AM. Temporal and spatial patterns in a conditional probability machine. In: Shannon CE, McCarthy J, eds. *Automata Studies*. Princeton: Princeton University Press; 1956:277-285.
5. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med*. 2000;19:453-473. doi:10.1002/(SICI)1097-0258(20000229)19:4<453::AID-SIM350>3.0.CO;2-5
6. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130:515-524. doi:10.7326/0003-4819-130-6-199903160-00016
7. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167-176. doi:10.1016/j.jclinepi.2015.12.005
8. Kull M, Silva Filho TM, Flach P. Beyond sigmoids: how to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electron J Statist*. 2017;11:5052-5080. doi:10.1214/17-EJS1338SI
9. Böken B. On the appropriateness of Platt scaling in classifier calibration. *Inf Syst*. 2021;95:101641. doi:10.1016/j.is.2020.101641
10. Platt J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Smola AJ, Bartlett PJ, Schölkopf B, Schuurmans D, eds. *Advances in Large Margin Classifiers*. Cambridge: MIT Press; 2000:61-74.
11. Fawcett T, Niculescu-Mizil A. PAV and the ROC convex hull. *Mach Learn*. 2007;68:97-106. doi:10.1007/s10994-007-5011-0
12. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In: Hand D, Keim DA, Ng R, eds. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery; 2002:694-699. doi:10.1145/775047.775151
13. Elkan C. The foundations of cost-sensitive learning. In: Nebel B, ed. *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*. Vol 2. San Francisco: Morgan Kaufmann; 2001:973-978.
14. Dankowski T, Ziegler A. Calibrating random forests for probability estimation. *Stat Med*. 2016;35:3949-3960. doi:10.1002/sim.6959
15. Dua D, Graff C. *UCI Machine Learning Repository*. Irvine, CA: School of Information and Computer Sciences, University of California; 2019. https://archive-beta.ics.uci.edu. Accessed June 1, 2023
16. R Core Team. R: a language and environment for statistical computing. 2022.
17. Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. *Stat Med*. 1991;10:1213-1226. doi:10.1002/sim.4780100805
18. Cox DR. Two further applications of a model for binary regression. *Biometrika*. 1958;45:562-565. doi:10.2307/2333203
19. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 2nd ed. Cham: Springer; 2019.
20. Lucena B. Spline-based probability calibration. *arXiv* 2018: 1809.07751. https://arxiv.org/abs/1809.07751. Accessed June 1, 2023.
21. Harrell, F. E., Jr.. *Regression Modeling Strategies. With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. Cham: Springer; 2015.

22. Zhang J, Yang Y. Probabilistic score estimation with piecewise logistic regression. In: Greiner R, Schuurmans D, eds. *Proceedings of the 21st International Conference on Machine Learning*. New York: ACM Press; 2004:115-123.
23. Dormann CF. Calibration of probability predictions from machine-learning and statistical models. *Glob Ecol Biogeogr*. 2020;29:760-765. doi:10.1111/geb.13070
24. Landwehr N, Hall M, Frank E. Logistic model trees. *Mach Learn*. 2005;59:161-205. doi:10.1007/s10994-005-0466-3
25. Leathart T, Frank E, Holmes G, Pfahringer B. Probability calibration trees. In: Min-Ling Z, Yung-Kyun N, eds. *Proceedings of the 9th Asian Conference on Machine Learning*. Cambridge, MA: ML Research Press; 2017:145-160. http://proceedings.mlr.press/v77/leathart17a.html. Accessed June 1, 2023.
26. de Leeuw J, Hornik K, Mair P. Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. *J Stat Softw*. 2009;32:1-24. doi:10.18637/jss.v032.i05
27. Bella A, Ferri C, Hernández-Orallo J, Ramírez-Quintana MJ. Calibration of machine learning models. In: Soria Olivas E, Martín Guerrero JD, Martinez Sober M, Magdalena Benedito JR, Serrano López AJ, eds. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. Hershey: IGI Global; 2010:128-146. doi:10.4018/978-1-60960-818-7.ch104
28. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc*. 2020;27:621-633. doi:10.1093/jamia/ocz228
29. Dimitriadis T, Gneiting T, Jordan AI. Stable reliability diagrams for probabilistic classifiers. *Proc Natl Acad Sci*. 2021;118:e2016191118. doi:10.1073/pnas.2016191118
30. Naeini MP, Cooper GF, Hauskrecht M. Obtaining well calibrated probabilities using Bayesian binning. *Proc Conf AAAI Artif Intell*. 2015;2015:2901-2907.
31. Zadrozny B, Elkan C. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In: Brodley CE, Danyluk AP, eds. *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*. Burlington: Morgan Kaufmann; 2001:609-2616.
32. Chen W, Sahiner B, Samuelson F, Pezeshk A, Petrick N. Calibration of medical diagnostic classifier scores to the probability of disease. *Stat Methods Med Res*. 2018;27:1394-1409. doi:10.1177/0962280216661371
33. Bella A, Ferri C, Hernández-Orallo J, Ramírez-Quintana MJ. Similarity-binning averaging: a generalisation of binning calibration. In: Corchado E, Yin H, eds. *Intelligent Data Engineering and Automated Learning – IDEAL 2009*. Berlin: Springer; 2009:341-349. doi:10.1007/978-3-642-04394-9_42
34. Biau G, Cérou F, Guyader A. Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Transact Inform Theor*. 2010;56:2034-2040. doi:10.1109/TIT.2010.2040857
35. Bella A, Ferri C, Hernández-Orallo J, Ramírez-Quintana MJ. On the effect of calibration in classifier combination. *Appl Intell*. 2013;38:566-585. doi:10.1007/s10489-012-0388-2
36. Jiang X, Osl M, Kim J, Ohno-Machado L. Calibrating predictive model estimates to support personalized medicine. *J Am Med Inform Assoc*. 2012;19:263-274. doi:10.1136/amiajnl-2011-000291
37. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Wea Rev*. 1950;78:1-3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
38. Kruppa J, Liu Y, Diener HC, et al. Probability estimation with machine learning methods for dichotomous and multi-category outcome: applications. *Biom J*. 2014;56:564-583. doi:10.1002/bimj.201300077
39. Vovk V. The fundamental nature of the log loss function. In: Beklemishev LD, Blass A, Dershowitz N, Finkbeiner B, Schulte W, eds. *Fields of Logic and Computation II: Essays Dedicated to Yuri Gurevich on the Occasion of his 75th Birthday*. Cham: Springer; 2015:307-318. doi:10.1007/978-3-319-23534-9_20
40. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc*. 2007;102:359-378. doi:10.1198/016214506000001437
41. Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A. Probability machines: consistent probability estimation using nonparametric learning machines. *Methods Inf Med*. 2012;51:74-81. doi:10.3414/ME00-01-0052
42. Demšar J. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res*. 2006;7:1-30.
43. Mease D, Wyner AJ, Buja A. Boosted classification trees and class probability/quantile estimation. *J Mach Learn Res*. 2007;8:409-439.
44. Weimar C, Ziegler A, König IR, Diener HC. On behalf of the German stroke study collaborators. Predicting functional outcome and survival after acute ischemic stroke. *J Neurol*. 2002;249:888-895. doi:10.1007/s00415-002-0755-8
45. Weimar C, König IR, Kraywinkel K, Ziegler A, Diener HC, German Stroke Study Collaboration. Age and National Institutes of Health stroke scale score within 6 hours after onset are accurate predictors of outcome after cerebral ischemia: development and external validation of prognostic models. *Stroke*. 2004;35:158-162. doi:10.1161/01.STR.0000106761.94985.8B
46. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Bmj*. 2015;350:g7594. doi:10.1136/bmj.g7594
47. König IR, Weimar C, Diener HC, Ziegler A. Vorhersage des Funktionsstatus 100 Tage nach einem ischämischen Schlaganfall: design einer prospektiven Studie zur externen Validierung eines prognostischen Modells. *Z Arztl Fortbild Qualitatssich*. 2003;97:717-722.
48. Mahoney FI, Barthel DW. Functional evaluation: the Barthel index. *Md Med J*. 1965;14:56-61.
49. König IR, Malley JD, Weimar C, Diener HC, Ziegler A. On behalf of the German stroke study collaboration. Practical experiences on the necessity of external validation. *Stat Med*. 2007;26:5499-5511. doi:10.1002/sim.3069
50. Watson DS, Wright MN. Testing conditional independence in supervised learning algorithms. *Mach Learn*. 2021;110:2107-2129. doi:10.1007/s10994-021-06030-6

51. Seiffert M, Ojeda F, Müllerleile K, et al. Reducing radiation exposure during invasive coronary angiography and percutaneous coronary interventions implementing a simple four-step protocol. *Clin Res Cardiol*. 2015;104:500-506. doi:10.1007/s00392-015-0814-7
52. Detrano R, Janosi A, Steinbrunn W, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. *Am J Cardiol*. 1989;64:304-310. doi:10.1016/0002-9149(89)90524-9
53. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B*. 2005;67:301-320. doi:10.1111/j.1467-9868.2005.00503.x
54. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33:1-22. doi:10.18637/jss.v033.i01
55. Bühlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting. *Stat Sci*. 2007;22:477-505. doi:10.1214/07-STS242
56. Hofner B, Mayr A, Robinzonov N, Schmid M. Model-based boosting in R: a hands-on tutorial using the R package mboost. *Comput Stat*. 2014;29:3-35. doi:10.1007/s00180-012-0382-5
57. Mayr A, Binder H, Gefeller O, Schmid M. The evolution of boosting algorithms – from machine learning to statistical modelling. *Methods Inf Med*. 2014;53:419-427. doi:10.3414/ME13-01-0122
58. Ziegler A, König IR. Mining data with random forests: current options for real-world applications. *WIRE Data Mining Knowl Discov*. 2014;4:55-63. doi:10.1002/widm.1114
59. Wright MN, Ziegler A. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw*. 2017;77:1.
60. Kruppa J, Liu Y, Biau G, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: theory. *Biom J*. 2014;56:534-563. doi:10.1002/bimj.201300068
61. Christmann A, Steinwart I. *Support Vector Machines*. New York: Springer; 2008.
62. Karatzoglou A, Smola A, Hornik K, Zeileis A. Kernlab – an S4 package for kernel methods in R. *J Stat Softw*. 2004;11:1-20.
63. Torgo L. An infra-structure for performance estimation and experimental comparison of predictive models in R. *arXiv* 2015: 1412.0436v4. 2015 https://arxiv.org/abs/1412.0436v4. Accessed June 1, 2023.
64. Xu P, Davoine F, Zha H, Denœux T. Evidential calibration of binary SVM classifiers. *Int J Approx Reason*. 2016;72:55-70. doi:10.1016/j.ijar.2015.05.002
65. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat Methods Med Res*. 2017;26:796-808. doi:10.1177/0962280214558972

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ojeda FM, Jansen ML, Thiéry A, et al. Calibrating machine learning approaches for probability estimation: A comprehensive comparison. *Statistics in Medicine*. 2023;42(29):5451-5478. doi: 10.1002/sim.9921