Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications

Jochen Kruppa[1], **Yufeng Liu**[2], **Hans-Christian Diener**[3], **Theresa Holste**[1], **Christian Weimar**[3], **Inke R. König**[1], and **Andreas Ziegler***[,1,4]

[1] Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Ratzeburger Allee 160, Haus 24, 23562 Lübeck, Germany
[2] Department of Statistics and Operations Research, Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, CB 3260, Chapel Hill, NC 27599, USA
[3] Klinik für Neurologie, Universitätsklinikum Essen, Hufelandstr. 55, 45147 Essen, Germany
[4] Zentrum für Klinische Studien Lübeck, Ratzeburger Allee 160, Haus 2, 23562 Lübeck, Germany

Received 23 April 2013; revised 29 September 2013; accepted 1 October 2013

Machine learning methods are applied to three different large datasets, all dealing with probability estimation problems for dichotomous or multicategory data. Specifically, we investigate *k*-nearest neighbors, bagged nearest neighbors, random forests for probability estimation trees, and support vector machines with the kernels of Bessel, linear, Laplacian, and radial basis type. Comparisons are made with logistic regression. The dataset from the German Stroke Study Collaboration with dichotomous and three-category outcome variables allows, in particular, for temporal and external validation. The other two datasets are freely available from the UCI learning repository and provide dichotomous outcome variables. One of them, the Cleveland Clinic Foundation Heart Disease dataset, uses data from one clinic for training and from three clinics for external validation, while the other, the thyroid disease dataset, allows for temporal validation by separating data into training and test data by date of recruitment into study. For dichotomous outcome variables, we use receiver operating characteristics, areas under the curve values with bootstrapped 95% confidence intervals, and Hosmer–Lemeshow-type figures as comparison criteria. For dichotomous and multicategory outcomes, we calculated bootstrap Brier scores with 95% confidence intervals and also compared them through bootstrapping. In a supplement, we provide R code for performing the analyses and for random forest analyses in `Random Jungle`, version 2.1.0. The learning machines show promising performance over all constructed models. They are simple to apply and serve as an alternative approach to logistic or multinomial logistic regression analysis.

Keywords: Brier score; German Stroke Study Collaboration; Probability estimation; Random forest; Random Jungle; Support vector machine.

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

For probability estimation of dichotomous and multicategory outcome variables, machine learning approaches serve as an alternative to logistic and multinomial logistic regression. The theoretical basis for some learning machines is described in a companion paper (Kruppa et al., 2014) with a focus on *k*-nearest neighbors (*k*-NN), bagged nearest neighbors (b-NN), random forests (RFs) using

*Corresponding author: e-mail: ziegler@imbs.uni-luebeck.de, Phone: +49-451-500-2780, Fax: +49-451-500-2999

probability estimation trees (RF-PETs), and support vector machines (SVMs). However, we have not shown how the approaches fare in real data applications. The aim of this work therefore is to apply the different learning machines to three different medical problems and compare them with logistic and multinomial logistic regression models for evaluating their performance.

The first application utilizes data from the German Stroke Study Collaboration (Weimar et al., 2002, 2004), which aimed at predicting functional independence or mortality 100 days after acute ischemic stroke. The training data were prospectively collected in 1998 and 1999 from seven centers; for the validation study, patients were enrolled during 2001 and 2002. Four hospitals participating in the training study also took part in the validation study, allowing for temporal validation. Nine hospitals ascertained patients for the validation study only and serve for temporal plus external validation.

The other two datasets are freely available from the University of California in Irvine (UCI) repository of machine learning database (http://archive.ics.uci.edu/ml/datasets.html). In the second application, the aim was to diagnose coronary artery disease (CAD) using the Cleveland Clinic Foundation Heart Disease data. Patients referred for coronary angiography at the Cleveland Clinic between May 1981 and September 1984 formed the training dataset, and data collected between 1983 and 1987 from three other centers were used for external validation (Detrano et al., 1984, 1989).

In the third application, the aim was to predict whether a subject was hypothyroid. Here, we used the so-called thyroid disease data, originally obtained from the Garvan Institute of Medical Research, St. Vincent's Hospital, Sydney. The dataset includes subjects measured at the Garvan clinical laboratory for endocrine analysis between 1984 and 1987. Data were divided into a training and a validation dataset by date of enrollment, allowing for temporal validation.

For the first application, data were ascertained prospectively, and the test data used for validation were collected after the initial study was completed. In the other two applications, data were cross-sectional. All three applications thus follow study designs allowing for the calculation of probabilities. In the first application, we used two dichotomous and one three-category outcome, while in applications 2 and 3 the outcome was dichotomous.

In Sections 2–4, we describe the datasets used for application in more detail. Results from logistic and multinomial logistic regression models are provided in Supporting Information S1. The different machine learning approaches, that is, b-NN, *k*-NN, RF, and SVM, and the utilized implementations are briefly described in Section 5. Statistical approaches for comparing the performance of different machine learning approaches are introduced in Section 6. In Sections 7–9, we present the results of the data analyses for the three applications. In Supporting Information S2 and S3, we provide R code for performing the analyses for the thyroid disease data (application dataset 3), and we also provide code chunks for running `Random Jungle` 2.1.0 on this dataset.

2 Application 1: Prognosis 100 days after stroke—German Stroke Study Collaboration

Disability and mortality represent the most relevant clinical outcomes after acute ischemic stroke. The aim of the clinical study from the German Stroke Study Collaboration was the development and validation of prognostic models for recovery, that is, functional independence and mortality (Weimar et al., 2002, 2004). Accurate comprehensive models do not only allow for prognosis of patients but also for correction of stratification of treatment groups, and prediction of the distribution of endpoint variables in a clinical trial. This in turn can increase the power to detect clinically relevant differences. The data have been used for methodological investigations before (König et al., 2007, 2008; Malley et al., 2011).

To predict functional independence, the Barthel index (BI) (Mahoney and Barthel, 1965) determined 100 days after the stroke event was used as outcome. This index evaluates individual abilities, such as mobility or personal hygiene that are immediately important to the patient. The BI can take values

Table 1 Final logistic and multinomial logistic regression models in the training data of the German Stroke Study Collaboration. Model I: complete restitution versus incomplete restitution or mortality. Model II: mortality versus survival. Model III: three-category (trinomial) logistic regression model using the most frequent category complete restitution as reference category.

Variable	Model I		Model II		Model III			
	$\hat{\beta}$	$SE_{\hat{\beta}}$	$\hat{\beta}$	$SE_{\hat{\beta}}$	Incomp rest		Mortality	
					$\hat{\beta}$	$SE_{\hat{\beta}}$	$\hat{\beta}$	$SE_{\hat{\beta}}$
Intercept	−8.378	0.510	−9.370	0.783	−7.920	0.508	−15.240	1.025
Neurological complications	1.289	0.332			1.237	0.336	1.649	0.465
Fever > 38°C	1.078	0.238	1.317	0.209	0.885	0.242	1.865	0.292
Lenticulostriate arteries infarction	0.743	0.216			0.773	0.215	0.202	0.400
Diabetes mellitus	0.655	0.152			0.632	0.152	0.815	0.239
Rankin scale[a)]	0.537	0.070			0.544	0.071	0.621	0.146
Prior stroke	0.517	0.161			0.492	0.163	0.708	0.257
Left arm paresis[a)]	0.488	0.103			0.388	0.089	0.438	0.125
Female gender	0.394	0.138			0.420	0.139	−0.089	0.230
Right arm paresis[a)]	0.393	0.089			0.494	0.103	0.368	0.147
NIH-SS total score at admission[a)]	0.074[b)]	0.024	0.133	0.013	0.059	0.024	0.154	0.033
Age (difference of one year)	0.066[c)]	0.006	0.076	0.010	0.059	0.006	0.150	0.012

$\hat{\beta}$, parameter estimate; $SE_{\hat{\beta}}$, standard error.
a) Difference of 1 scale score.
b) NIH-SS overall score is modeled as $[(\text{score}+1)/10]^{0.5} - 0.89$.
c) Age is modeled as $[(\text{age}/10)^3 - 315.60]/100$.

between 0 (total functional dependence) and 100 (total functional independence) in steps of five points to identify patients with complete restitution.

We consider the development of three models:

Model I: Complete restitution (BI $\geq$ 95) versus incomplete restitution (BI $<$ 95) or mortality,
Model II: mortality versus survival, and
Model III: incomplete restitution versus complete restitution versus mortality.

A systematic literature review was conducted in 1998 prior to data assessment to identify independent prognostic factors for outcome after ischemic stroke (`www.unidue.de/neurologie/stroke/free/lit_eng1.html`), and 49 variables were selected, which are assessable within the first 72 h after admission. For the training dataset (Weimar et al., 2002), complete data were available in 1737 (99.0%) of the patients. For the validation study (Weimar et al., 2004), complete case records were available for 1447 (98.4%) patients.

The development of the logistic regression models and the variable selection approach is briefly described in Supporting Information S1.1. Regression coefficients for the logistic regression models and the multinomial logistic regression model are provided in Table 1.

3 Application 2: Diagnosis of CAD—Cleveland Clinic data

CAD and its main complication, myocardial infarction, are leading causes of death and disability worldwide. Treatment of CAD depends on the specific symptoms and the severity of disease. The spectrum of treatment options is very wide and ranges from changes in life style, over lipid lowering therapies to surgical interventions. The reference standard for diagnosing CAD is coronary angiography, which is expensive and involves a small risk of complications and death (Genders et al., 2011). Therefore, noninvasive testing is recommended to select patients who will benefit from coronary angiography. The aim of the Cleveland Clinic data project was to determine whether an automate for probability estimation derived from the clinical and test characteristics of a relatively small group of 303 patients could accurately predict CAD. The training data consisted in 303 consecutive patients referred for coronary angiography to the Cleveland Clinic between May 1981 and September 1984. Patients underwent exercise electrocardiogram, thallium scintigraphy, and cardiac fluoroscopy, which are all noninvasive. Further details of these data collection can be found elsewhere (Detrano et al., 1984, 1989). In accordance with Detrano et al. (1989), data from three other centers with a total of 617 patients were used for external validation. These were (i) 200 patients drawn from all consecutive subjects undergoing cardiac catheterization between 1984 and 1987 at the Veterans Administration Medical Center in Long Beach, California, USA, (ii) 294 patients undergoing catheterization drawn from the Hungarian Institute of Cardiology (Budapest) between 1983 and 1987, and (iii) 123 patients drawn from subjects undergoing cardiac catheterization at the University Hospitals Zurich and Basel, Switzerland, in 1985. In these groups, noninvasive test results were not withheld from the treating physician and might have influenced the decision to perform coronary angiography. Detailed descriptive statistics of these patients can be found elsewhere (Detrano et al., 1989).

The dependent variable was the diagnosis of CAD as determined by the presence or absence of a $>50\%$ diameter in an angiography (Detrano et al., 1989), and independent variables are described in Supporting Information S1.2. Missing values of categorical variables were imputed by the most frequent category, and mean imputation for continuous independent variables.

The development of the logistic regression models and the variable selection approach is described in Supporting Information S1.2. Regression coefficients for the logistic regression model are provided in Table 2. All variables included in the final model were termed "clinically relevant" by Detrano et al. (1989).

4 Application 3: Diagnosis of hypothyroid conditions—Thyroid data

The thyroid produces several hormones, including triiodothyronine (T3) and thyroxine (T4). These hormones help oxygen entering the cells, and make the thyroid the master gland of metabolism (Shomon, 2013). The thyroid is part of a huge feedback process. The hypothalamus in the brain

Table 2 Final logistic regression developed using the training data of the Cleveland Clinic data.

Variable	$\hat{\beta}$	$SE_{\hat{\beta}}$
Intercept	−0.303	1.346
Sex	1.556	0.340
Chest pain type	0.871	0.169
Maximum heart rate	−0.030	0.008
ST depression induced by exercise relative to rest	0.699	0.152

$\hat{\beta}$, parameter estimate; $SE_{\hat{\beta}}$, standard error.

releases thyrotropin-releasing hormone (TRH), and the release of TRH effects the pituitary gland to release thyroid-stimulating hormone (TSH). Symptoms in case of a hypothyroid condition or even hypothyroidism range from sleepiness over depression to anemia. Hypothyroidism can be treated with the levorotatory forms of T3 and T4. The aim of the initial study was the diagnosis of thyroid conditions.

All 9172 subjects were referred to the Garvan clinical laboratory for endocrine analysis between 1984 and 1987. They had employed an expert system to generate diagnostic comments on each case in addition to approval by a qualified clinical endocrinologist (Quinlan et al., 1987). Subjects were classified as being hypothyroid (categories E, F, G, H in the last variable of the dataset) or nonhypothyroid (categories I, J, R, S, T), leaving 8021 subjects. Age and the following data measured at the laboratory are continuous: TSH, T3, total T4 (TT4), thyroxine uptake (T4U), free thyroxine index (FTI), and thyroxine-binding globuline (TBG). The referring center is the only categorical variable with values WEST, STMW, SVHC, SVI, SVHD, and "other." Sex (male/female), sick, pregnant, lithium, goiter, tumor, hypopituitary, psych, TSH measured, T3 measured, TT4 measured, T4U measured, FTI measured, and TBG measured (all yes/no) are dichotomous. Missing categorical data were imputed by the most frequent category, and missing continuous data were imputed by mean imputation of a single variable. We removed the variables T3 and TBG because they had more than 20% missing values in the training data.

Five major and several minor centers enrolled subjects into this study. Data collected before January 1, 1986 formed the training dataset, and all other data the test dataset. The name of the dataset on the UCI server is thyroid0387.

The development of the logistic regression models and the variable selection approach is described in Supporting Information S1.3. Regression coefficients for the logistic regression model are provided in Table 3. To the best of our knowledge, we are not aware of any other published logistic regression coefficients from the use of the Thyroid data. The signs of the regression coefficients are in the expected direction.

5 Machine learning methods and their implementations

The machine learning approaches employed are described in detail in the theoretical companion paper (Kruppa et al., 2014). In brief, we used logistic regression and multinomial logistic regression as described in the Supporting Information.

5.1 *k*-NN and b-NN

Probabilities were estimated for k-NN by first fixing k, and second estimating the proportion of affected subjects among the k-NN.

Table 3 Final logistic regression model in the training data of the Thyroid data.

Variable	$\hat{\beta}$	$SE_{\hat{\beta}}$
Intercept	−6.342	0.308
Age	−0.010	0.003
Thyrotropine measured (yes = 1)	−7.510	0.399
Thyrotropine	0.488	0.023
Total thyroxine measured (yes = 1)	1.389	0.240
Total thyroxine	0.032	0.002

$\hat{\beta}$, parameter estimate; $SE_{\hat{\beta}}$, standard error.

For b-NN, bootstrap samples are drawn with replacement and of the same size as the original dataset. For each bootstrap sample, the proportion of "1"s is determined for the k-NN, and the probability estimate is the average of these proportions over all bootstrap samples. The optimal number of nearest neighbors can be tuned using the training data for each model. It is the parameter k yielding the lowest error, in our case the lowest Brier score (BS); for details on the BS, see Section 6. For nearest neighbors, we varied k, calculated the BS, and picked the k with minimal BS for further analysis. k-NN and b-NN are expected to yield similar results as long as bootstrapping is done with replacement and with the same sample size as the original dataset (Domeniconi and Yan, 2004; Hall and Samworth, 2005; Samworth, 2012; Kruppa et al., 2014).

For k-NN and b-NN, we employed the `caret` package version 5.15044 in R (http://www.r-project.org/) using the functions `knn3()` for k-NN and `bag()` for b-NN. As working environment, we used R version 2.15.3 and connected packages. Supporting Information S2.3 presents a generic function in pseudo R code for the optimization of k for the nearest neighbor approaches.

The tuning revealed $k = 87$ for the German Stroke Study Collaboration data for all three models, $k = 5$ for the Cleveland Clinic data, and $k = 5$ for the Thyroid data. For b-NN, 200 bootstrap draws were drawn. All analyses were done using default options. In the case of b-NN, we took the mean as aggregate function for all predictions. No further adjustments were done to the training or the test data.

5.2 Random forests

RF is similar to b-NN in the way that bootstrapping is essential for this learning machine. However, a regression tree is built in each bootstrap step (Malley et al., 2012). Another difference to b-NN is that not all independent variables, also termed features, are used for building a tree, but a randomly selected proportion of features (Kruppa et al., 2013). The corresponding parameter is termed `mtry` in R and other RF packages. In applications, this parameter is often tuned in steps of 10% of the number of available features (Schwarz et al., 2010). The default is, however, the use of either the logarithm or the square root of available features. For estimating individual probabilities, the proportion of "1"s is determined in each terminal node. To determine the individual probability estimate over all trees, a subject is dropped down all trees, and the proportion of "1"s is averaged over the trees. A second tuning parameter is the size of the trees (Kruppa et al., 2013). A final important parameter is the number of bootstrap samples to be drawn, that is, the number of trees to be grown.

The R `randomForest` package and our stand-alone implementation `Random Jungle` have fewer adjustable parameters than k-NN and b-NN. However, RF may be easily combined with a back step procedure to select important features. This might be helpful if the number of independent variables is large and if independent variables are available with no effect on the dependent variable (Díaz-Uriarte and Alvarez de Andrés, 2006; Kruppa et al., 2012). Example code for the `randomForest` package version 4.6-7 and its use on the Thyroid data is provided in Supporting Information S3. However, for the analyses of the three application datasets we chose our own implementation `Random Jungle` 2.1.0 (http://www.randomjungle.de) because of its computational speed and the possibility of parallel processing (Schwarz et al., 2010). Furthermore, the multicategory approach to probability estimation is only available in `Random Jungle`, but not in R.

To tune the terminal node size for RF, we varied the terminal node size, calculated the BS for every terminal node size, and chose the terminal node size with the minimal BS (see Supporting Information S2.3 for code pieces). The tuning yielded a minimum terminal node size of 15, and the optimal number of features per tree to split on (`mtry`) was 3 for all German Stroke Study Collaboration models. The optimization of the Cleveland Clinic data yielded a minimum terminal node size of 13 and `mtry` = 3. Finally, for the Thyroid data, we observed a minimum terminal node size of 25, and we trained the model with `mtry` = 3. The number of trees grown in the RF was set to 10,000.

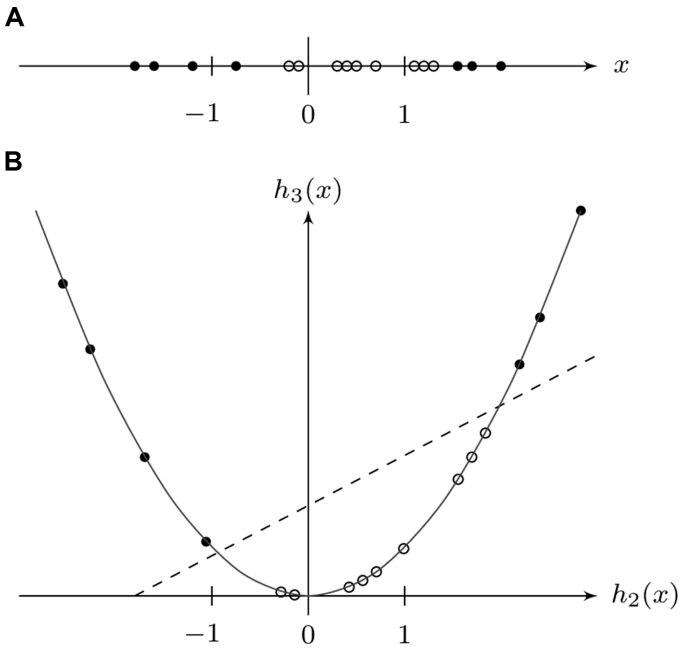

Figure 1 Example of (a) a one-dimensional space of independent variables, and (b) a higher dimensional feature space. No single cut separates closed and open bullets in the linear space (a), but the feature space (b) allows separation by a hyperplane. In (b), the kernel function is an inhomogeneous quadratic kernel, and the univariate independent variable x from (a) is transformed via this kernel to the two-dimensional feature space. Perfect separation is possible with the displayed data in this feature space. The functions from x in (a) to the x-axis and y-axis in (b) are $h_2(x) = \sqrt{2x}$ and $h_3(x) = x^2$, respectively. This figure is similar to Fig. 8 in König et al. (2008).

The importance of independent variables was determined using conditional variable importance (for detailed discussion on importance measures, see Nicodemus et al. (2010)). Unexpectedly, the backward elimination approach of Díaz-Uriarte and Alvarez de Andrés (2006) did not improve RF estimates for any datasets, and we therefore chose the complete set of independent variables for model building.

5.3 Support vector machines

While the use of b-NN, k-NN, and RF for probability estimation is straightforward, SVMs primarily are classifiers. In SVMs, hyperplanes are constructed in a multidimensional space to separate affected from unaffected subjects or, more formally, to separate subjects of different class labels. A formal discussion of SVMs can be found, for example, in the textbook by Schölkopf and Smola (2002) or the review papers by Moguerza and Muñoz (2006) or König et al. (2008). Intuitively, a good separation is achieved by the hyperplane that separates affected from unaffected subjects by maximizing the margin between the two groups. Since a perfect separation cannot be expected, this maximization of the margin is done subject to some error in the separation. Because subjects may not be separable in the original space by a hyperplane (Fig. 1), the original space is often mapped into a much higher dimensional space, usually termed feature space. To keep the computational effort low, only a small class of kernel functions is used for these transformations, and their most important property is that they can be

calculated by inner products in the original space. The list of kernel functions available in R packages can be found in Karatzoglou et al. (2006). Although kernel functions reduce the computational burden substantially because they can be computed through functions of an inner product, they affect the computational speed of the SVM as illustrated by Kruppa et al. (2014).

The standard SVM approach allows for classification. However, it has been shown that the solution of the standard SVM classifier is a consistent estimator for a specific probability (Lin, 2002). Wang et al. (2008) proposed to estimate probabilities over the entire range from 0 to 1 by repeatedly solving classification tasks. This requires the use of so-called weighted SVMs and allows a bracketed estimation of individual probabilities.

The SVMs were run using the `kernlab` package version 0.9-14. We used the function `ksvm()` in classification mode ("C-svc") with a cost of 0.01 constraints violation and the option `prob.model` to receive the class probability predictions. We applied different kernels: `rdfdot` as radial kernel, `vanilladot` as linear kernel, `laplacedot` as Laplacian kernel, and finally `besseldot` as Bessel kernel. In principle, one could try to improve the performance by tuning the cost parameter through cross-validation, but the computational effort is huge.

All machine learning approaches can be extended to the multicategory setting, as described in detail by Kruppa et al. (2014). For multicategory applications, the standard R code for b-NN, k-NN, and the SVM approaches does not need to be adjusted. Our only recommendations are to use numeric values for the categorical dependent variable and positive values for the codes. In addition to the detailed code for the analysis of the Thyroid data in Supporting Information S3, we provide extra code pieces in Supporting Information S2. Finally, we note that we replaced missing data by mean imputation or imputing the most frequent category in the first step. We did no further adjustments to the training or the test data.

6 Evaluating the performance of learning machines

6.1 Performance measures for a single learning machine

The standard approach for displaying results in classification approaches is the confusion matrix of predicted versus observed class membership. For dichotomous outcome variables, traditional performance measures include sensitivity (sens) and specificity (spec), that is, the proportion of correctly classified affected and unaffected subjects, respectively, and the percentage of correctly classified subjects (percentage correct; PC). All three measures are probabilities, and confidence intervals can be calculated as discussed by Newcombe (1998c).

If a threshold model is used for classification, such as the logistic regression model or SVMs, a quantitative score is available, such as a linear predictor. This allows estimation of the receiver operating characteristic (ROC) curve, the area under the curve (AUC), equivalent to the c-statistic, and confidence intervals for the AUC (Reiser and Guttman, 1986; Cortes and Mohri, 2005).

In the multiclass case, classification performance of a single machine is often measured using PC, but there are alternatives including the Heidke skill score, the Hanssen–Kuipers score (Wilks, 2006), or an entropy-based measure allowing to weight for a priori probabilities (Sindhwani et al., 2001).

One standard measure for evaluating the performance of a probability estimator is the BS (Brier, 1950), and the BS is sometimes termed probability score (PS). An alternative measure is the Brier skill score, BSS (Bradley et al., 2008).

BS and BSS are given by

$$BS = \mathbb{E}(y_i - \boldsymbol{P}(y_i = 1|\boldsymbol{x}_i))^2 \quad \text{and} \quad BSS = \frac{\mathbb{E}(y_i - \boldsymbol{P}(y_i = 1|\boldsymbol{x}_i))^2}{BS_{ref}},$$

where BS_{ref} is a reference BS. The reference BS may be the BS without the use of covariates, for example, $BS_{ref} = \mathbb{E}(y_i - I\!P(y_i = 1))^2$, and in this case the BSS is directly related to the coefficient of determination. BS and BSS can be consistently estimated by

$$\widehat{BS} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{I\!P}(y_i = 1|\boldsymbol{x}_i)\right)^2 \quad \text{and} \quad \widehat{BSS} = \frac{\sum_{i=1}^{n}(y_i - \widehat{I\!P}(y_i = 1|\boldsymbol{x}_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

The BS is a strictly proper score (Gneiting and Raftery, 2007), which means that it takes its minimal value only when the true probabilities are inserted as predictions. Phrased differently, it is not possible to improve the BS by systematically predicting probability values other than the best estimate. The BS is the mean square probability error, thus measures the same characteristics of a probability as the mean squared error measures for a continuous forecast (Stanski et al., 1989), and it can be estimated consistently if $I\!P(y_i = 1|\boldsymbol{x}_i)$ can be estimated consistently (Malley et al., 2012).

The sampling variances of BS and BSS have only been studied recently (Bradley et al., 2008). The BS is also often used in the multicategory case (Stanski et al., 1989), and their estimates are given by

$$\widehat{BS}_J = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{J}(y_{ij} - \widehat{I\!P}(y_i = j|\boldsymbol{x}_i))^2$$

and

$$\widehat{BSS}_J = \frac{\sum_{i=1}^{n}\sum_{j=1}^{J}(y_i j - \widehat{I\!P}(y_i = j|\boldsymbol{x}_i))^2}{\sum_{i=1}^{n}\sum_{j=1}^{J}(y_{ij} - \bar{y}_j)} \tag{1}$$

with $y_{ij} = 1$ if $y_i = j$, and 0 otherwise. The division by 2 in the first term of Eq. (1) is sometimes done for normalization purposes.

The extension of the BS to the ordered multicategory case, which we do not consider in this article, is the ranked PS, and it compares the cumulative density function (CDF) of a probabilistic forecast with the CDF of the corresponding observation over a given number of discrete probability categories (Wilks, 2006; Weigel et al., 2007).

Alternative measures for the dichotomous case include the mean absolute error (MAE) or the root mean square error (RMSE), which are estimated by

$$\widehat{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{I\!P}(y_i = 1|\boldsymbol{x}_i)| \quad \text{and} \quad \widehat{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{I\!P}(y_i = 1|\boldsymbol{x}_i))^2},$$

respectively.

As a graphical display, Hosmer and Lemeshow (H&L) type figures for dichotomous outcome can be created as described by Gillmann and Minder (2009).

6.2 Comparing two learning machines

If two learning machines are applied to the same dataset, classification accuracy for a dichotomous-dependent variable can be compared using estimates for differences of two correlated binomial distributions (Newcombe, 1998a; Tango, 2000; Zou and Donner, 2004; Zhou and Qin, 2005, 2007; Wenzel and Zapf, 2013).

For the German Stroke Study Collaboration data, temporal and external validation can be considered separately. If the aim is to compare the classification performance of a learning machine applied to temporal validation data with that of the same or a different learning machine applied to external

validation data, confidence intervals can be constructed either for differences (Newcombe, 1998b; Zhou et al., 2004; Wenzel and Zapf, 2013) or for ratios (Koopman, 1984) of two independent binomial distributions (König et al., 2008).

In the multicategory case, the construction of confidence intervals to compare classification performance is, however, substantially more complicated, unless only PC is considered.

For dichotomous outcome variables, a measure to compare the performance of probability estimates is the difference of two BS. Redelmeier et al. (1991) derived the variance of two correlated BS, say BS_1 and BS_2, and this variance is given by

$$\mathbb{V}\mathrm{ar}(\widehat{BS}_1 - \widehat{BS}_2) = \frac{4}{n^2}\sum_{i=1}^{n}((\widehat{\boldsymbol{P}}_1(y_i = 1|\boldsymbol{x}_i) - \widehat{\boldsymbol{P}}_2(y_i = 1|\boldsymbol{x}_i))^2\pi_i(1-\pi_i)).$$

This expression involves the a priori probability $\pi_i = \widehat{\boldsymbol{P}}_1(y_i = 1|\boldsymbol{x}_i)$, and the authors discussed the choice of π_i. In their application they chose π_i as the average $(\widehat{\boldsymbol{P}}_1(y_i = 1|\boldsymbol{x}_i) + \widehat{\boldsymbol{P}}_2(y_i = 1|\boldsymbol{x}_i))/2$. Redelmeier et al. (1991) recommend the Spiegelhalter (1986) test as a pretest before comparing two BS to determine if the estimated probabilities are compatible with the outcome.

An alternative approach to avoid the dependence on the a priori probability is to use resampling (Ferro, 2007), and we specifically used the bootstrap to obtain confidence intervals for the difference of two BS (Malley et al., 2012). Ikeda et al. (2001) proposed the following bootstrap test. First, using the out of bag bootstrap samples, the difference of the BS $\widehat{BS}_{1-2}(\cdot)$ of the two learning machines is estimated as is the standard error $\hat{\sigma}^B_{\widehat{BS}_{1-2}(\cdot)}$. The standardized test statistic

$$z_b = \frac{\widehat{BS}_{1-2}(\cdot)}{\hat{\sigma}^B_{\widehat{BS}_{1-2}(\cdot)}}$$

is asymptotically normally distributed. One advantage of this resampling approach is that it can directly be extended to the BSS in the dichotomous and the multicategory case as well as to the BS in the multicategory case.

We compared the different probability estimation methods with the ROC curves, the corresponding AUC values, the BS, and the H&L type figures. Because of the high number of observations, we chose 100 as bin size for the latter. The main criterion for judging H&L type plots is whether the values are on a line with slope 1. Furthermore, there should be no gaps.

In addition, we tested the difference of two BS using the Spiegelhalter (1986) test as a pretest to all binary models. As we obtained significant results for all comparisons, this approach was not advisable for the datasets. We therefore decided to use the bootstrap test for BS comparisons. The multiclass models do not allow ROC and H&L type figures. We thus report BS only. Finally, we compared the AUCs of two methods by the DeLong test (DeLong et al., 1988).

7 Results for application 1: Prognosis 100 days after stroke—German Stroke Study Collaboration

Figure 2 displays the ROC curves for the data from the German Stroke Study Collaboration of models I and II, respectively. For both models, logreg yielded the best results in both the temporal and the external validation data. k-NN and b-NN performed slightly worse in terms of the AUC (Table 4) and the BS (Table 5), but not significantly worse than logreg (lowest $p = 0.399$, Supporting Information Tables S1 and S2). RF did not perform well on model I, and the Bessel kernel SVM did not perform well for models I and II. However, the other SVM kernels had reasonable AUCs for both models, and the linear kernel unexpectedly performed best among the SVMs with AUCs very similar to the AUCs

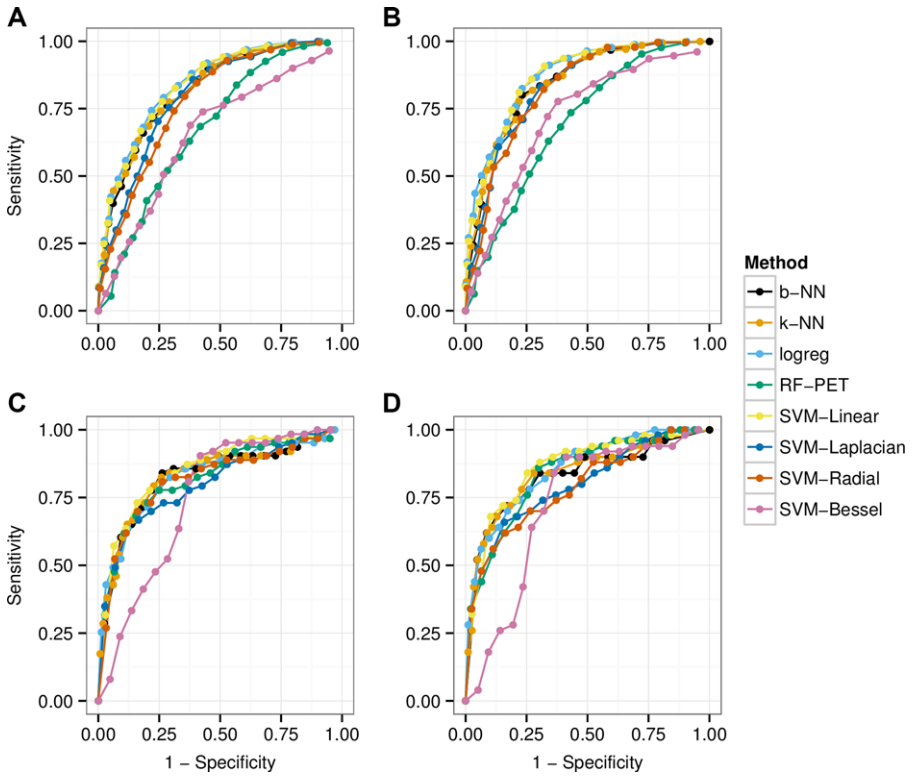

Figure 2 Receiver operating characteristic (ROC) curves for models I (complete restitution vs. incomplete restitution or mortality) and II (survival vs. mortality) of the German Stroke Study Collaboration data. (a) Temporal validation data, model I; (b) external validation data, model I; (c) temporal validation data, model II; (d) external validation data, model II.

of the logreg models. Table 5 also provides bootstrapped BS for model III, and the results are similar to those for models I and II for both the temporal and the external validation data.

Supporting Information Table S4 shows the ten most important variables for models I–III as obtained by RF. We stress that the numerical values of the conditional variable importance measure (VIM) are not comparable between the different models. However, the order of the variables can be compared, and it can be seen that the order of the most important variables varies substantially over the three models. For example, prior peripheral arterial disease is important for models I and III, which considers the probability of complete restitution. However, prior peripheral arterial disease seems to be less important for mortality.

Supporting Information Figs. S1 and S2 provide H&L type figures of the temporal and external validation, respectively, for model I. All learning machines but RF showed a good behavior, and values were reasonably spread. RF, however, had some gaps in the lower and higher predicted probabilities.

For model II, the corresponding displays are provided in Supporting Information Figs. S3 and S4. For these mortality data, all learning machines have some gaps, especially for the higher predicted probabilities.

The primary analysis of the data was a tuned logreg (Weimar et al., 2002, 2004). It performed very well, and no other machine outperformed logreg on this dataset. The Bessel kernel SVM, which has the greatest flexibility among the used kernels showed bad performance on all three models. Similarly,

Table 4 Area under the curve (AUC) values and 95% bootstrap confidence intervals (in brackets) after 2000 bootstrap draws for models I and II of the German Stroke Study Collaboration test data. Model I: complete restitution versus incomplete restitution or mortality; model II: mortality versus survival.

	Temporal validation		External validation	
	Model I	Model II	Model I	Model II
b-NN	0.845	0.832	0.822	0.832
	[0.800; 0.885]	[0.740; 0.92]	[0.787; 0.857]	[0.740; 0.910]
k-NN	0.846	0.837	0.823	0.829
	[0.803; 0.887]	[0.744; 0.922]	[0.787; 0.858]	[0.735; 0.912]
logreg	0.872	0.853	0.844	0.832
	[0.834; 0.91]	[0.812; 0.877]	[0.811; 0.875]	[0.755; 0.912]
RF-PET	0.699	0.837	0.677	0.822
	[0.642; 0.756]	[0.744; 0.910]	[0.630; 0.727]	[0.740; 0.900]
SVM-linear	0.871	0.860	0.840	0.854
	[0.834; 0.905]	[0.785; 0.934]	[0.804; 0.872]	[0.784; 0.928]
SVM-Laplacian	0.829	0.800	0.797	0.802
	[0.786; 0.874]	[0.702; 0.891]	[0.757; 0.835]	[0.713; 0.890]
SVM-radial	0.821	0.789	0.780	0.818
	[0.774; 0.864]	[0.689; 0.884]	[0.738; 0.822]	[0.721; 0.899]
SVM-Bessel	0.716	0.721	0.656	0.743
	[0.663; 0.776]	[0.639; 0.804]	[0.606; 0.705]	[0.674; 0.806]

b-NN, bagged nearest neighbors; *k*-NN, *k*-nearest neighbors; logreg, final logistic regression model; RF-PET, random forest using probability estimation trees; SVM, support vector machine (with corresponding kernel).

the RF approach did not perform well on models I and III. As a result, we have tried to improve the RF by adding variable selection because this had improved the performance of the logreg. Among others, we tried the backstepping approach proposed by Díaz-Uriarte and Alvarez de Andrés (2006), but it did not result in improvements (results not shown). All other approaches performed very similarly and reasonably well, and their results were more or less exchangeably.

8 Results for application 2: Diagnosis of CAD—Cleveland Clinic data

The ROC curves of the Cleveland Clinic data on CAD are provided in Fig. 3. While RF had performed poorly in application 1, it overall showed the best performance on these data. However, for low sensitivities logreg and SVM approaches performed slightly better so that there was no significant difference between these methods (Supporting Information Table S7). *k*-NN and b-NN were similar and performed poorly (Table 6). These conclusions are also supported by the H&L type figures (Supporting Information Fig. S5) and the AUC estimates (Supporting Information Tables S5 and S6). Furthermore, the BS of *k*-NN and b-NN were significantly larger than those of all other approaches (Supporting Information Table S7).

Supporting Information Table S8 shows the ten most important variables together with the VIM. The highest VIM was observed for the chest pain type, and this variable had a strong effect in the logistic regression too (Table 2). Substantial differences between variables included in the logistic regression model and the most important variables from RF were observed only for the variable age,

Table 5 Bootstrap Brier scores with 95% bootstrap confidence intervals for the temporal and external validation of the German Stroke Study Collaboration data. Model I: complete restitution (Barthel index (BI) $\geq$ 95) versus incomplete restitution (BI $<$ 95) or mortality; model II: mortality versus survival; model III: three-category (trinomial) logistic regression model for incomplete restitution versus complete restitution versus mortality ($B = 2000$).

	Temporal validation			External validation		
	Model I	Model II	Model III	Model I	Model II	Model III
b-NN	0.170	0.060	0.424	0.178	0.054	0.432
	[0.145; 0.196]	[0.042; 0.079]	[0.374; 0.477]	[0.158; 0.198]	[0.040; 0.069]	[0.388; 0.472]
k-NN	0.170	0.060	0.424	0.178	0.054	0.431
	[0.146; 0.195]	[0.042; 0.079]	[0.372; 0.478]	[0.157; 0.198]	[0.041; 0.069]	[0.388; 0.472]
logreg	0.153	0.058	0.395	0.168	0.052	0.419
	[0.126; 0.180]	[0.040; 0.077]	[0.341; 0.451]	[0.146; 0.192]	[0.037; 0.068]	[0.373; 0.466]
RF-PET	0.223	0.071	0.491	0.227	0.064	0.499
	[0.209; 0.236]	[0.054; 0.089]	[0.454; 0.529]	[0.216; 0.239]	[0.051; 0.078]	[0.469; 0.530]
SVM-linear	0.149	0.061	0.393	0.166	0.053	0.415
	[0.124; 0.176]	[0.054; 0.082]	[0.338; 0.445]	[0.145; 0.189]	[0.039; 0.067]	[0.371; 0.460]
SVM-Laplacian	0.164	0.064	0.411	0.185	0.057	0.446
	[0.140; 0.191]	[0.042; 0.085]	[0.361; 0.466]	[0.163; 0.207]	[0.042; 0.074]	[0.398; 0.490]
SVM-radial	0.172	0.065	0.416	0.201	0.061	0.463
	[0.146; 0.198]	[0.043; 0.087]	[0.363; 0.468]	[0.178; 0.227]	[0.044; 0.079]	[0.418; 0.508]
SVM-Bessel	0.214	0.085	0.502	0.243	0.078	0.538
	[0.188; 0.240]	[0.068; 0.103]	[0.457; 0.549]	[0.221; 0.226]	[0.065; 0.092]	[0.494; 0.577]

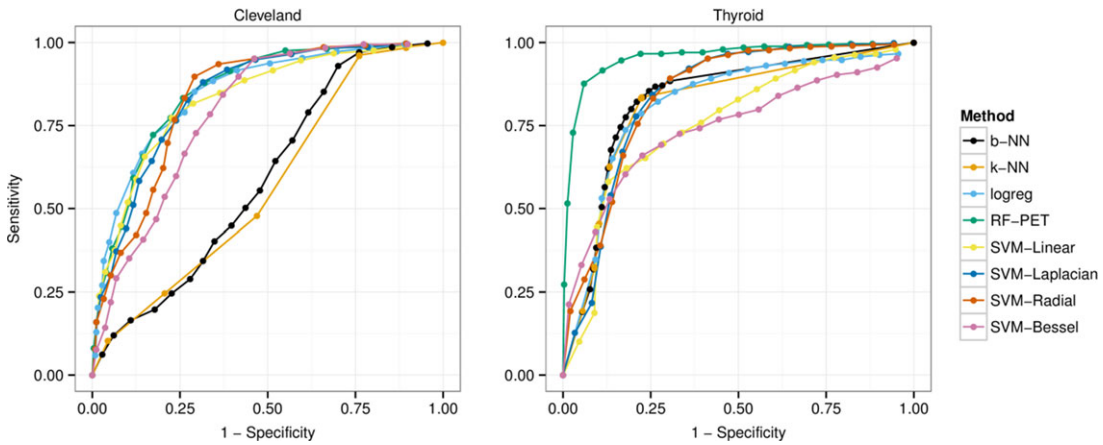

Figure 3 Receiver operating characteristic (ROC) curves for the test data of the Cleveland Clinic data and the Thyroid data.

which was not included in the logistic regression model but was the fourth most important variable in the RF analyses.

Although the diagnosis of CAD was far from perfect with the noninvasive tests, some of the machines showed good performance in estimating individual probabilities on the test data (Supporting Information Fig. S5). Only *k*-NN and b-NN failed completely on these data, and the Bessel kernel SVM as well as the radial basis kernel SVM showed an unwanted pattern in the H&L type figure

Table 6 Bootstrap Brier scores with 95% bootstrap confidence intervals for the test data of the Cleveland Clinic data and the Thyroid data ($B = 2000$).

	Cleveland Clinic	Thyroid
b-NN	0.252 [0.232; 0.271]	0.137 [0.125; 0.151]
k-NN	0.266 [0.244; 0.288]	0.141 [0.128; 0.154]
logreg	0.192 [0.166; 0.220]	0.150 [0.135; 0.164]
RF-PET	0.154 [0.135; 0.173]	0.058 [0.052; 0.064]
SVM-linear	0.166 [0.141; 0.193]	0.166 [0.151; 0.181]
SVM-Laplacian	0.155 [0.128; 0.182]	0.162 [0.148; 0.177]
SVM-radial	0.154 [0.127; 0.181]	0.133 [0.122; 0.146]
SVM-Bessel	0.174 [0.148; 0.200]	0.119 [0.108; 0.129]

Table 7 Area under the curve (AUC) values and 95% bootstrap confidence intervals (in brackets) after 2000 bootstrap draws for the Thyroid data.

Machine	AUC
b-NN	0.834 [0.813; 0.857]
k-NN	0.813 [0.790; 0.839]
logreg	0.809 [0.780; 0.833]
RF-PET	0.958 [0.944; 0.970]
SVM-linear	0.759 [0.731; 0.789]
SVM-Laplacian	0.841 [0.821; 0.860]
SVM-radial	0.844 [0.822; 0.862]
SVM-Bessel	0.744 [0.711; 0.778]

(Supporting Information Fig. S5). Nevertheless, there are several machines that could be chosen to estimate the individual probability for CAD using noninvasive tests, and these estimates, in turn, could be used for selecting patients who should undergo coronary angiography.

9 Results for application 3: Diagnosis of hypothyroid conditions—Thyroid data

The ROC curves of the Thyroid data are shown in Fig. 3, and the H&L type figure is provided in Fig. 4. The corresponding AUC values with bootstrapped 95% confidence intervals are summarized in Table 7; the corresponding BS values with 95% bootstrap confidence intervals are displayed in Table 6. When using the AUC as performance criterion, the linear kernel SVM and the Bessel kernel SVM did not perform well, and b-NN, *k*-NN, logreg, the Laplacian, and the radial kernel SVMs performed similarly. Most strikingly, RF outperformed all other machines on this dataset (Figs. 3 and 4, Tables 6 and 7, Supporting Information Tables S9 and S10). The AUC and the H&L type figure (Figs. 3 and 4) provide another unexpected finding. Although the Bessel kernel did not perform well in terms of the AUC (Fig. 3, Table 7), it had the second best BS (Fig. 4, Table 6), and it outperformed all machines (all nominal $p < 0.05$, see Supporting Information Table S9) but RF—to which it was inferior—and the radial kernel SVM. The discrepancy for the Bessel kernel SVM between the ROC curve on the one hand and the H&L type figures—both supported by AUC values and BS—might be explained by the low probability estimates for many subjects in the test data (Fig. 4).

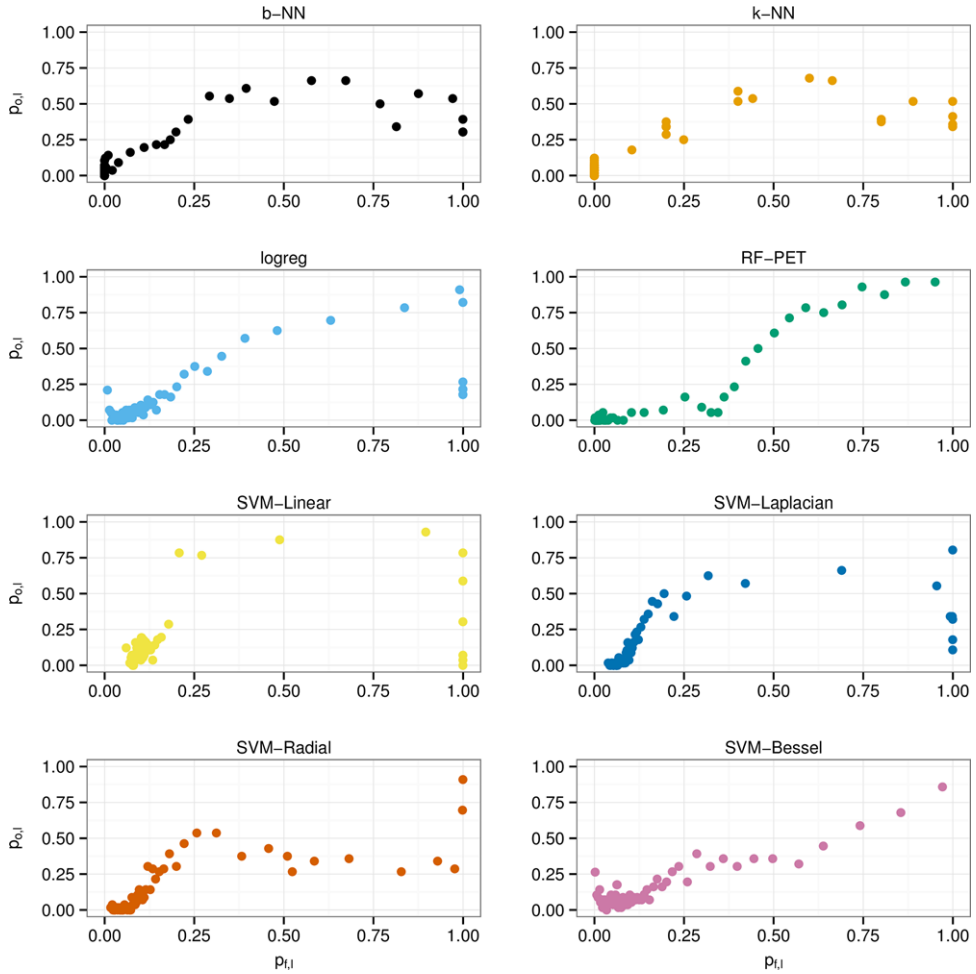

Figure 4 Hosmer–Lemeshow-type figures of the Thyroid data. The x-axis shows the estimated probabilities $p_{f,l}$ for each l bin and the y-axis the proportion $p_{o,l}$ of 1's in each l bin. Due to the many observations we decided to choose an l of 100. In ideal, we would expect a straight line without any gaps, indicating a good predictive power.

The highest VIM was observed for sex (Supporting Information Table S11), but this variable was not included in the final logreg model. Total thyroxine measured was the only variable among the ten most important variables from the RF analyses, which was also selected for the final logreg model. This might at least partly explain the difference between the performance of logreg and RF on the Thyroid data.

RF outperformed all other machines on this dataset, and RF is the only machine that might be used for diagnostic purposes on the Thyroid data. For example, when the sensitivity was 90%, the specificity was also approximately 90%. The H&L type figure also showed the best pattern for RF—which was confirmed by the BS estimates. All other machines cannot be recommended for probability estimation on the Thyroid data.

10 Discussion

No single machine performed best on all three datasets. Although this result has been expected from the simulation in the companion paper (Kruppa et al., 2014), several questions remain, and one important is whether all or at least one machine can be recommended for applications.

In all three real data examples and the simulated examples from the companion paper, there were negligible differences between *k*-NN and b-NN. This observation is in line with the literature (Domeniconi and Yan, 2004; Hall and Samworth, 2005; Samworth, 2012), although b-NN has a lower variance and greater stability in theory. This difference may depend on the specific bootstrap approach chosen. Here, we have followed the classical bootstrap method, where samples are drawn with replacement up to the original sample size. We cannot exclude that performance of the learning machines improves when a different resampling approach would be used. In general, given the similar performance of *k*-NN and b-NN, *k*-NN might be the better choice if the data are very large or if computational capacities are limited because b-NN is computationally intensive. Nevertheless, both approaches are limited for use in practice. Assume that a diagnostic test has been developed at one laboratory and that it consists in multiple biomarkers and a set of clinical variables. The multimarker rule has been developed using a nearest neighbor approach. Assume further that the test is to be carried out at a different laboratory. To determine the nearest neighbors of a new case with laboratory values generated at the new laboratory, the raw data need to be passed on from the initial laboratory to the new one.

Another aspect is interpretability of results. The fundamental question is whether a clinician will trust the findings obtained with a fancy noninterpretable machine and use this in clinical routine. Here, the logistic regression model has the clear advantage of yielding parameter estimates that are simple to interpret. RF has the advantage that VIMs can be estimated easily so that variables can be ranked by their importance. However, the magnitude of these estimates has no meaning. It would therefore be extremely helpful if approaches identifying representative trees are available. A first step in this direction is the work by Banerjee et al. (2012). Interpretability is generally difficult with SVMs, except for the linear kernel SVM and its linear extensions (Huang et al., 2012). The restriction in interpretability and the need to transfer data for nearest neighbor approaches could easily lead to the decision that only logreg and RF are used for an application.

However, the applications in this paper and the simulations from the theoretical paper (Bradley et al., 2008) show that results from logreg and RF may be conflicting and that the use of a few more machines is indicated. For example, it remains unclear why RF performed poorly on models I and III from the German Stroke Study Collaboration data but outperformed all other machines on the Thyroid data. This extremely varying performance of RF has been observed before, see, for example, Kruppa et al. (2013). The bad performance of logreg on the Thyroid data might, however, be explained by a suboptimal selection of variables. Specifically, only one variable in the final logreg model was among the ten most important variables identified by RF. Performance of machines may also depend on tuning of the machines, and the best choice of tuning parameters may be found by bootstrapping or other procedures. However, the associated computational cost can be substantial.

The need for manual manipulations before the final model is obtained might also lead to a selection of specific machines. In logreg, for example, the user needs to check the convergence of a model before the next step is taken in the analysis. Such difficulties do not occur with *k*-NN or b-NN. The nearest neighbors can always be determined automatically, and no checks are required by the investigator.

Missing data, such as missing follow-up data in the German Stroke Study Collaboration example can also be challenging. They can lead to substantial bias if the proportion of missing data in the dependent variable or the independent variables is too high. Users should therefore carefully investigate the dependency of probability estimates on the missing data structure, and they should take into consideration the use of computationally demanding multiple imputation approaches.

The ability to calibrate a machine to a different dataset might also be important, and predicted probabilities are on average adequate in a well-calibrated model (König et al., 2007). If baseline

probabilities differ, however, between training and test data, the intercept of a logreg might need to be estimated anew for good calibration of the model. Such a procedure seems to be impossible for other learning machines.

Since it is impossible to identify the best performing machine for a specific dataset a priori, it is advisable to use a set of different machines, because performance improvements over logreg can be obtained in some applications even when standard modeling approaches are followed (Harrell et al., 1996). We therefore recommend the use of other machines in addition to the biostatistical standard model, which is logreg. The machines can be compared easily using user-specified criteria, such as AUC, BS, or reclassification indices. However, the choice of the performance measure needs to be made carefully, as shown, for example, for BS and AUC in the Thyroid data example and, for example, for reclassification indices and AUC and BS by Hilden and Gerds (2013).

Since there is not a single best learning machine, future research should identify the conditions under which a learning machine can be expected to perform good or bad on a specific dataset. This would allow the selection of a well performing learning machine in advance.

Acknowledgments This paper is the summary from a workshop, entitled "Probability Estimation With Data Mining Methods" held at the Meeting of the Central European Network 2011 in Zurich, Switzerland. The authors are grateful to the German Region of the International Biometric Society for their generous funding of the Workshop. The authors are grateful to G. Biau, M. Kohler, and J. D. Malley for helpful discussions. AZ acknowledges funding from the German Science Foundation Excellence Cluster "Inflammation at Interfaces," and the European Union (BiomarCare, grant number: HEALTH-2011-278913), and the German Ministry of Education and Research (CARDomics, grant numbers: 01KU0908A and 01KU0908B; Phenomics, grant number: 0315536F). The German Stroke Study Collaboration was funded in part by the German Ministry of Education and Research (BMBF) as part of the Competence Net Stroke. YL acknowledges funding from the US National Institute of Health, grant NIH/NCI R01 CA-149569. We are grateful to the members of the German Stroke Study Collaboration for data collection. It includes the following neurology departments and responsible study investigators (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), University of Essen (G. Mörger-Kiefer, C. Weimar), University of Ulm (M. Riepe), University of Leipzig (D. S. Schneider), University of Jena (V. Willig). The authors thank K. Kraywinkel, M.D., M.Sc., and P. Dommes, Ph.D., for central data collection and management.

Conflict of interest

CW has received honorarium for presentations and membership on advisory boards from Bayer-Schering, Böhringer Ingelheim, Bristol-Myers Squib, and Sanofi-Aventis. He has also received travel funds from Teva and Biogen Idec, and research grants from Johnson & Johnson. HCD received honoraria for participation in clinical trials, contribution to advisory boards or oral presentations from Abbott, Allergan, AstraZeneca, Bayer Vital, BMS, Boehringer Ingelheim, CoAxia, Corimmun, Covidien, Daichii-Sankyo, D-Pharm, EV3, Fresenius, GlaxoSmithKline, Janssen Cilag, Johnson & Johnson, Knoll, MSD, Medtronic, MindFrame, Neurobiological Technologies, Novartis, Novo-Nordisk, Paion, Parke-Davis, Pfizer, Sanofi-Aventis, Schering-Plough, Servier, Solvay, Thrombogenics, Wyeth, and Yamanouchi. Financial support for research projects was provided by Astra/Zeneca, GSK, Boehringer Ingelheim, Lundbeck, Novartis, Janssen-Cilag, Sanofi-Aventis, Syngis, and Talecris. The Department of Neurology at the University Duisburg-Essen received research grants from the German Research Council (DFG), German Ministry of Education and Research (BMBF), European Union, NIH, Bertelsmann Foundation, and Heinz-Nixdorf Foundation. Within the past year, HCD served as editor of Aktuelle Neurologie, Arzneimitteltherapie, Kopfschmerznews, Stroke News, and the Treatment Guidelines of the German Neurological Society, as co-editor of Cephalalgia, and on the editorial board of Lancet

Neurology, European Neurology, and Cerebrovascular Disorders. AZ is statistical consultant for the Protagen AG, Dortmund. He is Associate Editor of the Biometrical Journal, Statistics in Medicine, and Methods of Information in Medicine. All other authors declare no conflict of interest.

References

Banerjee, M., Ding, Y. and Noone, A. M. (2012). Identifying representative trees from ensembles. *Statistics in Medicine* **31**, 1601–1616.

Bradley, A. A., Schwartz, S. S. and Hashino, T. (2008). Sampling uncertainty and confidence intervals for the Brier score and Brier skill score. *Weather and Forecasting* **23**, 992–1006.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1–3.

Cortes, C. and Mohri, M. (2005). Confidence intervals for the area under the ROC curve. In: Saul, L. K., Weiss, Y. and Lon, B. (Eds.), *Advances in Neural Information Processing Systems*, Vol. **17**. A Bradford Book, Cambridge, MA, pp. 305–312.

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S. and Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology* **64**, 304–310.

Detrano, R., Yiannikas, J., Salcedo, E. E., Rincon, G., Go, R. T., Williams, G. and Leatherman, J. (1984). Bayesian probability analysis: a prospective demonstration of its clinical utility in diagnosing coronary disease. *Circulation* **69**, 541–547.

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3.

Domeniconi, C. and Yan, B. (2004). Nearest neighbor ensemble. In: Kittler, J., Petrou, M., Nixon, M. S. and Hancock, E. R. (Eds.), *Proceedings of the 17th International Conference on Pattern Recognition, 2004*. IEEE Computer Society Press, Cambridge, UK, pp. 228–231.

Ferro, C. A. T. (2007). Comparing probabilistic forecasting systems with the Brier score. *Weather and Forecasting* **22**, 1076–1088.

Genders, T. S., Steyerberg, E. W., Alkadhi, H., Leschka, S., Desbiolles, L., Nieman, K., Galema, T. W., Meijboom, W. B., Mollet, N. R., de Feyter, P. J., Cademartiri, F., Maffei, E., Dewey, M., Zimmermann, E., Laule, M., Pugliese, F., Barbagallo, R., Sinitsyn, V., Bogaert, J., Goetschalckx, K., Schoepf, U. J., Rowe, G. W., Schuijf, J. D., Bax, J. J., de Graaf, F. R., Knuuti, J., Kajander, S., van Mieghem, C. A., Meijs, M. F., Cramer, M. J., Gopalan, D., Feuchtner, G., Friedrich, G., Krestin, G. P., Hunink, M. G. and Consortium, C. A. D. (2011). A clinical prediction rule for the diagnosis of coronary artery disease: validation, updating, and extension. *European Heart Journal* **32**, 1316–1330.

Gillmann, G. and Minder, C. E. (2009). On graphically checking goodness-of-fit of binary logistic regression models. *Methods of Information in Medicine* **48**, 306–310.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378.

Hall, P. and Samworth, R. J. (2005). Properties of bagged nearest neighbour classifiers. *Journal of Royal Statistical Society, Series B* **67**, 363–379.

Harrell, F. E., Jr., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361–387.

Hilden, J. and Gerds, T. A. (2013). A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Statistics in Medicine*, doi: 10.1002/sim.5804.

Huang, H., Liu, Y. and Marron, J. S. (2012). Bidirectional discrimination with application to data visualization. *Biometrika* **99**, 851–864.

Ikeda, M., Itoh, S., Ishigaki, T. and Yamauchi, K. (2001). Application of resampling techniques to the statistical analysis of the Brier score. *Methods of Information in Medicine* **40**, 259–264.

Karatzoglou, A., Meyer, D. and Hornik, K. (2006). Support vector machines in R. *Journal of Statistical Software* **15**. Available at www.jstatsoft.org/v15/i09/paper.

König, I. R., Malley, J. D., Pajevic, S., Weimar, C., Diener, H. C. and Ziegler, A. (2008). Patient-centered yes/no prognosis using learning machines. *International Journal of Data Mining and Bioinformatics* **2**, 289–341.

König, I. R., Malley, J. D., Weimar, C., Diener, H. C., Ziegler, A. and on behalf of the German Stroke Study Collaborators. (2007). Practical experiences on the necessity of external validation. *Statistics in Medicine* **26**, 5499–5511.

Koopman, P. A. R. (1984). Confidence intervals for the ratio of two binomial proportions. *Biometrics* **40**, 513–517.

Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I. R., Malley, J. D. and Ziegler, A. (2014). Probability estimation with machine learning methods for dichotomous and multi-category outcome: theory. *Biometrical Journal* **56**, 534–563.

Kruppa, J., Schwarz, A., Arminger, G. and Ziegler, A. (2013). Consumer credit risk: individual probability estimates using machine learning. *Expert Systems with Applications* **40**, 5125–5131.

Kruppa, J., Ziegler, A. and König, I. R. (2012). Risk estimation and risk prediction using machine learning methods. *Human Genetics* **131**, 1639–1654.

Lin, Y. (2002). Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery* **6**, 259–275.

Mahoney, F. I. and Barthel, D. W. (1965). Functional evaluation: the Barthel index. *Maryland State Medical Journal* **14**, 61–65.

Malley, D. J., Malley, K. G. and Pajevic, S. (2011). *Statistical Learning for Biomedical Data*. Cambridge University Press, Cambridge.

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G. and Ziegler, A. (2012). Probability machines: consistent probability estimation using nonparametric learning machines. *Methods of Information in Medicine* **51**, 74–81.

Moguerza, J. M. and Muñoz, A. (2006). Support vector machines with applications. *Statistical Science* **21**, 322–336.

Newcombe, R. G. (1998a). Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* **17**, 2635–2650.

Newcombe, R. G. (1998b). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* **17**, 873–890.

Newcombe, R. G. (1998c). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* **17**, 857–872.

Nicodemus, K. K., Malley, J. D., Strobl, C. and Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics* **11**, 110.

Quinlan, J. R., Compton, P. J., Horn, K. A. and Lazarus, L. (1987). Inductive knowledge acquisition: a case study. In: Quinlan, J. R. (Ed.), *Applications of Expert Systems*. Addison Wesley, London, UK, pp. 157–173.

Redelmeier, D. A., Bloch, D. A. and Hickam, D. H. (1991). Assessing predictive accuracy: how to compare Brier scores. *Journal of Clinical Epidemiology* **44**, 1141–1146.

Reiser, B. and Guttman, I. (1986). Statistical inference for $p(y < x)$: the normal case. *Technometrics* **28**, 253–257.

Samworth, R. J. (2012). Optimal weighted nearest neighbour classifiers. *Annals of Statistics* **40**, 2733–2763.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.

Schwarz, D. F., König, I. R. and Ziegler, A. (2010). On safari to Random Jungle: a fast implementation of Random Forests for high dimensional data. *Bioinformatics* **26**, 1752–1758.

Shomon, M. (2013) Thyroid disease. Basic information on hypothyroidism, hyperthyroidism, nodules, cancer. Available at `http://thyroid.about.com/cs/basics_starthere/a/thyroid101.htm`, accessed November 28, 2013.

Sindhwani, V., Bhattacharya, P. and Rakshit, S. (2001). Information theoretic feature crediting in multiclass support vector machines. In: Grossman, R. and Kumar, V. (Eds.), *Proceedings of the First SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, Chicago, IL, pp. 5–7.

Spiegelhalter, D. J. (1986). Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine* **5**, 421–433.

Stanski, H. R., Wilson, L. J. and Burrows, W. R. (1989). Survey of common verification methods in meteorology. World Weather Watch Technical Report 8, WMO/TD no. 358. World Meteorological Organization. Available at `http://www.cawcr.gov.au/projects/verification/Stanski_et_al/Stanski_et_al.html`, accessed April 1, 2013.

Tango, T. (2000). Confidence intervals for differences in correlated binary proportions. *Statistics in Medicine* **19**, 133–139.

Wang, J., Shen, X. and Liu, Y. (2008). Probability estimation for large-margin classifiers. *Biometrika* **95**, 149–167.
Weigel, A. P., Liniger, M. A. and Appenzeller, C. (2007). The discrete Brier and ranked probability skill scores. *Monthly Weather Review* **135**, 118–124.
Weimar, C., König, I. R., Kraywinkel, K., Ziegler, A., Diener, H. C. and on behalf of the German Stroke Study Collaboration. (2004). 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* **35**, 158–162.
Weimar, C., Ziegler, A., König, I. R., Diener, H. C. and on behalf of the German Stroke Study Collaborators. (2002). Predicting functional outcome and survival after acute ischemic stroke. *Journal of Neurology* **249**, 888–895.
Wenzel, D. and Zapf, A. (2013). Difference of two dependent sensitivities and specificities: comparison of various approaches. *Biometrical Journal* **55**, 705–718.
Wilks, D. S. (2006). *Statistical Methods in the Atmospherical Sciences* (2nd edn.). Academic Press, Burlington, MA.
Zhou, X. H. and Qin, G. S. (2005). A new confidence interval for the difference between two binomial proportions of paired data. *Journal of Statistical Planning and Inference* **128**, 527–542.
Zhou, X. H. and Qin, G. S. (2007) A supplement to: "a new confidence interval for the difference between two binomial proportions of paired data." *Journal of Statistical Planning and Inference* **137**, 357–358.
Zhou, X. H., Tsao, M. and Qin, G. S. (2004). New intervals for the difference between two independent binomial proportions. *Journal of Statistical Planning and Inference* **123**, 97–115.
Zou, G. and Donner, A. (2004). A simple alternative confidence interval for the difference between two proportions. *Controlled Clinical Trials* **25**, 3–12.

