LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Giuseppe Casalicchio, Bernd Bischl, Anne-Laure Boulesteix & Matthias Schmid

The Residual-based Predictiveness Curve - A Visual Tool to Assess the Performance of Prediction Models

Technical Report Number 178, 2015
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

The Residual-based Predictiveness Curve - A Visual Tool to Assess the Performance of Prediction Models

Giuseppe Casalicchio[*], Bernd Bischl[†], Anne-Laure Boulesteix[‡], Matthias Schmid[§]

January 13, 2015

Abstract

It is agreed among biostatisticians that prediction models for binary outcomes should satisfy two essential criteria: First, a prediction model should have a high discriminatory power, implying that it is able to clearly separate cases from controls. Second, the model should be well calibrated, meaning that the predicted risks should closely agree with the relative frequencies observed in the data. The focus of this work is on the predictiveness curve, which has been proposed by Huang *et al.* (Biometrics 63, 2007) as a graphical tool to assess the aforementioned criteria. By conducting a detailed analysis of its properties, we review the role of the predictiveness curve in the performance assessment of biomedical prediction models. In particular, we demonstrate that marker comparisons should not be based solely on the predictiveness curve, as it is not possible to consistently visualize the added predictive value of a new marker by comparing the predictiveness curves obtained from competing models. Based on our analysis, we propose the "residual-based predictiveness curve" (RBP curve), which addresses the aforementionened issue and which extends the original method to settings where the evaluation of a prediction model on independent test data is of particular interest. Similar to the predictiveness curve, the RBP curve reflects both the calibration and the discriminatory power of a prediction model. In addition, the curve can be conveniently used to conduct valid performance checks and marker comparisons.

Keywords: Calibration; Classification; Discrimination; Predictiveness curve; Risk prediction.

[*]Department of Statistics, University of Munich, Germany, email: giuseppe.casalicchio@stat.uni-muenchen.de
[†]Department of Statistics, TU Dortmund, Germany, email: bischl@statistik.tu-dortmund.de
[‡]Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Germany, email: boulesteix@ibe.med.uni-muenchen.de
[§]Department of Medical Biometry, Informatics and Epidemiology, University of Bonn, Germany, email: matthias.schmid@ukb.uni-bonn.de

1 Introduction

The development of prediction models for binary outcomes is an important issue in biomedical research (Moons et al., 2009). Regarding the development of statistical methodology, key issues are not only to develop methods for *deriving* new marker combinations, which usually combine statistical modelling techniques with feature selection, but also to develop reliable methods to *evaluate* and assess the practical value of new prediction models.

In the literature it is commonly agreed that prediction models for a binary outcome D should satisfy two major criteria: First, they should have a high *discriminatory* power, meaning that they are able to separate the categories of the binary outcome. For example, prediction models should be able to well separate cases ($D = 1$) from controls ($D = 0$) in case-control studies. Second, they should be *well calibrated*, meaning that the predicted risks should closely agree with the relative frequencies observed in the data. Discrimination and calibration can be evaluated by several measures, such as the area under the curve (AUC), the Brier score, and the Hosmer-Lemeshow statistic (Hilden & Gerds, 2014; Crowson et al., 2014).

The focus of this paper is on the predictiveness curve, which was first proposed by Huang, Pepe & Feng (2007) and has since been used by many authors as a graphical tool to visualize and evaluate calibration and discriminatory power (Pepe et al., 2008b; Gu & Pepe, 2009; Huang & Pepe, 2009b; Pepe, Gu & Morris, 2010; Moons et al., 2012; Steyerberg et al., 2014). The predictiveness curve depicts the risk distribution of a marker (or a marker combination) and is formally defined as follows: Let F be the cumulative distribution function (cdf) of the marker Y, and let $v = F(Y) \in (0, 1)$ be the v-th percentile of Y. Then the risk at a specific value v is given by

$$R(v) = risk(F^{-1}(v)) = \mathbb{P}(D = 1 \,|\, Y = F^{-1}(v)), \tag{1}$$

and the predictiveness curve is obtained by plotting $R(v)$ versus v. When $R(\cdot)$ is not known and estimation of the risk is required, for example, by using logistic regression, the predictiveness curve is obtained by plotting the predicted risk $\hat{R}(v) = \widehat{risk}(\hat{F}^{-1}(v))$ vs. $v = \hat{F}(\eta)$, which is equivalent to plotting $\widehat{risk}(\eta)$ versus $\hat{F}(\eta)$, where $\hat{F}$ is, for example, the empirical cdf and η is an observed value of the marker Y (see also Gu & Pepe, 2009). In other words, the curve plots the ordered predicted risks (based on the ordered observed values η of the marker Y) versus the cumulative percentage of individuals with marker values smaller than or equal to η.

As an example, Figure 1 shows the predictiveness curve for a simulated data set of size $n = 10000$. The estimated prevalence in this data set is $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} D_i = 0.3$, where D_i denotes the binary outcome of observation i. That is, the simulated data contain 30% cases and 70% controls. The values of the marker combination Y (simulated as the combination of five independent normally distributed markers) range from -24.5 to 22.3. The dashed gray curve represents the predictiveness curve that was obtained from fitting a logistic regression model to the data. It can be easily seen from Figure 1 that the predictiveness curve characterizes the distribution of the risk. Each horizontal line at $\hat{R}(v) = t$ intersects the curve at

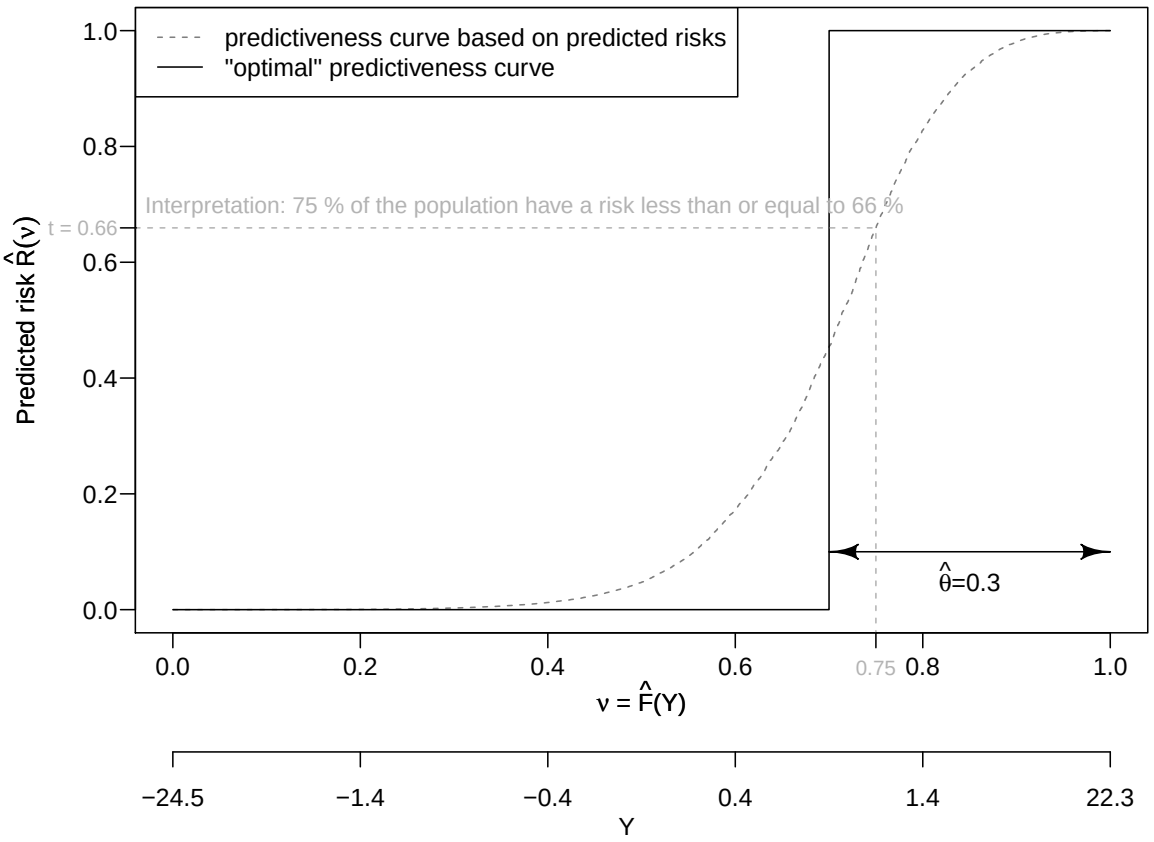

Figure 1: Predictiveness curve obtained from the simulated data set of Section 1 with estimated prevalence $\hat{\theta} = 0.3$. The dashed gray line represents the predictiveness curve of a logistic regression model while the solid black line corresponds to the optimal predictiveness curve that was obtained by using the binary outcome values.

a specific abscissa v, where $v = \hat{R}^{-1}(t)$ is interpreted as the proportion of the population that has a risk of being a case less than or equal to t. Since the predicted risks should be as close as possible to 1 for cases and 0 for controls, the "optimal" predictiveness curve is a step function that jumps up from 0 to 1 at $v = 1 - \hat{\theta}$, where $\hat{\theta}$ is an estimate of the prevalence $\theta = \mathbb{P}(D = 1)$.

Under the assumption that the underlying prediction model is well calibrated, Huang et al. (2007) suggested interpreting the predictiveness curve as follows: Because a steep slope of the curve indicates that cases and controls are well separated, models with a predictiveness curve that is close to the "optimal" curve in Figure 1 should have a high discriminatory power. Consequently, the "optimal" prediction model is attained when the resulting predictiveness curve is equal to the step function in Figure 1. Based on this interpretation, Pepe et al. (2008b) suggested investigating the performance of prediction models by comparing their predictiveness curves. This strategy was adopted by Pencina, D'Agostino & Vasan (2010) and Moons et al. (2012), who assessed the predictive capacity of a new marker by comparing the resulting predictiveness curves with and without the new marker. In particu-

lar, Pencina et al. (2010) noticed that in their data the predictiveness curves with and without the new marker hardly differed. Specifically, the authors pointed out that there is a need in future research to explore whether this is "caused by the inadequacy of the markers under consideration, or if it is an inherent property of the predictiveness curve itself" (Pencina et al., 2010, p. 6).

In this paper, we carry out a detailed analysis of the properties of the predictiveness curve and review its role in the performance assessment of biomedical prediction models. A key result of our analysis is that the predictiveness curve should not be used as a criterion for marker comparison unless one has made sure that all prediction models under consideration are well calibrated. In particular, the predictiveness curve does not necessarily provide information on the added predictive value of a new marker, a result that is in line with the observation of Pencina et al. (2010). The main reason for these problems is contained in the definition of the predictiveness curve itself: From Equation (1), it is obvious that the curve is solely based on the distribution of the marker Y but not on the observed values of the binary outcome D in the data. Consequently, the predictiveness curve alone does not reveal whether a model fits the data well, and it is possible to derive inferior prediction models that nevertheless result in the "optimal" predictiveness curve of Figure 1. Examples are given in Section 6.

Because it has been recognized that the predictiveness curve does not adequately measure calibration (Cook, 2010), several authors have suggested first checking whether all prediction models under consideration are well calibrated (Cook, 2010; Pepe et al., 2013). This could be done, for example, with the help of goodness-of-fit tests such as the Hosmer-Lemeshow test (Pepe et al., 2008b; Pepe, 2010). In the next step, predictive performance would be based on a comparison of the slopes of the resulting predictiveness curves. Although this approach solves part of the problem, it does not overcome the original shortcomings in the definition of the predictiveness curve. In particular, we will demonstrate in Section 6 that the Hosmer-Lemeshow test is not sensitive enough to reliably separate well calibrated from badly calibrated models (see also Peek et al., 2007; Cook, 2010). Consequently, artefacts are still possible, and it is unclear how one should compare markers via their predictiveness curves in these situations.

To overcome these problems, we propose a new graphical tool, the residual-based predictiveness curve (hereinafter abbreviated as *RBP curve*), which does not depend on the results of previously conducted calibration checks. The idea is to incorporate the binary outcome values into the definition of the predictiveness curve *itself*, thereby defining a new evaluation criterion for marker performance that also reveals how well a model is calibrated. The properties of this RBP curve will be assessed in detail in Section 5, and it will be demonstrated that it can address the aforementioned problems. Specifically, it is possible to obtain valid performance assessments via graphical checks of the RBP curve. Also, several well-known performance criteria, such as the true positive rate (TPR), the false positive rate (FPR) and the proportion of explained variation (PEV) (see Sachs & Zhou, 2013) can be derived graphically from the RBP curve.

2 Notation and Definitions

In the biomedical sciences, binary outcomes often refer to the status of a disease, with diseased ($D = 1$) and healthy ($D = 0$) subjects being considered cases and controls, respectively. Unless otherwise stated, these terms will be used synonymously in the rest of this paper. To predict the disease status D, we will consider a continuous marker (or marker combination) that will be denoted by Y. The observed value of Y will be denoted by η. The risk of being diseased given a particular value of Y will be defined as $risk(\eta) := \mathbb{P}(D = 1 \,|\, Y = \eta)$. Generally, there are two possibilities to obtain predictions for D: The first possibility is to use a single marker Y, for example the total prostate specific antigen (PSA) to detect prostate cancer or the CA-125 antigen to detect ovarian cancer. The second possibility is to define Y as a score, which is often constructed as a linear combination of single markers. Scores for the early detection of prostate cancer can, for example, be defined by combinations of PSA and other blood-based markers (see Pepe et al., 2008b; Shariat et al., 2011). The values of the score are usually derived by fitting a *prediction model* to a data set $\mathcal{D} = \{(\mathbf{x}_i, D_i), i = 1, \ldots, n\}$, where D_i denotes the disease status of subject i and $\mathbf{x}_i = (1, x_{i,1}, \ldots, x_{i,P})^\top$ is the associated vector of P marker values that may include the constant 1 for the intercept term. The score values are often assumed to be a linear combination of the marker values, i.e. $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$, $i = 1, \ldots, n$, where $\boldsymbol{\beta}$ is a vector of coefficients that has to be estimated. Typically, a prediction model is obtained by modelling the risk

$$risk(\eta_i) = \mathbb{P}(D_i = 1 \,|\, Y_i = \eta_i) = G(\mathbf{x}_i^\top \boldsymbol{\beta}), \qquad i = 1, \ldots, n, \tag{2}$$

using a parametric function G that transforms the range of the score to the unit interval $(0, 1)$. This parametrization is widely used in parametric modelling, where in general G has the form of a cumulative distribution function (cdf). For example, in logistic regression, G is the logistic distribution function, and the risk is given by

$$risk(\eta_i) = \mathbb{P}(D_i = 1 \,|\, Y_i = \eta_i) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}.$$

3 Calibration and Discrimination

In the literature, a large variety of approaches have been suggested to measure the predictive capacity of a marker Y (see Pepe et al., 2008b; Steyerberg et al., 2010). Among the most important principles are *calibration* and *discrimination*. Generally, a well calibrated prediction model is achieved when, for appropriate groupings of the observations, the predicted risks closely agree with the proportion of diseased in each group. An appropriate grouping, which we will consider, is grouping the observations via quantiles of the predicted risks. A well discriminating model is able to separate cases from controls, i.e., when both sensitivity (the proportion of diseased subjects that are correctly identified as having the disease) and specificity (the proportion of healthy subjects that are correctly identified as not having the disease) are high. While both concepts, calibration and discrimination, are well-established quality

criteria for markers in biomedical research, an important issue is how to quantify them, i.e., how to define appropriate measures for evaluating calibration and discrimination. In this respect, Huang et al. (2007) and Pepe et al. (2008b) have shown that many relevant properties of a marker can be summarized through the *predictiveness curve*.

3.1 Measures of Calibration

Pepe et al. (2013) distinguish between *good calibration* of a model and a *well calibrated* model. A measure for *good calibration* is given by the mean calibration, which is also known as *calibration-in-the-large* and measures how well the average risk (over all observations) agrees with the proportion of diseased subjects. In other words, *good calibration* is satisfied when the mean predicted risk, i.e. $\mathbb{E}(risk(Y)) = \frac{1}{n}\sum_{i=1}^{n}\widehat{risk}(\eta_i)$, is close to the estimated prevalence $\hat{\theta}$. Interestingly, the calibration-in-the-large is equal to the area under the predictiveness curve, which should therefore be as close as possible to $\hat{\theta}$ to ensure *good calibration* (Huang & Pepe, 2009b). In the case of the "optimal" predictiveness curve, for example, the area under the curve is simply the area of a rectangle with width $\hat{\theta}$ and height 1.

Although the calibration-in-the-large is a valuable tool for calibration assessment, Cook (2010) pointed out that evaluating this measure is not sufficient to fully assess the calibration of a prediction model. This is because the predicted risks of a well calibrated model should also be close to the proportions of diseased individuals when the observations are grouped and only a subset of observations is considered. Therefore, a measure for a *well calibrated model* can be obtained by grouping the observations via quantiles of the predicted risks and comparing the mean of the predicted risk in each group with the respective proportion of diseased within each group. This strategy gives rise to the Hosmer-Lemeshow statistic (see Crowson et al., 2014). Pepe et al. (2008b) suggested visualizing the components of the Hosmer-Lemeshow statistic by additionally plotting the observed proportions of diseased at the midpoint of each decile of the predicted risks (see Figure 2). If the points are close to the predictiveness curve, the prediction model in question is likely to be well calibrated.

3.2 Measures of Discrimination

A well discriminating model is obtained when the prediction model is able to separate cases from controls, i.e., when both the sensitivity (the proportion of diseased subjects that are correctly identified as having the disease) and the specificity (the proportion of healthy subjects that are correctly identified as not having the disease) are high. The discriminatory power of a *well calibrated* prediction model can be visually assessed by inspecting the slope of the predictiveness curve. Because the observations with higher risks are located to the right side of the vertical line at $v = 1 - \hat{\theta}$, a well discriminating model should assign risk values close to 1 to these observations. Conversely, it should assign risk values close to 0 to the observations to the left side (i.e., if $v < 1 - \hat{\theta}$). A well discriminating model should therefore result in a predictiveness curve with steep slope (see also Huang et al., 2007; Huang & Pepe, 2009a; Moons et al., 2012). In Figure 2, for example, the step function that represents the "optimal" predictiveness curve jumps from zero to one at $1 - \hat{\theta} = 0.7$. This implies that the step function has the

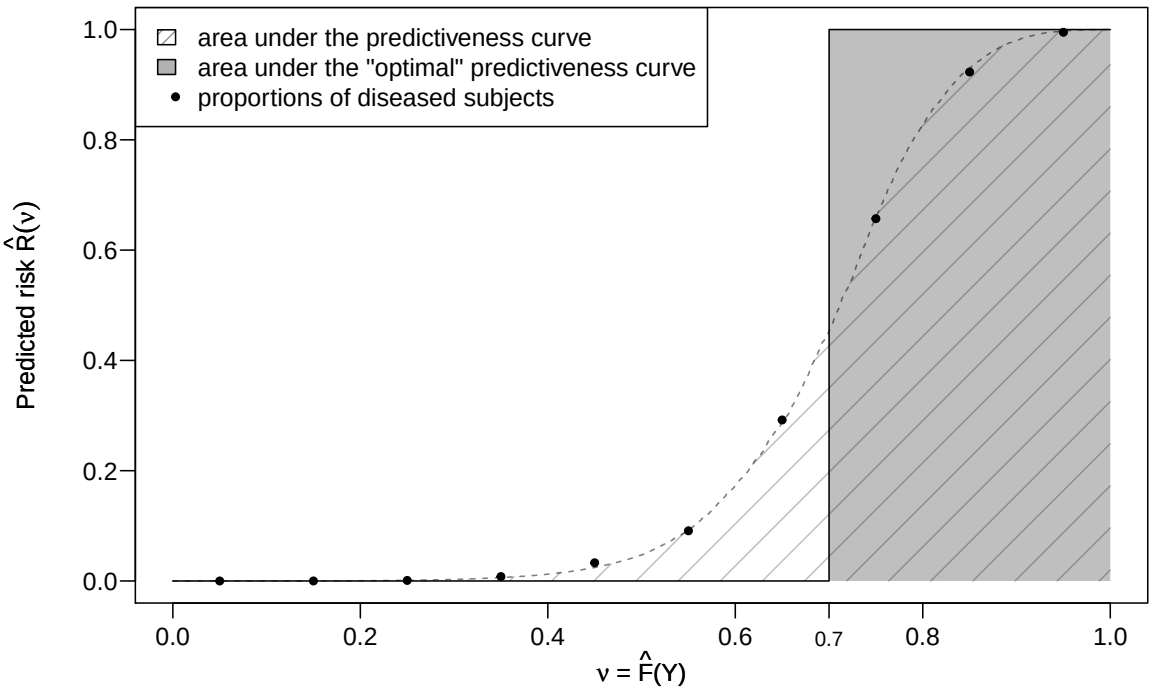

Figure 2: Predictiveness curves obtained from the simulated data set of Section 1. The dashed gray curve was obtained from a logistic regression model, whereas the step function corresponds to the optimal predictiveness curve that was obtained from the binary outcome values. The black points correspond to the proportion of diseased within each decile of predicted risks. Because the data were generated with the help of a logistic regression model, the black points are close to the respective predictiveness curve. This indicates that the logistic regression model is well calibrated.

steepest possible slope and therefore the maximum possible discriminatory power.

A popular measure to quantify the discriminatory power of a prediction model is the proportion of explained variation (PEV), which can be shown to be equivalent to PEV $:= \mathbb{E}(risk(Y)\,|\,D = 1) - \mathbb{E}(risk(Y)\,|\,D = 0)$ (see Pepe et al., 2008a). The PEV measure is the difference between the conditional expectation of $risk(Y)$ in the diseased group and the respective expectation in the healthy group (Pepe et al., 2008a). Because the first expectation should be close to 1 and the second one close to 0 in a well discriminating prediction model, the PEV is directly related to the slope of the predictiveness curve. The steeper the slope, the better the population is separated into two groups. If, at the same time, the model is *well calibrated*, a steep slope of the curve implies a large value of PEV. Generally, PEV ranges from 0 (implying minimum possible discrimination) to 1 (implying optimal discrimination). Empirically, the measure can be estimated by

$$\widehat{\text{PEV}}(\boldsymbol{\beta}) = \frac{1}{n_1} \sum_{i:D_i=1} G(\mathbf{x}_i^\top \boldsymbol{\beta}) - \frac{1}{n_0} \sum_{i:D_i=0} G(\mathbf{x}_i^\top \boldsymbol{\beta}), \tag{3}$$

where $G(\mathbf{x}_i^\top \boldsymbol{\beta})$ is the parametrization of the risk as in Equation (2) and $n_j := \sum_{i=1}^{n} I(D_i = j)$, $j = 0, 1$, is the number of healthy and diseased subjects, respectively. Our following analysis will be mainly based on the PEV measure; the relationships of other measures to the predictiveness curve were discussed in Gu & Pepe (2009).

4 Shortcomings of the Predictiveness Curve

Although the predictiveness curve has proved to be a valuable tool for the characterization of biomarker combinations, several shortcomings regarding the evaluation of prediction models remain. In particular, it can be misleading to rely on the predictiveness curve when the aim is to *compare* different marker combinations. This problem is mainly due to the fact that the predictiveness curve does not involve any of the true values of the outcome variable D. Consequently, except for the calibration-in-the-large (which can be evaluated by inspecting the area under the predictiveness curve), the curve does not provide any information on how well the prediction models under consideration are calibrated. It is therefore possible to construct algorithms that "optimize" marker combinations such that they result in the "optimal" step function shown in Figure 1. In Section 6.2, we will present such an algorithm and illustrate that the resulting marker combination is impractical and badly calibrated.

The aforementioned problems can also be viewed from a different angle: Recently, Hilden & Gerds (2014) analyzed the *integrated discrimination improvement* (IDI) measure and argued that IDI is insufficient for comparing the predictive capacity of two prediction models. The IDI can be defined as the difference of two PEV values of two prediction models (see Pepe et al., 2008a). Because the PEV reflects the discriminatory power of a model and does not provide information on whether the model is well calibrated, the PEV only measures part of the predictive capacity of a marker. As there is a relationship between PEV and the slope of the predictiveness curve (see Section 3.2), the problems identified by Hilden & Gerds (2014) naturally apply to the predictiveness curve as well.

Another problem is that the predictiveness curve cannot be used to evaluate whether a prediction model overfitted the data it was derived from. Evaluating the curve on an external test data set will not help in this respect, as the predictiveness curve only displays the distribution of $risk(Y)$. Consequently, if the covariates in the test data follow the same distribution as the covariates in the original data, both curves will be highly similar regardless of whether the model overfits the original data or not. Again, this problem arises from the fact that the predictiveness curve does not involve the true outcome status D that is needed to evaluate prediction accuracy. We will illustrate this issue in detail in Section 6.1.

As noted in Section 1, several authors have suggested conducting marker comparisons only if the marker combinations under consideration are well calibrated. The current state of the art is therefore to test whether prediction models are well calibrated and then to compare the predictiveness curves with respect to their slopes (Pepe et al., 2013). Calibration can, for example, be investigated by the Hosmer-Lemeshow test (Pepe et al., 2008b; Crowson et al., 2014). Although this strategy solves part of the problem, wrong conclusions are still possible: Because of power issues of the Hosmer-Lemeshow test (Peek et al., 2007; Cook, 2010), it is likely that prediction models with very different levels of

calibration will enter the final slope comparison. Also, the two-step approach automatically discards models that do not pass the calibration test, which implies that no graphical checks of these models are provided. Again we will illustrate these issues in detail in Section 6.3.

To address these shortcomings, we propose the *residual-based predictiveness (RBP) curve*, which is an improved graphical tool to evaluate and compare the accuracy of prediction models. We will start with the definition of the RBP curve and then show how various other popular measures of calibration and discrimination can be derived from it. In Section 6, we will present the results of several simulation studies that demonstrate how the RBP curve can be used to remedy the shortcomings of the original curve.

5 The RBP Curve

Instead of plotting $R(v)$ versus v or, equivalently $risk(Y)$ versus $F(Y)$ (see Gu & Pepe, 2009), we propose using the residuals $\epsilon = D - risk(Y)$ and plot ϵ versus $F_\epsilon(\epsilon)$. The index ϵ in $F_\epsilon(\epsilon)$ (the cdf of the residuals) was introduced to emphasize the difference between $F_\epsilon(\cdot)$ and $F(\cdot)$, which is the cdf of Y. This strategy allows us to directly assess the calibration of a prediction model (without plotting any additional points) because the real outcome D is included in the definition of the curve. In particular, considering the residuals ϵ makes it easier to assess the predictive capacities when comparing RBP curves. When statistical estimation is involved, we consider the estimated residuals $\hat{\epsilon}_i = D_i - \hat{p}_i$, where $\hat{p}_i := \widehat{risk}(\eta_i) \in [0, 1]$ and $D_i \in \{0, 1\}$.

To generate Figure 3, we used the simulated data and the predictions from the logistic model described in Section 1. The figure shows the RBP curve, and also how different criteria (calibration-in-the-large, calibration across deciles, PEV, FPR, TPR) can be obtained from the RBP curve. The horizontal line corresponds to the optimal RBP curve, where all residuals are zero. By definition, all non-diseased subjects $\{i : D_i = 0\}$ are located below the horizontal zero line, as $\hat{\epsilon}_i = D_i - \hat{p}_i = -\hat{p}_i \leq 0$. Accordingly, all diseased subjects $\{i : D_i = 1\}$ are located above the horizontal line, as $\hat{\epsilon}_i = D_i - \hat{p}_i = 1 - \hat{p}_i \geq 0$. The vertical line that splits diseased and non-diseased subjects is located at $1 - \hat{\theta}$ (one minus the prevalence). Therefore, we can derive the proportion of diseased from this line (which is $\hat{\theta} = 0.3$ in Figure 3(a)). Detailed descriptions of the relation between the RBP curve and several performance measures are given in the following sections.

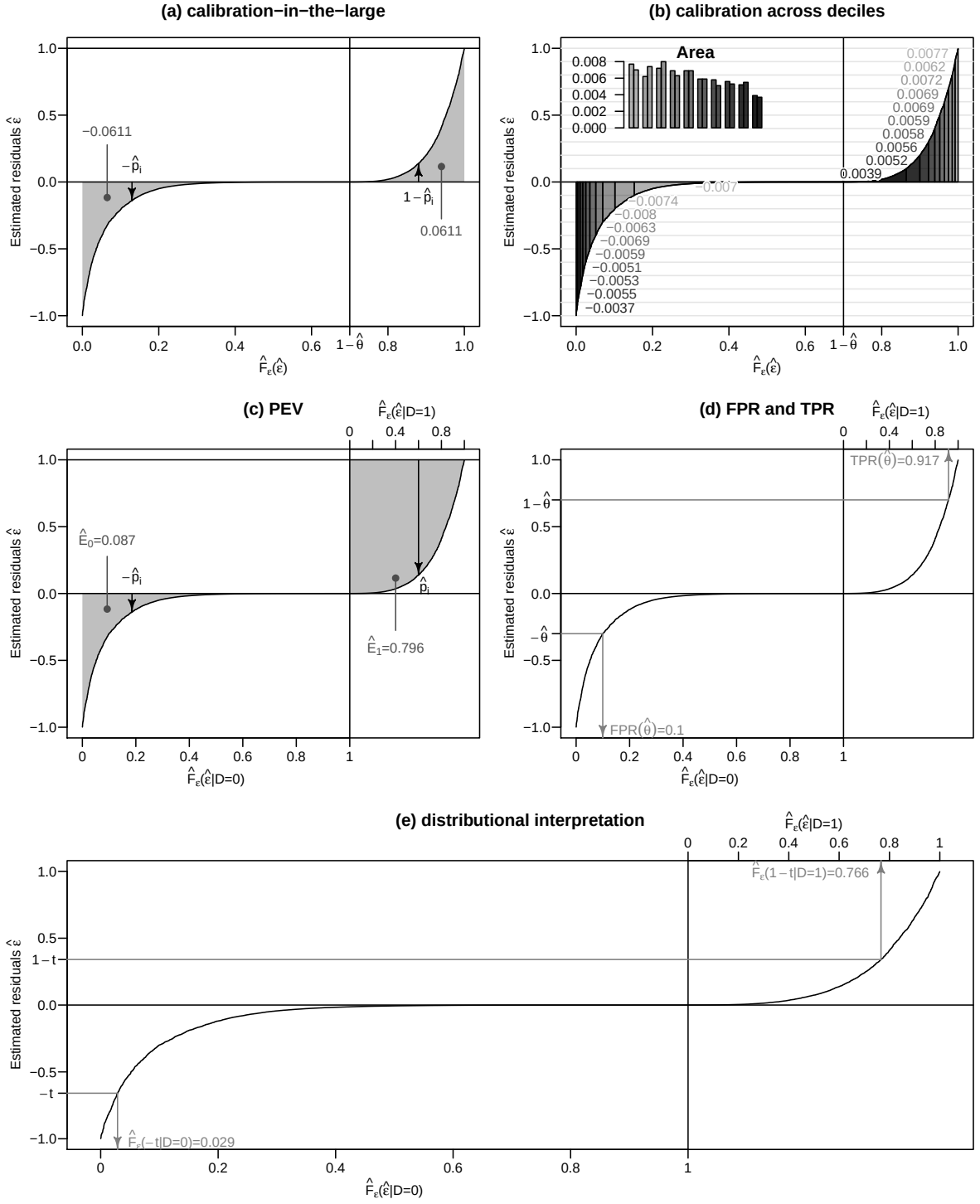

Figure 3: Derivation of performance measures from the RBP curve as described in Section 5. The numbers in panels (a) and (b) correspond to the areas of the respective regions. The bar plots in the second figure facilitate the comparison of areas with the same gray tones, which should be equal in magnitude.

5.1 Relation to the Calibration-in-the-large

The calibration-in-the-large reflects whether a prediction model satisfies *good calibration*, meaning that the expected value of the risk, which can be estimated by $\frac{1}{n}\sum_{i=1}^{n}\hat{p}_i$, should be as close as possible to the prevalence $\theta = P(D = 1)$, which is estimated by $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} D_i$. It follows that the risks should satisfy $\theta - E(risk(Y)) = 0$ when *good calibration* is satisfied. The empirical counterpart

$$\frac{1}{n}\sum_{i=1}^{n} D_i - \frac{1}{n}\sum_{i=1}^{n}\hat{p}_i$$

can be interpreted as the integral below the RBP curve, which can be approximated by the sample mean

$$\hat{A}_{\text{RBP}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i = \frac{1}{n}\sum_{i=1}^{n}(D_i - \hat{p}_i) = \frac{1}{n}\sum_{i=1}^{n} D_i - \frac{1}{n}\sum_{i=1}^{n}\hat{p}_i.$$

Therefore, $\hat{A}_{\text{RBP}} = 0$ ensures *good calibration*. Here, we prefer using the term *integral* rather than *area* because $\hat{A}_{\text{RBP}}$ can also take negative values. When the term *area* is used, we are referring to the positive valued integral. Visually, *good calibration* implies that the area above the horizontal zero line and the area below the horizontal zero line should be approximately equal, or equivalently, the sum of both integrals should be as close as possible to zero. This is because the integral below the RBP curve can be rewritten as

$$\hat{A}_{\text{RBP}} = \frac{1}{n}\sum_{i=1}^{n}(D_i - \hat{p}_i) = \frac{1}{n}\left\{\sum_{i:D_i=0}(D_i - \hat{p}_i) + \sum_{i:D_i=1}(D_i - \hat{p}_i)\right\} = \frac{1}{n}\sum_{i:D_i=0}(0 - \hat{p}_i) + \frac{1}{n}\sum_{i:D_i=1}(1 - \hat{p}_i),$$

where the first sum corresponds to the integral below the horizontal zero line and the second sum corresponds the integral above the horizontal zero line. For example, in Figure 3(a), the integral below the horizontal zero line is $\frac{1}{n}\sum_{i:D_i=0}(-\hat{p}_i) = -0.0611$ and the integral above the horizontal zero line is $\frac{1}{n}\sum_{i:D_i=1}(1 - \hat{p}_i) = 0.0611$.

5.2 Relation to the Calibration across Deciles

In a similar manner to the calibration-in-the-large, the calibration across deciles is visually obtained by contrasting the area above the horizontal zero line with the area below the horizontal zero line. For the calibration across deciles, we need to additionally split the integral according to deciles of the predicted risks. The relationship between the RBP curve and the calibration across deciles can then be derived as follows: Let $Q_q = \{i : \hat{p}_i \in (q - 0.1, q)\}$ with $q = 0.1, 0.2, \ldots, 1$, be the sets of observations whose predicted risks are included in intervals that are bounded by the $((q-1)*10)$-th and $(q*10)$-th deciles, respectively. For example, $Q_{0.1} = \{i : \hat{p}_i \in (0, 0.1)\}$ is the set of observations whose predicted risks are below the first decile. Furthermore, let $\hat{A}^q_{RBP}$, $q = 0.1, 0.2, \ldots, 1$, be the integrals that are obtained by splitting $\hat{A}_{RBP}$ according to deciles of the predicted risk, so that $\hat{A}_{RBP} = \hat{A}^{0.1}_{RBP} + \hat{A}^{0.2}_{RBP} + \ldots + \hat{A}^{1}_{RBP}$. For

example, the integral for the first decile is estimated by

$$\hat{A}^{0.1}_{RBP} = \frac{1}{n_{0.1}} \sum_{i \in Q_{0.1}} (D_i - \hat{p}_i) = \frac{1}{n_{0.1}} \sum_{i \in Q_{0.1}} D_i - \frac{1}{n_{0.1}} \sum_{i \in Q_{0.1}} \hat{p}_i,$$

where $n_{0.1} = |Q_{0.1}|$ is the number of observations below the first decile of predicted risks. Analogously to the previous section, this integral can be rewritten as

$$\hat{A}^{0.1}_{RBP} = \frac{1}{n_{0.1}} \sum_{\{i: i \in Q_{0.1} \wedge D_i = 0\}} (-\hat{p}_i) + \frac{1}{n_{0.1}} \sum_{\{i: i \in Q_{0.1} \wedge D_i = 1\}} (1 - \hat{p}_i),$$

where the first sum corresponds to the integral below the horizontal zero line (using only the observations, whose predicted risks are below the first decile, i.e., for $i \in Q_{0.1} = \{i : \hat{p}_i \in (0, 0.1)\}$). The second sum corresponds to the integral above the horizontal zero line, also using only the observations $i \in Q_{0.1} = \{i : \hat{p}_i \in (0, 0.1)\}$. When both integrals are equal in magnitude, they sum up to zero, yielding $\hat{A}^{0.1}_{RBP} = 0$, so that perfect calibration for the first decile is satisfied. A well calibrated prediction model in terms of the calibration across deciles is given when each integral $\hat{A}^{0.1}_{RBP}, \hat{A}^{0.2}_{RBP}, \ldots, \hat{A}^{1}_{RBP}$ is close to zero. In Figure 3(b), this is visualized by contrasting areas with the same gray tones, where each area with the same gray tone contains the predicted risks for a specific decile. In the case of a well calibrated prediction model, the areas with the same gray tones above the horizontal zero line and below the horizontal zero line should be approximately equal.

5.3 Relation to PEV

The PEV measure is visually obtained by the difference of the gray areas in Figure 3(c). By definition, the gray shaded area above the horizontal line corresponds to the conditional expectation $E_1 := E(risk(Y) \,|\, D = 1)$ in the diseased group (estimated by $\hat{E}_1 = \frac{1}{n_1} \sum_{i:D_i=1} \hat{p}_i$), whereas the gray shaded area below the horizontal line corresponds to the conditional expectation $E_0 := E(risk(Y) \,|\, D = 0)$ in the non-diseased group (estimated by $\hat{E}_0 = \frac{1}{n_0} \sum_{i:D_i=0} \hat{p}_i$). It follows that the PEV measure, which is the difference of the two conditional expectations, is estimated by $\widehat{\text{PEV}} = \hat{E}_1 - \hat{E}_0$. For example, in Figure 3(c) one obtains

$$\widehat{\text{PEV}} = \hat{E}_1 - \hat{E}_0 = 0.796 - 0.087 = 0.709.$$

5.4 Relation to FPR and TPR

By definition, true and false positive rates are given by

$$TPR(\theta) = P(risk(Y) > \theta \,|\, D = 1) = P(1 - risk(Y) < 1 - \theta \,|\, D = 1) = F_\epsilon(1 - \theta \,|\, D = 1)$$

and

$$FPR(\theta) = P(risk(Y) > \theta \,|\, D = 0) = P(0 - risk(Y) < 0 - \theta \,|\, D = 0) = F_\epsilon(0 - \theta \,|\, D = 0),$$

respectively. Therefore, estimated values of FPR and TPR are given by the intersection of the RBP curve with the horizontal lines at $0 - \hat{\theta}$ and $1 - \hat{\theta}$, respectively (see Figure 3(d)).

5.5 Distributional Interpretation

As shown in Section 1, the predictiveness curve contains the full distribution of the risk $R(\nu)$, which is depicted as a function of the quantiles of the marker Y. It follows that one can make interpretations about the proportion of the population having a risk of being diseased less than or equal to a specific value t (see also Figure 1). Analogous distributional interpretations can be derived from the RBP curve. This is most easily seen when writing the *proportion of the population having a risk of being diseased less than or equal to* t, $P(risk(Y) \leq t)$, in terms of the TPR, FPR and the prevalence θ, whose values can all be obtained from the RBP curve:

$$\begin{aligned} P(risk(Y) \leq t) &= 1 - P(risk(Y) > t) \\ &= 1 - [P(risk(Y) > t \,|\, D = 1) \cdot P(D = 1) + P(risk(Y) > t \,|\, D = 0) \cdot \{1 - P(D = 1)\}] \\ &= 1 - \{TPR(t) \cdot \theta + FPR(t) \cdot (1 - \theta)\}\,. \end{aligned}$$

The values for $TPR(t)$ and $FPR(t)$ are given by the intersection of the curve and the horizontal lines at $1 - t$ and $-t$.

In Figure 3(e), for example, we used $t = 0.66$ and obtained $TPR(t) = 0.766$ and $FPR(t) = 0.029$. Therefore, it follows that $P(risk(Y) \leq t) = 0.75$, which yields the same distributional interpretation as in Figure 1, namely that 75% of the population have a risk less than or equal to 66%.

5.6 Relation to the L_1 and L_2 Loss

We finally note that there is a direct relationship between the RBP curve and the mean absolute error (MAE) based on the L_1 loss

$$MAE = \frac{1}{n}\sum_{i=1}^{n} L_1(D_i, \hat{p}_i) = \frac{1}{n}\sum_{i=1}^{n} |D_i - \hat{p}_i| = \frac{1}{n}\sum_{i=1}^{n} |\epsilon_i|. \tag{4}$$

As seen from the definition of the residuals $\epsilon_i = D_i - \hat{p}_i$, the RBP curve depicts the signed summands of the L_1 loss in Equation (4). Visually, the MAE reflects the sum of the two positive valued integrals below and above the RBP curve. For example, in Figure 3(a) the area below the horizontal zero line (that is, the positive valued integral) is 0.0611, and the area above the horizontal zero line is 0.0611, yielding $MAE = 0.0611 + 0.0611 = 0.1222$. As a consequence, changes in the MAE will automatically be reflected in changes of the RBP curve, rendering the RBP curve a convenient tool for a detailed graphical illustration of the MAE.

In addition, there is an indirect relation between the RBP curve and the Brier score, which is based

on the L_2 loss and is obtained by averaging the squared residuals ϵ_i^2, $i = 1, \ldots, n$:

$$BS = \frac{1}{n}\sum_{i=1}^{n} L_2(D_i, \hat{p}_i) = \frac{1}{n}\sum_{i=1}^{n}(D_i - \hat{p}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2. \tag{5}$$

By definition, the RBP curve depicts the signed square roots of the summands in Equation (5). Therefore, changes in the Brier score are also reflected in changes of the RBP curve.

6 Simulations

To illustrate the importance of calibration checks before assessing the predictive capacity of a prediction model, we start with a model that massively overfits the data. By evaluating this model with the predictiveness curve, we demonstrate that calibration checks are not possible without consideration of the outcome variable D. Furthermore, we show that it is possible to construct well calibrated prediction models with almost completely overlapping predictiveness curves (suggesting similar predictive capacities), even in the case in which one of the models ignores an influential marker. In contrast to the original predictiveness curves, examination of the RBP curves can reveal both overfitting and the inclusion/exclusion of influential markers.

6.1 Overfitting Issues

An overfitting prediction model, which often occurs when there are too many markers, is able to fit a specific data set well, but fails in predicting future values of D (implying that the model has a poor predictive capacity). To construct an overfitting model, we generated a data set $\mathcal{D} = \{(\mathbf{x}_i, D_i), i = 1, \ldots, n\}$ with $n = 3000$ observations and 1000 non-informative markers $\mathbf{x}_i^\top = (x_{i,1}, \ldots, x_{i,1000})^\top$.

The data were subdivided into a training set $\mathcal{L} = \{(\mathbf{x}_i, D_i), i = 1, \ldots, 1800\}$ and a test set $\mathcal{T} = \mathcal{D} \setminus \mathcal{L}$. Based on the training data, we estimated a logistic regression model that overfitted the training data. The model was subsequently evaluated on the test data set. Panels (a) and (b) in Figure 4 show the predictiveness curves for the predicted risks of the training and test data. In particular, one can see that the shapes of both predictiveness curves are similar to the respective "optimal" predictiveness curves, suggesting an optimal predictive capacity of the model. On the other hand, when considering the calibration points in Figure 4(b) (corresponding to the test data), one sees that the prediction model is not well calibrated at all and has therefore a poor predictive capacity. One can see that no valid calibration checks are possible without adding the calibration points to the predictiveness curve. Further, the predictive capacity of a prediction model cannot be adequately assessed without plotting the points obtained from the test data.

Figure 4(c) shows the RBP curves obtained from the training and test data. As expected, the RBP curve of the training data is a horizontal line at zero. As it corresponds to a prediction model with extreme overfitting, it is very similar to the "optimal" RBP curve defined in Section 5. Unlike the predictiveness curve of the test data (depicted in Figure 4(b)), the RBP curve obtained from the test

data does not look optimal at all. By definition, the RBP curve depends on the real outcome values; hence it can reveal bad calibration and large prediction errors.

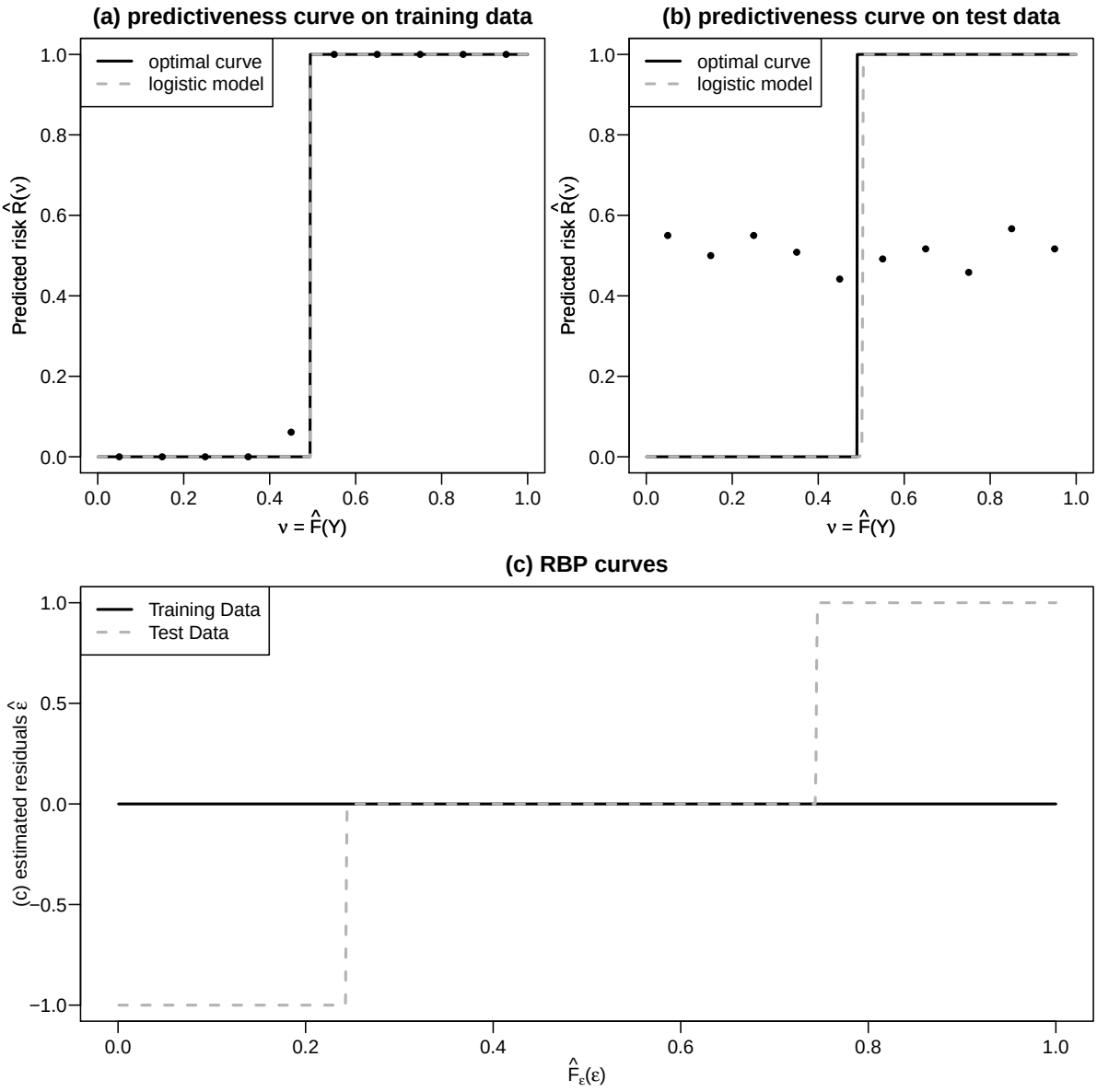

Figure 4: Predictiveness curves and RBP curves of the overfitting prediction model considered in Section 6.1. The points correspond to the observed proportions of diseased subjects in the risk deciles.

6.2 Pathological Optimization of the Predictiveness Curve

Here we demonstrate that it is even possible to optimize the predictiveness curve of a model with low predictive capacity such that it still looks "optimal", as defined by Figure 1. This can be achieved by direct maximization of the empirical PEV measure, which reflects the steepness of the slope of the predictiveness curve (see Section 3.2).

To ensure that the optimized predictiveness curve results in an adequate calibration-in-the large that satisfies $\mathbb{E}(risk(Y)) = \theta$ and hence corresponds to the area under the "optimal" predictiveness curve in Figure 1, we use the constraint

$$\frac{1}{n}\sum_{i=1}^{n} G(\mathbf{x}_i^\top \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} I(D_i = 1), \tag{6}$$

where the first sum is the empirical counterpart of the expected risk and the second sum is an estimate of the prevalence $\theta = \mathbb{P}(D = 1)$ based on samples $(\mathbf{x}_i, D_i)$, $i = 1, \ldots, n$. The resulting optimization problem can be written as

$$\begin{aligned} \max_{\boldsymbol{\beta}} \widehat{\text{PEV}}(\boldsymbol{\beta}) \;\text{ s.t. }\; & \frac{1}{n}\sum_{i=1}^{n} G(\mathbf{x}_i^\top \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} I(D_i = 1), \\ & \boldsymbol{\beta}^\top \boldsymbol{\beta} \leq t, \end{aligned} \tag{7}$$

where $\boldsymbol{\beta}^\top \boldsymbol{\beta} = \sum_{p=1}^{P} \beta_p^2 \leq t$ is an additional constraint that penalizes the size of the coefficients to address convergence issues and to include an optional shrinking effect for the coefficients as in ridge regression (see Tibshirani, 1996). In Figure 5, we illustrate that this additional constraint is able to control the steepness of the predictiveness curve.

To illustrate that solving (7) yields an "optimal" looking predictiveness curve for large t (implying large coefficients), we conducted a simulation study with $n = 10000$ observations and $P = 5$ standard normally distributed markers $X_1, \ldots, X_5 \sim N(0, 1)$. The disease status D was generated from the underlying model

$$\text{logit}(P(D = 1 \,|\, \mathbf{x}_i)) = \mathbf{x}_i^\top \boldsymbol{\beta}, \tag{8}$$

where the (randomly generated) coefficients $\boldsymbol{\beta} = (-2.71, -2.21, -2.59, -2.12, -2.06, -2.95)^\top$ were used. The first coefficient corresponds to the intercept and the others correspond to the influential markers.

The simulated data $\mathcal{D} = \{(\mathbf{x}_i, D_i), i = 1, \ldots, n\}$, with $n = 10000$ observations were subdivided into a training set $\mathcal{L} = \{(\mathbf{x}_i, D_i), i = 1, \ldots, 6000\}$, on which the estimation of the prediction model was based, and a test set $\mathcal{T} = \mathcal{D} \setminus \mathcal{L}$, on which the model was evaluated. The estimated coefficients that resulted from solving the optimization problem in (7) with three different boundaries for the squared sum of the coefficients ($t = 36$, $t = 360$ and $t = 3600$) are presented in Table 1. The resulting predictiveness curves for training and test sets are shown in Figure 5. The predictiveness curve for $t = 3600$ suggests a well discriminating prediction model because of its steep slope. At the same time, the area under the curve

(being close to the area under the "optimal" predictiveness curve) suggests a perfect calibration-in-the-large. As a consequence, the curve is close to the "optimal" predictiveness curve and suggests a perfect predictive capacity.

However, as emphasized in Section 3, one should also consider the calibration across deciles using the Hosmer-Lemeshow test. As noted by Pepe et al. (2008b), the components of the test statistic can be visualized through the predictiveness curve by additionally plotting the proportions of diseased for each decile and by comparing these proportions to the predictiveness curve.

Figure 5 suggests that only the prediction model with $t = 36$ was well calibrated. Obviously, this confirms that an "optimal" looking predictiveness curve does not necessarily imply a perfect predictive capacity of the model. Therefore, care must be taken when making comparisons of prediction models by solely comparing their predictiveness curves.

An improved strategy was proposed by Pepe et al. (2013), who recommended first checking the calibration of the prediction models under consideration (for example by applying the Hosmer-Lemeshow test) and then comparing the predictiveness curves to the models that passed the calibration tests. While this strategy addresses part of the problem, we will show in the next section that it can still yield misleading results.

	β_0	β_1	β_2	β_3	β_4	β_5
true coefficients	-2.71	-2.21	-2.59	-2.12	-2.06	-2.95
$t = 36$	-2.67	-2.11	-2.66	-2.04	-2.10	-2.97
$t = 360$	-8.13	-6.57	-8.58	-6.66	-6.68	-9.40
$t = 3600$	-25.54	-20.43	-27.12	-21.82	-21.02	-29.62

Table 1: Estimated coefficients obtained from the simulation study in Section 6.2. It can be seen that the regularization parameter t controls the magnitude of the coefficient estimates, thereby determining the steepness of the slope of the resulting predictiveness curve.

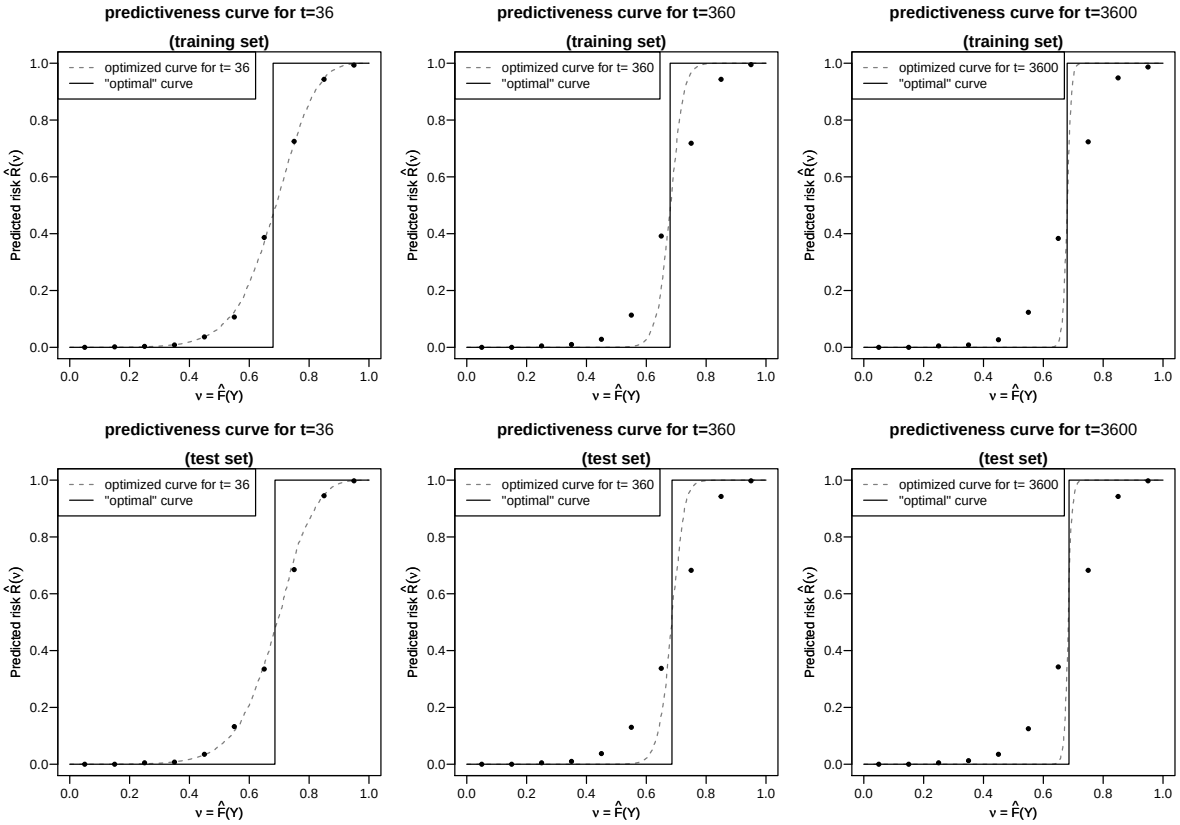

Figure 5: Simulation study of Section 6.2. The predictiveness curves are based on the predicted risks of the training and test data using different values of the regularization parameter t. As seen from the black points (which correspond to the proportions of diseased across deciles), only the model with $t = 36$ is well calibrated.

6.3 Model Comparisons

In this section we compare two well calibrated prediction models (one model with and the other model without a new marker) by their predictiveness curves. The aim of our simulation study is to demonstrate that making conclusions about the predictive capacity by solely using the predictiveness curve can be misleading – even if the prediction model is well calibrated.

We used the randomly generated coefficients $\boldsymbol{\beta} = (-2.71, -2.21, -2.59, -2.12, -2.06, 1)^\top$, where the first coefficient corresponds to the intercept and the others correspond to the influential markers, and simulated 1000 data sets with five markers $X_1, \ldots, X_5 \sim N(0, 1)$. Each of the 1000 data sets was split into a training set and a test set (containing 60% and 40% of the observations, respectively). Moreover, in each of the 1000 simulation runs we fitted one model with marker X_5 (model "M_1") and one model without X_5 (model "M_0") by solving the optimization problem in (7). The choice of the boundary t was based on the coefficients that resulted from a logistic model (denoted by $\hat{\boldsymbol{\beta}}_{\text{glm}}$). For model M_0

we used $t = 1.07 \cdot \hat{\boldsymbol{\beta}}_{\text{glm}}{}^{\top}\hat{\boldsymbol{\beta}}_{\text{glm}}$. This strategy resulted in a steeper predictiveness curve than the one resulting from the logistic model since the coefficients were allowed to be larger. For model M_1, we used $t = 0.91 \cdot \hat{\boldsymbol{\beta}}_{\text{glm}}{}^{\top}\hat{\boldsymbol{\beta}}_{\text{glm}}$ in order to obtain predictiveness curves with flatter slopes. We then applied the Hosmer-Lemeshow test in order to select those data sets whose predicted risks (for both the training and the test data) were *well calibrated*. In our simulation study, 114 out of the 1000 data sets passed the Hosmer-Lemeshow test. We used these 114 data sets for the following computations.

Figure 6 shows the averaged predictiveness curves obtained from the test data (panel (a)) as well as the averaged pointwise differences between the predictiveness curves obtained from the 114 test data sets (panel (c)). Since the curves are almost completely overlapping and the pointwise differences scatter around the zero line, the predictive capacities indicated by the predictiveness curves are similar for the two models. It is therefore difficult to tell whether the model with or without X_5 performed better, although, according to the data generating process, the model that includes X_5 should have resulted in a higher predictive capacity.

In contrast to panels (a) and (c), panels (b) and (d) of Figure 6 (which are based on the RBP curve instead of the predictiveness curve) indicate a difference between models M_1 and M_0. Specifically, panel (d) indicates that model M_1 has a higher predictive capacity than model M_0. This result was confirmed by the evaluation of various other performance measures on the test data (Figure 7).

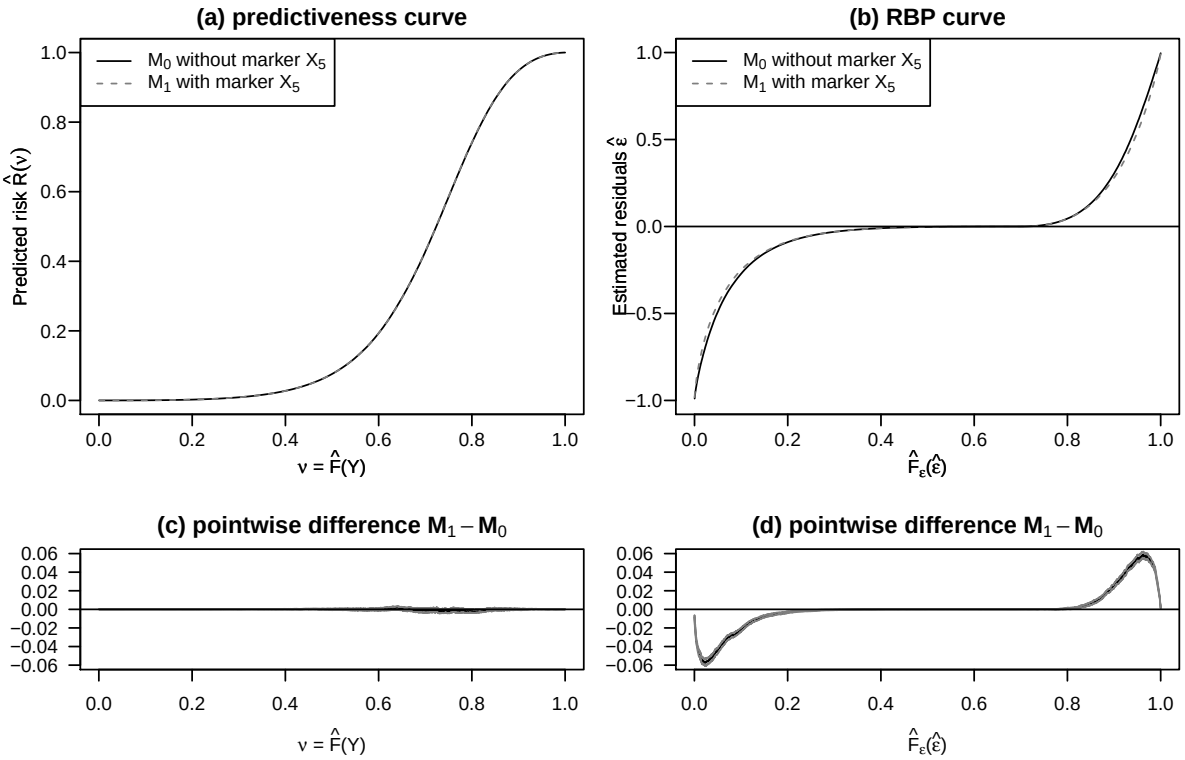

Figure 6: Results of the simulation study of Section 6.3. Panels (a) and (c) show the averaged predictiveness curves for models M_0 and M_1 and the averaged pointwise differences (with 95% bootstrap confidence bands) between the two predictiveness curves, respectively. Panels (b) and (d) are comparisons of the averaged RBP curves and the averaged pointwise differences (with 95% confidence bands), respectively.

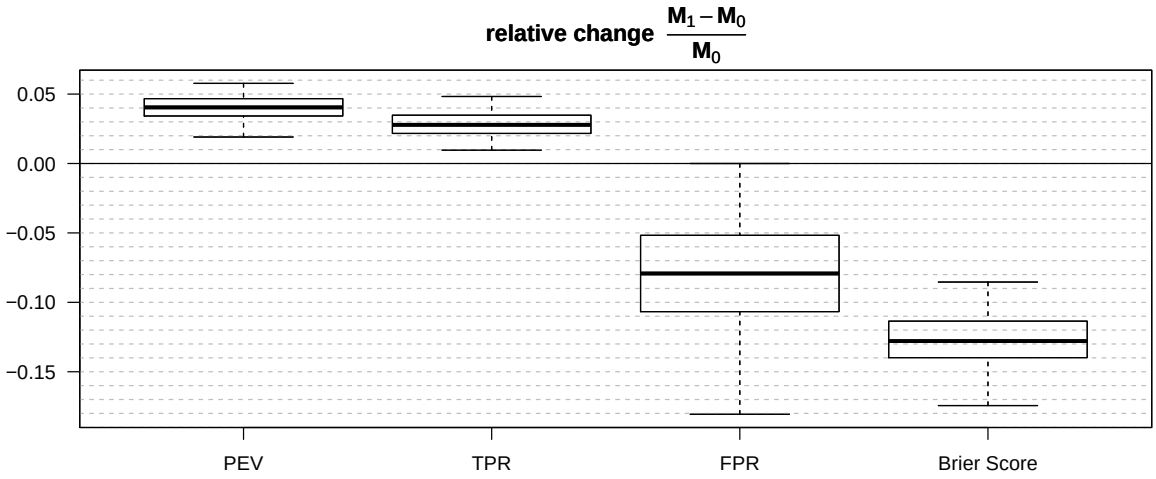

Figure 7: Results of the simulation study in Section 6.3. The box plots show the relative changes of various performance measures that were evaluated on the 114 test data sets that passed the Hosmer-Lemeshow calibration test. Note that lower false positive rates (FPR) and lower Brier scores indicate a better model performance. The respective box plots are therefore below the horizontal zero line.

7 Evaluation on Real Data

The data set considered in this section was collected by Hatzis et al. (2011). It contains information on patients with newly diagnosed ERBB2-negative breast cancer and is publicly available from the GEO repository (http://www.ncbi.nlm.nih.gov/geo). Covariate information is contained in a clinical data set and a high-throughput molecular data set (omics information). The data are stored in two separate data sets (GEO accession numbers GSE25055 and GSE25065) that, in our analysis, are merged together and randomly split into training set and test set containing $2/3$ and $1/3$ of the observations, respectively. The response variable considered here is the residual cancer burden (RCB) class. The levels RCB-0 and RCB-I (referring to no and minimal residual disease, respectively) were coded as $D = 0$, whereas levels RCB-II and RCB-III (referring to moderate and extensive residual disease, respectively) were coded as $D = 1$. Observations with missing values in the clinical data and/or the response (RCB class) were omitted, yielding training and test sets of sizes 254 and 128, respectively.

To construct a meaningful model that combines clinical and omics information, we used an approach that was originally proposed by De Bin, Sauerbrei & Boulesteix (2014) ("Strategy 4a"). The main idea of this strategy is to first fit a model to the omics data and to use the linear predictor obtained from this model (the so-called *omics score*) as an additional explanatory variable in a final logistic regression model that also contains the clinical predictors age, nodal status, tumor size, grade, estrogen receptor status, and progesterone receptor status. For the derivation of the omics score, we applied a modelling strategy that was able to handle the high-dimensional omics data. Specifically, we used a component-

wise gradient boosting algorithm with linear base-learners (Friedman, 2001; Bühlmann & Yu, 2003; Mayr et al., 2014) that automatically selected the most relevant predictors from the data. To avoid overfitting, the number of boosting iterations, which is the main tuning parameter of the algorithm, was optimized using 10-fold cross-validation. In our analysis, the optimal number of boosting iterations was 15.

Figure 8 shows the RBP curve and the predictiveness curve for a *meaningful* final model, where the omics score was computed using 15 boosting iterations, and for an *overfitting* final model, where the omics score was computed using 300 boosting iterations. It can be seen that the RBP curve displays the different predictive capacities of the models more clearly than the predictiveness curve, especially for the *overfitting* model (black curves). This can be illustrated by deriving the PEV, TPR and FPR measures from the RBP curves for training and test data for both models (see Figure 9). The first panel (top left) of the figure refers to the *overfitting* model and shows an optimal RBP curve, having $TPR = 1$, $FPR = 0$ and $\widehat{\text{PEV}} = \hat{E}_1 - \hat{E}_0 = 1$, which suggests a perfect fit on the training data. It can also be seen that the *overfitting* model did not perform well when predicting the test data (see top right panel) as opposed to the training data. For the test data, it is seen that the TPR is higher and the FPR is lower for the *meaningful* model than for the *overfitting* model, suggesting that the *meaningful* model performed better on the test data than the *overfitting* model.

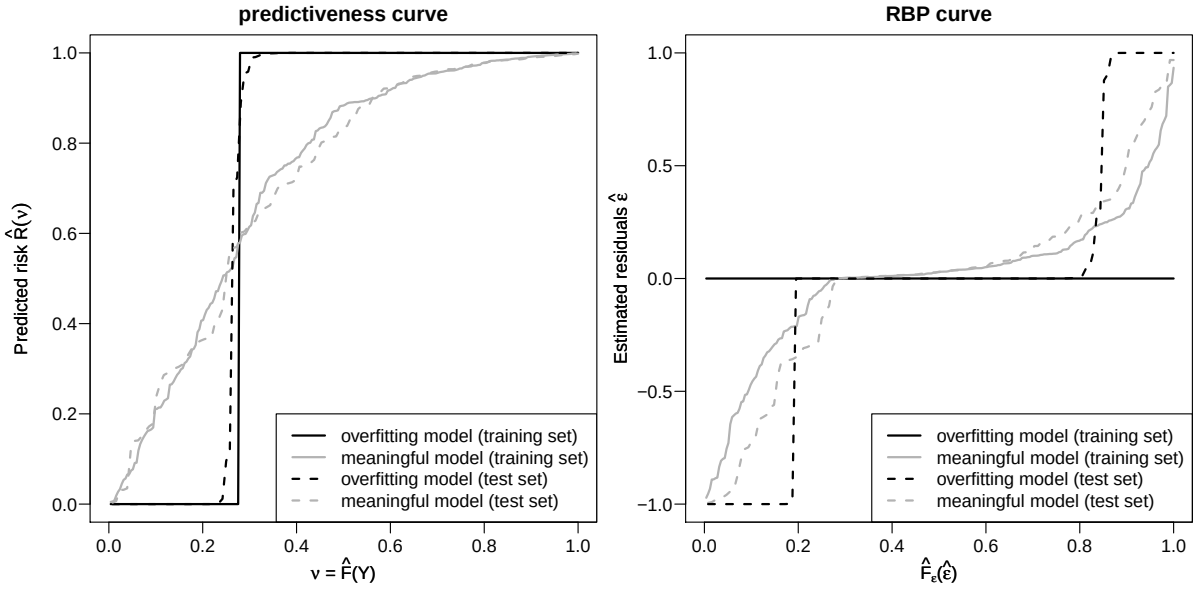

Figure 8: Predictiveness curves for training and test data of the *overfitting* and the *meaningful* model using the real data from Section 7.

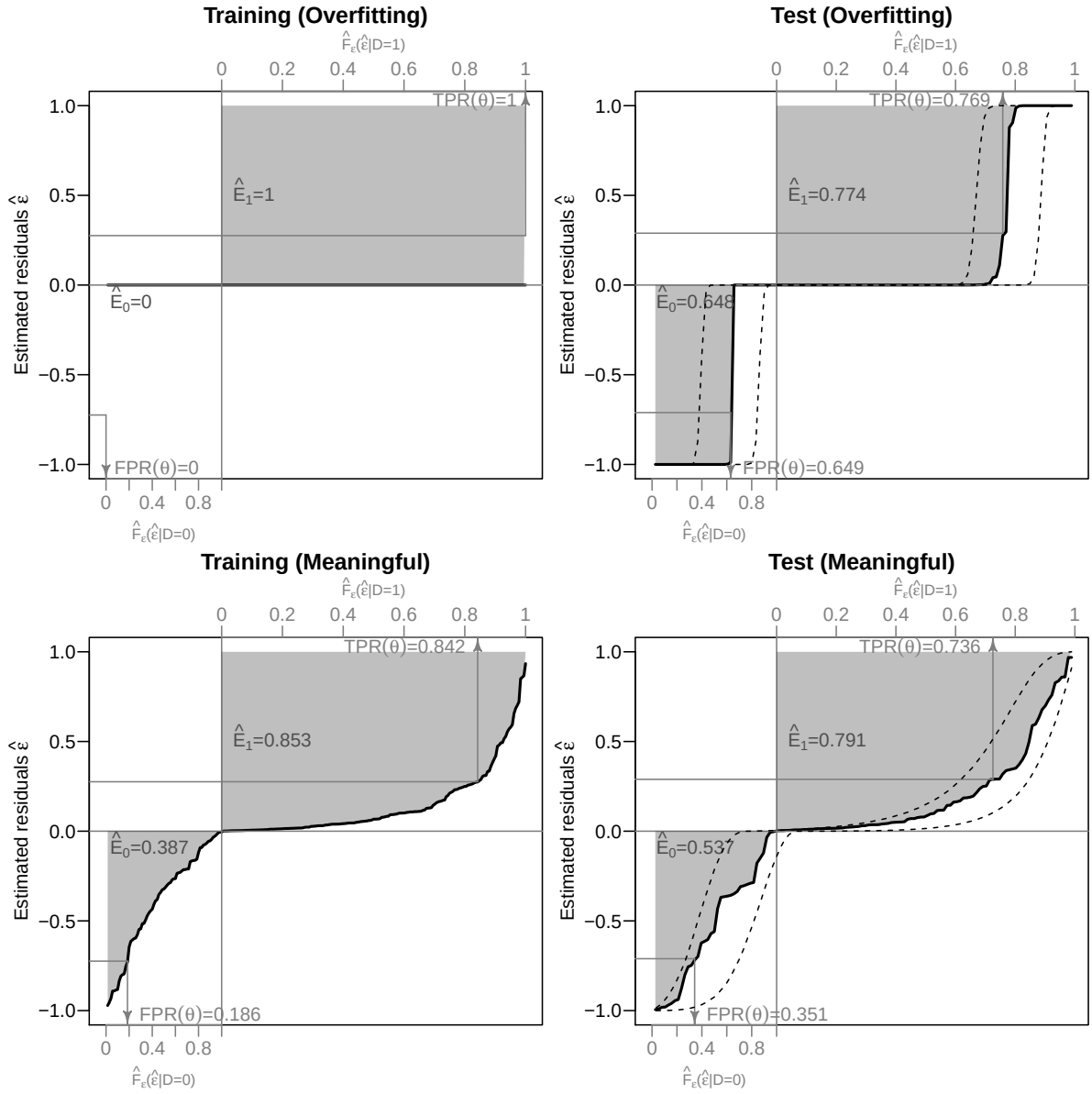

Figure 9: RBP curves for training and test data of the *overfitting* and the *meaningful* model using the real data from Section 7. The dashed lines correspond to 95% bootstrap confidence bands.

8 Summary and Discussion

Because the prediction of binary outcomes has become increasingly important in biomedical research, it is essential for practitioners to have reliable and unbiased measures of prediction accuracy. Regarding the derivation of summary measures, a remarkable number of strategies and easy-to-interpret coefficients (such as AUC and PEV) are available.

Equally important for practitioners is the availability of suitable *graphical* tools for visualizing prediction accuracy. In this respect, the RBP curve is a natural choice, since it contains the full distribution of the risks and several summary measures can be derived from it (see Section 5). It therefore reveals the most relevant properties of a prediction model. Moreover, we demonstrated that the RBP curve visualizes both discrimination and calibration and is therefore a suitable graphical tool for unbiased marker comparisons. In contrast to strategies for the original predictiveness curve, the RBP curve does not rely on (asymptotically valid) calibration tests and is therefore insensitive to power issues and convergence problems.

The RBP curve is implemented in the `R` add-on package `RBPcurve`, which is publicly available at http://cran.r-project.org (R Core Team, 2014). The package provides a user-friendly interface to generate RBP curves and enables users to visualize the relationships between the RBP curve and the performance measures discussed in Section 5.

Acknowledgements

Financial support from LMUexcellent is gratefully acknowledged.

References

Bühlmann, P. & Yu, B. (2003). Boosting with the L_2-Loss: Regression and Classification, *Journal of the American Statistical Association* **98**, 324–339.

Cook, N. R. (2010). Comment: Measures to Summarize and Compare the Predictive Capacity of Markers, *The International Journal of Biostatistics* **6**, Article 22.

Crowson, C. S., Atkinson, E. J. & Therneau, T. M. (2014). Assessing Calibration of Prognostic Risk Scores, *Statistical Methods in Medical Research*. DOI: 10.1177/0962280213497434.

De Bin, R., Sauerbrei, W. & Boulesteix, A.-L. (2014). Investigating the Prediction Ability of Survival Models based on both Clinical and Omics Data: Two Case Studies, *Statistics in Medicine* **33**, 5310–5329.

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine, *Annals of Statistics* **29**, 1189–1232.

Gu, W. & Pepe, M. S. (2009). Measures to Summarize and Compare the Predictive Capacity of Markers, *The International Journal of Biostatistics* **5**, 1–49.

Hatzis, C., Pusztai, L., Valero, V., Booser, D. J., Esserman, L., Lluch, A., et al. (2011). A Genomic Predictor of Response and Survival Following Taxane-anthracycline Chemotherapy for Invasive Breast Cancer, *Journal of the American Medical Association* **305**, 1873–1881.

Hilden, J. & Gerds, T. A. (2014). A Note on the Evaluation of Novel Biomarkers: Do Not Rely on Integrated Discrimination Improvement and Net Reclassification Index, *Statistics in Medicine* **33**, 3405–3414.

Huang, Y & Pepe, M. S. (2009a). A Parametric ROC Model Based Approach for Evaluating the Predictiveness of Continuous Markers in Case-Control Studies, *Biometrics* **65**, 1133–1144.

Huang, Y. & Pepe, M. S. (2009b). Semiparametric Methods for Evaluating Risk Prediction Markers in Case-Control Studies, *Biometrika* **96**, 991–997.

Huang, Y., Pepe, M. S. & Feng, Z. (2007). Evaluating the Predictiveness of a Continuous Marker, *Biometrics* **63**, 1181–1188.

Mayr, A, Binder, H, Gefeller, O & Schmid, M (2014). The Evolution of Boosting Algorithms. From Machine Learning to Statistical Modelling. *Methods of Information in Medicine* **53**, 419–27.

Moons, K. G. M., Royston, P., Vergouwe, Y., Grobbee, D. E. & Altman, D. G. (2009). Prognosis and Prognostic Research: What, Why, and How?, *BMJ: British Medical Journal* **338**, 1317–1320.

Moons, K. G. M., Groot, J. a. H. de, Linnet, K., Reitsma, J. B. & Bossuyt, P. M. M. (2012). Quantifying the Added Value of a Diagnostic Test or Marker, *Clinical Chemistry* **58**, 1408–1417.

Peek, N, Arts, D. G. T., Bosman, R. J., Voort, P. H. J. van der & Keizer, N. F. de (2007). External Validation of Prognostic Models for Critically Ill Patients Required Substantial Sample Sizes, *Journal of Clinical Epidemiology* **60**, 491–501.

Pencina, M., D'Agostino, R. & Vasan, R. (2010). Statistical Methods for Assessment of Added Usefulness of New Biomarkers, *Clinical Chemistry and Laboratory Medicine* **48**, 1703–1711.

Pepe, M. S. (2010). Rejoinder to Nancy Cook's Comment on "Measures to Summarize and Compare the Predictive Capacity of Markers", *The International Journal of Biostatistics* **6**, 16–18.

Pepe, M. S., Feng, Z & Gu, J. W. (2008a). Comments on 'Evaluating the Added Predictive Ability of a New Marker: From Area Under the ROC Curve to Reclassification and Beyond' by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929), *Statistics in Medicine* **27**, 173–181.

Pepe, M. S., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I. M., et al. (2008b). Integrating the Predictiveness of a Marker with its Performance as a Classifier, *American Journal of Epidemiology* **167**, 362–368.

Pepe, M. S., Gu, J. W. & Morris, D. E. (2010). The Potential of Genes and other Markers to Inform about Risk, *Cancer Epidemiology, Biomarkers & Prevention* **19**, 655–665.

Pepe, M. S., Kerr, K. F., Longton, G. & Wang, Z. (2013). Testing for Improvement in Prediction Model Performance, *Statistics in Medicine* **32**, 1467–1482.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Sachs, M. C. & Zhou, X.-H. (2013). Partial Summary Measures of the Predictiveness Curve, *Biometrical Journal* **55**, 589–602.

Shariat, S. F., Semjonow, A., Lilja, H., Savage, C., Vickers, A. J. & Bjartell, A. (2011). Tumor Markers in Prostate Cancer I: Blood-Based Markers, *Acta Oncologica* **50**, 61–75.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., et al. (2010). Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology* **21**, 128–138.

Steyerberg, E. W., Vedder, M. M., Leening, M. J. G., Postmus, D., D'Agostino, R. B., Van Calster, B., et al. (2014). Graphical Assessment of Incremental Value of Novel Markers in Prediction Models: From Statistical to Decision Analytical Perspectives, *Biometrical Journal*. DOI: 10.1002/bimj.201300260.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Series B* **58**, 267–288.