Boosting the Concordance Index for Survival Data – A Unified Framework To Derive and Evaluate Biomarker Combinations

Andreas Mayr[1]*, Matthias Schmid[2]

1 Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany, **2** Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany

Abstract

The development of molecular signatures for the prediction of time-to-event outcomes is a methodologically challenging task in bioinformatics and biostatistics. Although there are numerous approaches for the derivation of marker combinations and their evaluation, the underlying methodology often suffers from the problem that different optimization criteria are mixed during the feature selection, estimation and evaluation steps. This might result in marker combinations that are suboptimal regarding the evaluation criterion of interest. To address this issue, we propose a unified framework to derive and evaluate biomarker combinations. Our approach is based on the concordance index for time-to-event data, which is a non-parametric measure to quantify the discriminatory power of a prediction rule. Specifically, we propose a gradient boosting algorithm that results in linear biomarker combinations that are optimal with respect to a smoothed version of the concordance index. We investigate the performance of our algorithm in a large-scale simulation study and in two molecular data sets for the prediction of survival in breast cancer patients. Our numerical results show that the new approach is not only methodologically sound but can also lead to a higher discriminatory power than traditional approaches for the derivation of gene signatures.

Introduction

Recent technological developments in the fields of genomics and biomedical research have led to the discovery of large numbers of gene signatures for the prediction of clinical survival outcomes. In cancer research, for example, gene expression signatures are nowadays used to predict the time to occurrence of metastases [1,2] as well as the time to progression [3] and overall patient survival [4,5]. While the importance of molecular data in clinical and epidemiological research is expected to grow considerably in the next years [6–8], the detection of clinically useful gene signatures remains a challenging problem for bioinformaticians and biostatisticians, especially when the outcome is a survival time.

After normalization and data pre-processing, the development of a new gene signature usually comprises three methodological tasks:

Task 1: Select a subset of genes that is associated with the clinical outcome.

Task 2: Derive a marker signature by finding the "optimal" combination of the selected genes.

Task 3: Evaluate the prediction accuracy of the optimal combination using future or external data.

Task 1, the selection of a clinically relevant subset of genes, is often addressed by calculating scores to rank the univariate association between the survival outcome and each of the genes [8,9]. In a subsequent step, the genes with the strongest associations are selected to be included in the gene signature.

Task 2, the derivation of an optimal combination of the selected genes, is usually fulfilled by forming linear combinations of gene expression levels based on Cox regression. Due to multicollinearity problems and the high dimensionality of molecular data, a direct optimization of the Cox partial likelihood is often unfeasible [8]. Consequently, marker combinations are often derived by combining coefficients of univariate Cox regression models [10], or by applying regularized Cox regression techniques (such as the Lasso [11,12] or ridge-penalized regression [13,14]).

Task 3, the evaluation of prediction accuracy, is considered to be a challenging problem in survival analysis. This is because traditional performance measures for continuous outcomes (such as the mean squared error) are no longer applicable in the presence of censoring. In the literature, several approaches to address this problem exist (see, e.g., [15] for an overview). In this article, we focus on the *concordance index for time-to-event data* (*C*-index [16–18]), which has become a widely used measure of the performance of biomarkers in survival studies [19–22]. Briefly spoken, the *C*-index can be interpreted as the probability that a patient with a small survival time is associated with a high value of

a biomarker combination (and vice versa). Consequently, it measures the concordance between the rankings of the survival times and the biomarker values and therefore the ability of a biomarker to discriminate between patients with small survival times and patients with large survival times. This strategy is especially helpful if the aim is to subdivide patients into groups with good or poor prognosis (as applied in many articles in the medical literature, e.g., [10]). By definition, the C-index has the same scale as the classical area under the curve (AUC) in binary classification: While prediction rules based on random guessing yield $C = 0.5$, a perfectly discriminating biomarker combination leads to $C = 1$.

Interestingly, the derivation of new gene signatures for survival outcomes via Tasks 1–3 is often addressed by combining completely different methodological approaches and estimation techniques. For example, the estimation of biomarker combinations is usually based on Cox regression and is hence carried out via the optimization of a partial likelihood criterion. On the other hand, the resulting combinations are often evaluated by using the C-index [21,22] which has its roots in the receiver operating characteristics (ROC) methodology. This methodological inconsistency is also problematic from a practical point of view, as the marker combination that optimizes the partial log likelihood criterion is not necessarily the one that optimizes the C-index. In other words, if the C-index and therefore the discriminatory power is the evaluation criterion of interest, it may be suboptimal to use a likelihood-based criterion to optimize the marker combination. This issue is, of course, not only problematic in survival analysis but also in regression and classification. A theoretical discussion on the differences between performance measures for binary classification can, e.g., be found in [23].

To overcome the aforementioned inconsistencies, we propose a unified framework for survival analysis that is based on the same statistical methodology for gene selection (Task 1), derivation of an optimal biomarker combination (Task 2) *and* the evaluation of a new gene signature (Task 3). As will be demonstrated, all three tasks can be addressed by using the concordance index for time-to-event data as performance criterion. While the C-index has already been proposed for gene selection (Task 1) and the evaluation of prediction accuracy (Task 3) [9,15], the main contribution of this article is a new estimation technique that addresses the development of optimal combinations of genes (Task 2). To achieve this goal, we propose a method for finding gene combinations that is based on the gradient boosting framework [24]. As will be shown, it is possible to use boosting to derive prediction-optimized gene combinations via direct optimization of the C-index. Because this new approach allows for using the C-index to address all three tasks, the proposed method leads to a consistent framework for the derivation of gene signatures in biomarker studies where the C-index is the main performance criterion.

Methods

Notation

We first introduce some basic notation that will be used throughout the article. Let $T \in \mathbb{R}^+$ be the survival time of interest and $X = (x_1,...,x_p) \in \mathbb{R}^p$ a vector of predictor variables. In addition to the gene expression levels, X may contain the measurements of some clinical predictor variables. Denote the conditional survival function of T given X by $S(t|x) = P(T > t|X = x)$. The probability density and survival functions of T are denoted by $f(t)$ and $S(t)$, respectively. Further let $T_{\text{cens}} \in \mathbb{R}^+$ be a random censoring time and denote the

observed survival time by $\tilde{T} := \min(T, T_{\text{cens}})$. The random variable $\Delta := I(T \leq T_{\text{cens}})$ indicates whether $\tilde{T}$ is right-censored ($\Delta = 0$) or not ($\Delta = 1$).

A prediction rule for T will be formed as a linear combination

$$\eta := \beta_0 + \sum_{l=1}^{p} \beta_l \cdot x_l = X^{\mathrm{T}}\beta, \qquad (1)$$

where β is an unknown vector of coefficients. We generally assume that the estimation of $\hat{\beta}$ is based on an i.i.d. learning sample $\{(\tilde{T}_i^L, \Delta_i^L, X_i^L), i = 1,...,n\}$. In case of the Cox regression model, for example, η is related to T by the equation

$$S(t|x) = \exp(-\Lambda_0(t) \cdot \exp(\eta)), \qquad (2)$$

where $\Lambda_0(t)$ is the cumulative baseline hazard function. Because there is a one-to-one relationship between η and the expected survival time $E(T|X)$, the linear combination η can be used as a biomarker to predict the survival of individual patients.

Concordance index

Our proposed framework to derive and evaluate biomarker combinations is based on the *concordance index* ("C-index") which is a general discrimination measure for the evaluation of prediction models [16,17]. It can be applied to continuous, ordinal and dichotomous outcomes [25]. For time-to-event outcomes, the C-index is defined as

$$C := P(\eta_{j_1} > \eta_{j_2} \mid T_{j_1} < T_{j_2}), \qquad (3)$$

where T_{j_1}, T_{j_2} and η_{j_1}, η_{j_2} are the event times and the predicted marker values, respectively, of two observations in an i.i.d. test sample $\{(\tilde{T}_j, \Delta_j, X_j), j = 1,...,N\}$. By definition, the C-index for survival data measures whether large values of η are associated with short survival times T and vice versa. Moreover, it can be shown that the C-index is equivalent to the area under the time-dependent ROC curve, which is a measure of the discriminative ability of η at each time point under consideration (see [26], p. 95 for a formal proof).

During the last decades, the C-index has gained enormous popularity in biomedical research; for example, searching for the terms "concordance index" and "c-index" in PubMed [27] resulted in 1156 articles by the time of writing this article. Generally, a value of C close to 1 indicates that the marker η is close to a perfect discriminatory power, while a marker that does not perform better than chance results in a value of 0.5. For example, the famous Gail model [28] for the prediction of breast cancer is estimated to yield a value of $C = 0.67$ [29].

Being a flexible discrimination measure, the C-index is especially useful for selecting and ranking genes from a pre-processed set of high-dimensional gene expression data (Task 1 described in the Introduction). In other words, Task 1 can be addressed by computing the C-index (and hence the marginal discriminatory power) for each individual gene or biomarker, where only those genes with the highest C-index are incorporated into the yet-to-derive optimal combination (Task 2). Although there exist various other ways to rank genes and select the most influential ones, the C-index has been demonstrated to be especially advantageous for this task [9].

An estimator of the C-index for survival data is given by

$$\hat{C}_{\text{surv}} : =$$

$$\frac{\sum_{j<k} \mathrm{I}(\tilde{T}_j < \tilde{T}_k) \mathrm{I}(\hat{\eta}_j > \hat{\eta}_k)\Delta_j + \mathrm{I}(\tilde{T}_k < \tilde{T}_j)\mathrm{I}(\hat{\eta}_k > \hat{\eta}_j)\Delta_k}{\sum_{j<k} \mathrm{I}(\tilde{T}_j < \tilde{T}_k)\Delta_j + \mathrm{I}(\tilde{T}_k < \tilde{T}_j)\Delta_k} \quad (4)$$

with $j,k \in \{1,\ldots,n\}$ ("Harrell's C for survival data", [30]). Generally, $\hat{C}_{\text{surv}}$ is a consistent estimator of the C-index in situations where no censoring is present. However, because pairs of observations where the smaller observed survival time is censored are ignored in formula (4), $\hat{C}_{\text{surv}}$ is known to show a notable upward bias in the presence of censoring. This bias tends to be correlated with the censoring rate [15,30].

To overcome the censoring bias of $\hat{C}_{\text{surv}}$, Uno et al. [18] proposed a modified version of $\hat{C}_{\text{surv}}$, which is defined as

$$\hat{C}_{\text{Uno}} : = \frac{\sum_{j,k} (\hat{G}_n^L(\tilde{T}_j))^{-2} \mathrm{I}(\tilde{T}_j < \tilde{T}_k)\mathrm{I}(\hat{\eta}_j > \hat{\eta}_k)\Delta_j}{\sum_{j,k} (\hat{G}_n^L(\tilde{T}_j))^{-2}\mathrm{I}(\tilde{T}_j < \tilde{T}_k)\Delta_j} , \quad (5)$$

where $\hat{G}_n^L(t)$ denotes the Kaplan-Meier estimator of the unconditional survival function of T_{cens} (estimated from the learning data). In the following, we will assume that the censoring times are independent of X. Under this assumption, $\hat{C}_{\text{Uno}}$ is a consistent and asymptotically normal estimator of C (see [18], pp. 1113–1115). Consistency is ensured by applying inverse probability weighting (using the weights $\Delta_j/(\hat{G}_n^L(\tilde{T}_j)^2)$, which account for the inverse probability that an observation in the test data is censored [31]). Numerical studies suggest that $\hat{C}_{\text{Uno}}$ is remarkably robust against violations of the random censoring assumption [32].

Apart from the estimators $\hat{C}_{\text{surv}}$ and $\hat{C}_{\text{Uno}}$, there exist various other approaches to estimate the probability in (3) (see, e.g., [15] for an overview). Most of these approaches are based on the assumptions of a Cox proportional hazards model, so that they are not valid in case these assumptions are violated. Because $\hat{C}_{\text{Uno}}$ is model-free and because the consistency of $\hat{C}_{\text{Uno}}$ is guaranteed even in situations where censoring rates are high (in contrast to the estimator $\hat{C}_{\text{surv}}$), we will base our boosting method on $\hat{C}_{\text{Uno}}$.

Boosting the concordance index

The core of our proposed framework to address Tasks $1 - 3$ is the derivation of a prediction-optimized linear combination of genes that is optimal w.r.t. to the C-index for time-to-event data. Our approach will be based on a component-wise gradient boosting algorithm [24] that uses the C-index as optimization criterion.

Gradient boosting algorithms [33] are generally based on a loss function $\rho(T,\eta)$ that is assumed to be differentiable with respect to the predictor $\eta \equiv \eta(X)$. The aim is then to estimate the "optimal" prediction function

$$\eta^* : = \underset{\bar{\eta}}{\mathrm{argmin}}\, \mathbb{E}_{T,X}[\rho(T,\eta(X))] \quad (6)$$

by using gradient descent techniques. As the theoretical mean in (6) is usually unknown in practice, gradient boosting algorithms minimize the empirical risk $\mathcal{R} : = \sum_{i=1}^{n} \rho(t_i,\eta(x_i))$ over η instead.

When considering the C-index for survival data, the aim is to determine the optimal predictor η^* that maximizes the concordance measure $C = \mathrm{P}(\eta_i^* > \eta_k^* \,|\, T_i < T_k)$ – which is essentially the solution to Task 2. Hence a natural choice for the empirical risk function $\mathcal{R}$ would be the negative concordance index estimator

$$-\hat{C}_{\text{Uno}}(T,\eta) = - \frac{\sum_{i,k} \Delta_i (\hat{G}_n^L(\tilde{T}_i))^{-2} \mathrm{I}(\tilde{T}_i < \tilde{T}_k)\mathrm{I}(\eta_i > \eta_k)}{\sum_{i,k} \Delta_i (\hat{G}_n^L(\tilde{T}_i))^{-2}\mathrm{I}(\tilde{T}_i < \tilde{T}_k)} . \quad (7)$$

Setting $\mathcal{R} = -\hat{C}_{\text{Uno}}(T,\eta)$, however, is unfeasible because $\hat{C}_{\text{Uno}}(T,\eta)$ is not differentiable with respect to η_i and can therefore not be used in a gradient boosting algorithm. To solve this problem, we follow the approach of Ma and Huang [34] and approximate the indicator function in (7) by the sigmoid function $K(u) = 1/(1 + \exp(-u/\sigma))$. Here, σ is a tuning parameter that controls the smoothness of the approximation (details on the choice of σ will be given in the Numerical Results section). Replacing the indicator function in (7) by its smoothed version results in the smoothed empirical risk function

$$-\hat{C}_{\text{smooth}}(T,\eta) = - \sum_{i,k} w_{ik} \cdot \frac{1}{1 + \exp\left(\frac{\hat{\eta}_k - \hat{\eta}_i}{\sigma}\right)} \quad (8)$$

with weights

$$w_{ik} : = \frac{\Delta_i(\hat{G}_n^L(\tilde{T}_i))^{-2}\mathrm{I}(\tilde{T}_i < \tilde{T}_k)}{\sum_{i,k} \Delta_i(\hat{G}_n^L(\tilde{T}_i))^{-2}\mathrm{I}(\tilde{T}_i < \tilde{T}_k)} . \quad (9)$$

By definition, the smoothed empirical risk $-\hat{C}_{\text{smooth}}(T,\eta)$ is differentiable with respect to the predictor η_i. Its derivative is given by

$$-\frac{\partial \hat{C}_{\text{smooth}}(T,\eta)}{\partial \eta_i} = - \sum_k w_{ik} \frac{-\exp\left(\frac{\hat{\eta}_k - \hat{\eta}_i}{\sigma}\right)}{\sigma\left(1 + \exp\left(\frac{\hat{\eta}_k - \hat{\eta}_i}{\sigma}\right)\right)} . \quad (10)$$

In the proposed gradient boosting algorithm, the derivative in (10) is iteratively fitted to a set of base-learners. Typically, an individual base-learner (simple regression tool, e.g., a tree or a simple linear model) is specified for each marker. To ensure that the estimate of the optimal predictor η^* is a *linear* combination of the components of X, we will apply simple linear models as base-learners (cf. [35]). In other words, each base-learner is a simple linear model with one component of X as input variable. Consequently, there are p base-learners, which will be denoted by b_l, $l = 1,\ldots,p$. Each base-learner refers to one component of X and therefore to one marker (or gene).

The component-wise gradient boosting algorithm for the optimization of the smoothed C-index is then given as follows:

1. **Initialize** the estimate of the marker combination $\hat{\boldsymbol{\eta}}^{[0]}$ with offset values. For example, set $\hat{\boldsymbol{\eta}}^{[0]} = \mathbf{0}$, leading to $\hat{\beta}_l^{[0]} = 0$ for all components $l = 1,\ldots,p$. Choose a sufficiently large maximum number of iterations m_{stop} and set the iteration counter m to 1.

2. **Compute** the negative gradient vector by using formula (10) and evaluate it at the marker combination $\hat{\boldsymbol{\eta}}^{[m-1]}$ of the previous iteration:

$$U^{[m]} = \left(U_i^{[m]}\right)_{i=1,\ldots,n} := \left(\frac{\partial \hat{C}_{\text{smooth}}(T,\hat{\boldsymbol{\eta}}^{[m-1]})}{\partial \eta_i}\right)_{i=1,\ldots,n}.$$

3. **Fit** the negative gradient vector $U^{[m]}$ separately to each of the components of X via the base-learners $b_l(\cdot)$:

$$U^{[m]} \xrightarrow{\text{fitted by}} \hat{b}_l^{[m]}(x_l) \quad \text{for } l=1,\ldots,p.$$

4. **Select** the component l^* that best fits the negative gradient vector according to the least squares criterion, i.e., select the base-learner b_{l^*} defined by

$$l^* = \underset{1 \le l \le p}{\text{argmin}} \sum_{i=1}^n \left(U_i^{[m]} - \hat{b}_l^{[m]}(x_l)\right)^2 .$$

5. **Update** the marker combination $\hat{\boldsymbol{\eta}}$ for this component:

$$\hat{\boldsymbol{\eta}}^{[m]} \leftarrow \hat{\boldsymbol{\eta}}^{[m-1]} + \text{sl}\cdot\hat{b}_{l^*}^{[m]}(x_{l^*}),$$

where sl is a small step length $(0 < \text{sl} \ll 1)$. For example, if $\text{sl}=0.1$, only 10% of the fit of the base-learner is added to the current marker. This procedure shrinks the effect estimates towards zero, effectively increasing the numerical stability of the update step [23,25].

As only the base learner $\hat{b}_{l^*}$ was selected, only the effect of component l^* is updated $(\hat{\beta}_{l^*}^{[m]} = \hat{\beta}_{l^*}^{[m-1]} + \text{sl}\cdot\hat{b}_{l^*}^{[m]}(x_{l^*}))$ while all other effects stay constant $(\hat{\beta}_l^{[m]} = \hat{\beta}_l^{[m-1]}$ for $l \ne l^*)$.

6. **Stop** if $m=m_{\text{stop}}$. Else increase m by one and go back to step (2).

By construction, the proposed boosting algorithm automatically estimates the optimal linear biomarker combination that maximizes the smoothed C-index. The principle behind the proposed algorithm is to minimize the empirical risk $\mathcal{R} = -\hat{C}_{\text{smooth}}(T,\eta)$ by using gradient descent in function space, where the function space is spanned by the base-learners b_l, $l=1,\ldots,p$. In other words, the algorithm iteratively descents the empirical risk by updating $\hat{\boldsymbol{\eta}}^{[m]}$ via the best fitting base-learner. Because the base-learners are simple linear models (each containing only one possible biomarker as predictor variable) and because the update in step (5) of the algorithm is additive, the final solution $\hat{\boldsymbol{\eta}}^{[m_{\text{stop}}]}$ effectively becomes a linear combination of these markers.

The two main tuning parameters of gradient boosting algorithms are the stopping iteration m_{stop} and the step length sl. In the literature it has been argued that the choice of the step length is of minor importance for the performance of boosting algorithms [36]. Generally, a larger step length leads to faster convergence of the algorithm. However, it also increases the risk of overshooting near the minimum of $\mathcal{R}$. In the following sections we will use a fixed step-length of $\text{sl}=0.1$, which is a common recommendation in the literature on gradient boosting and which is also the default value in the R package mboost [37].

The stopping iteration m_{stop} is considered to be the most important tuning parameter of boosting algorithms [38]. The optimal value of m_{stop} is usually determined by using cross-validation techniques [24]. Small values of m_{stop} reduce the complexity of the resulting linear combination $\hat{\boldsymbol{\eta}}^{[m_{\text{stop}}]}$ and avoid overfitting via shrinking the effect estimates. In case of boosting the C-index, however, overfitting is less problematic as the predictive performance of η is not related to the actual size of the coefficients but to the concordance of the *rankings* between marker values and the observed survival times. As a result, the stopping iteration m_{stop} in this specific case is less relevant and can be also specified by a fixed large value (e.g., $m_{\text{stop}}=50000$).

Regarding the boosting algorithm for the smoothed C-index, an additional tuning parameter is given by the smoothing parameter σ. While too large values of σ will lead to a poor approximation of the indicator functions in (7), too small values of σ might overfit the data (and might therefore result in a decreased prediction accuracy). Details on how to best choose the value of σ will be given in the Numerical Results section.

The boosting algorithm presented above is implemented in the add-on software package mboost of the open source statistical programming environment R [39]. The specification of the new Cindex() family and a short description of how to apply the algorithm in practice are given in Text S1.

Evaluation

After having applied the C-index to select the most influential genes (Task 1), and after having used the proposed boosting algorithm to combine the selected genes (Task 2), a final challenge is to evaluate the prediction accuracy of the resulting gene combination (Task 3). Since the C-index was used for Tasks 1 and 2, it is also a natural criterion to evaluate the derived marker combination in Task 3. As argued before, it is advantageous from both a methodological perspective as well as from a practical one to use the same criterion for estimation and evaluation of a biomarker combination.

To avoid over-optimistic estimates of prediction accuracy, it is crucial to use different observations for the optimization and evaluation of the marker signature [25,40]. This can be achieved either by using two completely different data sets (*external evaluation*) or by splitting one data-set into a learning sample $\{(\tilde{T}_i^L, \Delta_i^L, X_i^L), i=1,\ldots,n\}$ and a test sample $\{(\tilde{T}_j, \Delta_j, X_j), j=1,\ldots,N\}$. The learning sample is used to optimize the marker combination while the "external" test sample serves only for evaluation purposes. A more elaborate strategy is *subsampling* (such as five-fold or ten-fold cross-validation), which is based on multiple random splits of the data. In our numerical analysis, we will use subsampling techniques in combination with stratification to divide the underlying data sets into learning and test samples (for details, see the next section).

When it comes to the C-index, two additional points have to be taken into consideration: First, as the task is to obtain the most precise estimation for the discriminatory power, it is no longer necessary to use the smoothed version $\hat{C}_{\text{smooth}}$ (which was included for numerical reasons in the boosting algorithm) for evaluation. Consequently, we propose to apply the original estimator $\hat{C}_{\text{Uno}}$ for evaluating biomarker combinations in Task 3. Second, when applying the estimator $\hat{C}_{\text{Uno}}$ to the observations in a test sample, a natural question is how to calculate the Kaplan-Meier estimator $\hat{G}_n^L(t)$ of the unconditional survival function of T_{cens}. In principle, there are three possibilities for the calculation of $\hat{G}_n^L(t)$: The Kaplan-Meier estimator can be computed from

either the test or from the training data, or, alternatively, from the combined data set containing all observations in the learning and test samples. Following the principle that all estimation steps should be carried out prior to Task 3, we will base computation of the Kaplan-Meier estimator on the learning data.

Numerical results

Simulation study. We first investigated the performance of our approach based on simulated data. The aim of our simulation study was:

(i)   To analyze if the proposed framework is able to select a small amount of informative markers from a much larger set of candidate variables.

(ii)   To check if gradient boosting is able to derive the optimal combination η of the selected markers, and to compare its performance to competing Cox-based estimation schemes.

(iii)   To investigate the effect of the smoothing parameter σ that controls the smoothness inside the sigmoid function, as well as potential effects of the sample size and the censoring rate on the performance of our approach.

The simulated survival times are generated via a log-logistic distribution for accelerated failure time (AFT) models [41]. They are based on the model equation $\log(T) = \mu + \phi W$, where T is the survival time, μ and ϕ are location and scale parameters, and W is a noise variable, following a standard logistic distribution. As a result, the density function for realizations t_i can be written as

$$f_{\mathrm{dens}}(t_i|\mu_i,\phi_i) = \frac{\exp((t_i - \mu_i)/\phi_i)}{\phi_i(1 + \exp((t_i - \mu_i)/\phi_i))^2} \quad (11)$$

with $\mathbb{E}(T) = \mu$ and $\mathrm{Var}(T) = \frac{\pi^2}{3\phi^2}$. The $p = 1000$ possible markers $X_1, ..., X_{1000}$ were drawn from a multivariate normal distribution with pairwise correlation ($\rho = 0.5$). The *true* underlying combinations of the predictors were given by

$$\mu_i = \eta_\mu = 1.5 + 1.5x_1 + x_2 - x_3 - 1.5x_4 ,$$

$$\log(\phi_i) = \eta_\phi = -1 + 2x_1 - 2x_2 + x_3 - x_4 .$$

Note that only four out of 1000 possible markers have an actual effect on the survival time. Furthermore, those four markers do not only contribute to the location parameter μ but also to the scale parameter ϕ (cf. [42]) – a setting where standard survival analysis clearly becomes problematic. Additionally to the survival times T, we generated an independent sample of censoring times T_{cens} and defined the observed survival time by $\tilde{T} := \min(T, T_{\mathrm{cens}})$ leading to independent censoring of 50% of the observations.

In order to answer the above questions, we investigated the performance of our framework in all three tasks that are necessary to develop new gene signatures in practice (Tasks 1–3 described in the Introduction). To avoid over-optimism and biased results, we simulated separate data sets for the different tasks. In $B = 100$ simulation runs, we first simulated a data set $\{(\tilde{T}_i, \Delta_i, X_i), i = 1, ..., 1000\}$ with 1000 observations to pre-select the most influential predictors based on the C-index (Task 1). The optimal combination η of those predictors was later estimated

(Task 2) by our boosting algorithm based on smaller training samples $\{(\tilde{T}_i^L, \Delta_i^L, X_i^L), i = 1, ..., n\}$ of size n. For the final external evaluation of the prediction accuracy (Task 3) we simulated a separate test data set $\{(\tilde{T}_j, \Delta_j, X_j), j = 1, ..., N\}$ with $N = 1000$.

For Task 1, we first pre-selected a subset of p^* predictors from the $p = 1000$ available markers. We ranked the predictors based on their individual values of $\hat{C}_{\mathrm{Uno}}$ and included only the $p^* = \{5, 10, 30\}$ best-performing markers in the boosting algorithm. The results suggest that the C-index is clearly able to identify markers that are truly related to the outcome: Although all predictors had a relatively high pairwise correlation ($\rho = 0.5$), the four informative markers had a selection probability of 98.5% for $p^* = 5$ (99% for $p^* = 10$ and 99.5% for $p^* = 30$).

To find the optimal combination η of the pre-selected markers (Task 2), we applied the proposed boosting approach on training samples with size $n = 100$. The resulting coefficients for $p^* = 5$ and smoothing parameter $\sigma = 0.1$ are presented in Figure 1. The boosting algorithm seems to be able to derive the optimal combination of the pre-selected markers, as the structure displayed by the coefficients is essentially the same as the one of the underlying *true* combination η_μ. The discriminatory power of the resulting biomarker does not depend on the absolute size of the coefficients: As the C-index is based solely on the concordance between biomarker and survival time, what matters in practice is the *relative* size of the coefficients. As can be seen from Figure 1, the estimated positive effect for x_1 is larger than the one for x_2. On the other hand, the negative effect of x_4 is correctly estimated to be more pronounced than the the one of x_3. The coefficient of the falsely selected marker is on average close to zero.

In a third step, we evaluated the performance of the resulting optimized marker combinations (Task 3) on separate test samples. The resulting estimates $\hat{C}_{\mathrm{Uno}}$ for different simulation settings are presented in Table 1. The highest discriminatory power (median $\hat{C}_{\mathrm{Uno}} = 0.763$, range = 0.559–0.819) can be observed for $p^* = 5$, which is closest to the true number of informative markers. We compared the performance of our proposed algorithm to penalized Cox regression approaches such as Cox-Lasso [11,12] and Cox regression with ridge-penalization [13,14] – see Figure 2. The proposed boosting approach clearly outperforms the competing estimation schemes, supporting our view that applying traditional Cox regression might be suboptimal if the discriminatory power is the performance criterion of interest. We additionally computed the optimal C-index resulting from the *true* marker combination η_μ with known coefficients. The values of the true C-index on the test samples are on average only slightly better than the ones of boosting the concordance index (median $\hat{C}_{\mathrm{Uno}} = 0.778$ – see Table 1).

To evaluate the possible effects of different sample sizes and censoring rates we modified the mean censoring time leading to approximate censoring rates of 30% and 70% and generated training samples of size $n = \{50, 200, 500\}$. Results are included in Table 1. As expected, the C-index resulting from our framework increases as censoring rates become small (median $\hat{C}_{\mathrm{Uno}} = 0.820$, range = 0.736–0.858) and decreases in settings with a large proportion of censored observations (median $\hat{C}_{\mathrm{Uno}} = 0.668$, range = 0.421–0.776). However, the same effect can be observed for the true C-index resulting from the true marker combination η_μ (30% censoring $\hat{C}_{\mathrm{Uno}} = 0.830$, 70% censoring $\hat{C}_{\mathrm{Uno}} = 0.690$). For larger training samples, the variance of the coefficient estimates decreases (see Figure 1). As a result, the discriminatory power resulting from our boosting algorithm improves (for $n = 500$, median $\hat{C}_{\mathrm{Uno}} = 0.778$, range = 0.614–0.818) and gets nearly as good as
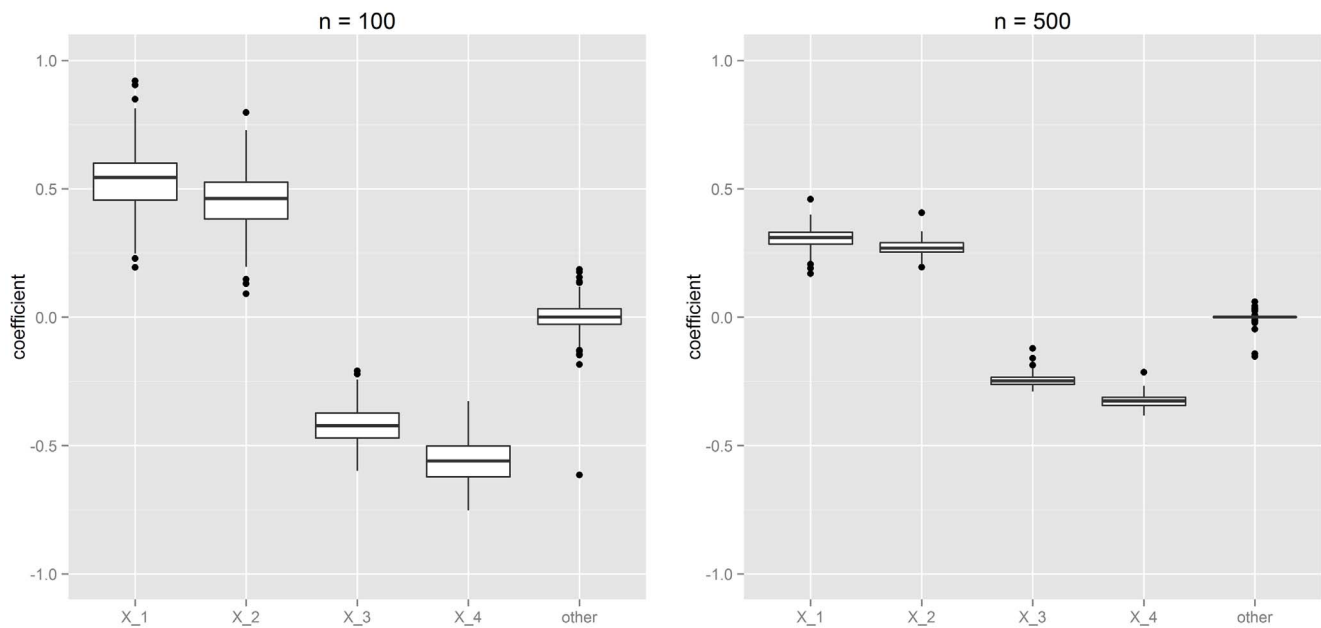
Figure 1. Coefficient estimates for $p^* = 5$ pre-selected markers obtained from 100 simulation runs. The marker combinations were optimized via gradient boosting based on training samples of size $n = 100$ (left) and $n = 500$ (right). Boxplots represent the empirical distribution of the resulting coefficients. Only markers X_1 to X_4 had an actual effect on the survival time.
doi:10.1371/journal.pone.0084483.g001

the true C-index ($\hat{C}_{\mathrm{Uno}} = 0.781$). This finding further confirms the ability of our approach to find the most optimal marker combination possible – see Figure 2. Note that also the true C-index differs slightly between the different sample sizes, as the training sample enters in $\hat{C}_{\mathrm{Uno}}$ via the Kaplan-Meier estimator $\hat{G}_n^L(t)$.

To investigate the effect of the smoothing parameter inside the sigmoid function, we additionally applied our boosting procedure for every simulation setting with different values of σ. The resulting estimates $\hat{C}_{\mathrm{Uno}}$ are presented in Table 2. Compared to the effects of the sample size or the number of pre-selected markers p^*, the smoothing parameter σ only seems to have a minor effect on the performance of our algorithm. In light of these empirical results, we recommend to apply a fixed small value (e.g., $\sigma = 0.1$,

which is also the default value in the Cindex() family for the mboost package [37] – see Text S1).

For both approaches to fit penalized Cox regression (Cox lasso, Cox ridge), we applied the R add-on package penalized [43]. In order to evaluate $\hat{C}_{\mathrm{Uno}}$, we used the UnoC() function implemented in the survAUC package [44].

Applications to predict the time to distant metastases

In the next step, we further analyzed the performance of our gradient boosting algorithm in two applications to estimate and evaluate the optimal combination of pre-selected biomarkers. All markers are used to predict the time to distant metastases in breast cancer patients. As in the simulation study, we compared the results of our proposed algorithm to Cox regression with lasso and ridge penalization. Additionally, we considered four competing

Table 1. Results of the simulation study.

setting			method			
n	p^*	*cens.*	C-index boosting	Cox lasso	Cox ridge	*true* C-index
100	5	50%	0.764 (0.04)	0.731 (0.06)	0.739 (0.04)	0.779
100	10	50%	0.746 (0.06)	0.709 (0.08)	0.707 (0.06)	0.779
100	30	50%	0.689 (0.07)	0.673 (0.11)	0.637 (0.07)	0.779
100	5	30%	0.820 (0.02)	0.774 (0.04)	0.724 (0.04)	0.830
100	5	70%	0.668 (0.10)	0.628 (0.10)	0.593 (0.11)	0.690
50	5	50%	0.741 (0.07)	0.662 (0.18)	0.722 (0.09)	0.772
200	5	50%	0.774 (0.02)	0.748 (0.04)	0.752 (0.04)	0.782
500	5	50%	0.778 (0.03)	0.759 (0.03)	0.760 (0.02)	0.781

Comparison of the discriminatory power resulting from boosting the C-index and competing approaches. Numbers refer to the median value and interquartile range (in parentheses) of the final $\hat{C}_{\mathrm{Uno}}$ on 100 simulation runs. The *true* C-index refers to the discriminatory power resulting from the true combination of predictors with known coefficients. The amount of pre-selected genes is denoted as p^*, n is the size of the training samples and *cens.* refers to the censoring rate.
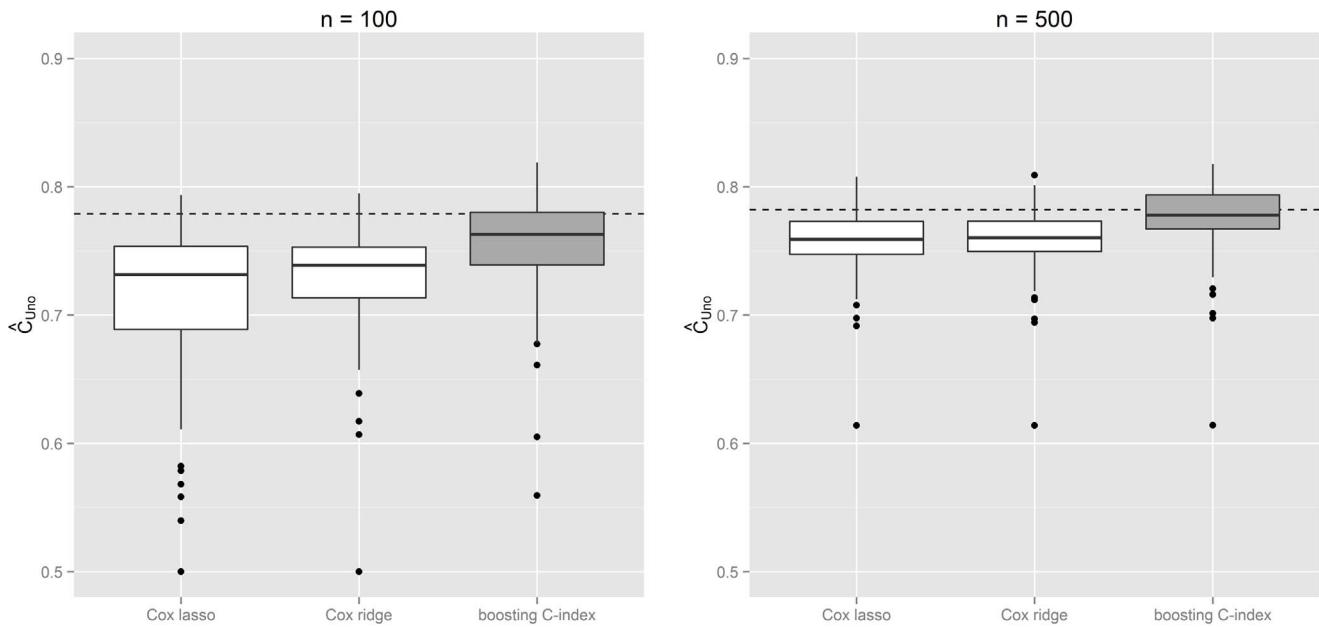doi:10.1371/journal.pone.0084483.t001

Figure 2. Simulation results for the discriminatory power obtained via the proposed C-index boosting approach and competing Cox-based estimation schemes. The marker combinations were optimized via the different approaches based on training samples of size $n = 100$ (left) and $n = 500$ (right). Boxplots represent the empirical distribution of the resulting $\hat{C}_{\mathrm{Uno}}$ on corresponding test samples. The dotted line corresponds to the discriminatory power resulting from the *true* combination of predictors with known coefficients.
doi:10.1371/journal.pone.0084483.g002

boosting approaches for survival analysis, which do not directly optimize the C-index. The first is classical Cox regression, estimated via gradient boosting, while the other three are parametric accelerated failure-time (AFT) models assuming a Weibull, log-normal or log-logistic distribution [45,46]. For all boosting approaches (Weibull AFT boosting, loglog-AFT boosting and Cox boosting) we used the corresponding pre-implemented functions of the mboost package. To ensure comparability, we used the same linear base-learners as described above for all boosting approaches.

To ensure that the combined predictor did not only work on the data it was derived from but also on "external" validation data, we carried out a subsampling procedure for both data sets to generate $B = 100$ different learning samples $\{(\tilde{T}_i^L, \Delta_i^L, X_i^L), i = 1, \ldots, n\}$ and test samples $\{(\tilde{T}_j, \Delta_j, X_j), j = 1, \ldots, N\}$, respectively. Consequently, we randomly split the corresponding data sets to use $2/3$ of the observations as learning sample in order to optimize the biomarker combination $\hat{\eta}$. To ensure an equal distribution of patients with and without an observed development of distant metastases we applied stratified sampling. Correspondingly, the $1/3$ of the observations not included in the learning sample were used to evaluate the resulting predictions via the C-index. As a result, for every data set and every method we computed 100 different combinations η and 100 corresponding values of $\hat{C}_{\mathrm{Uno}}$.

Breast cancer data by Desmedt et al.

Desmedt et al. [1] collected a data set of 196 node-negative breast cancer patients to validate a 76-gene expression signature developed by Wang et al. [10]. The signature, which is based on Affymetrix microarrays, was developed separately for estrogen-receptor (ER) positive patients (60 genes) and ER-negative patients (16 genes). In addition to the expression levels of the 76 genes, four clinical predictor variables were considered (tumor size, estrogen receptor (ER) status, grade of the tumor and patient age). The data

are publicly available on GEO (http://www.ncbi.nlm.nih.gov/geo, accession number GSE 7390).

Similar to Wang et al. [10], we used the time from diagnosis to distant metastases as primary outcome and considered the 76 genes together with the four clinical predictors. Observed metastasis-free survival ranged from 125 days to 3652 days, with 79.08% of the survival times being censored.

The main results of our analysis are presented in Figure 3. As expected, the unified framework to estimate and evaluate the optimal marker signature based on the C-index is not only methodologically more consistent than the Cox and AFT approaches, but also leads to to marker signatures that show a higher discriminatory power on external or future data (median $\hat{C}_{\mathrm{Uno}} = 0.736$, range = 0.467–0.854). As discussed in the methodological section, it is crucial to evaluate the discriminatory power on external data: the estimated C-index on the training sample was more than 35% higher (median $\hat{C}_{\mathrm{Uno}} = 0.986$) and hence extremely over-optimistic [25,40].

Considering the interpretation of the resulting coefficient estimates for the clinical predictors, it is crucial to note that boosting methods for the C-index and the AFT models yield biomarker combinations η^* where larger values indicate *longer* predicted survival times. On the other hand, classical Cox regression models rely on the hazard; higher values are hence associated with *smaller* survival times. If this is taken into account, results from the different approaches were in fact very similar. Both age of the patients and size of the tumor had a negative effect on the time to recurrence for all seven approaches. The same holds true for the tumor grade *poor differentiation* which resulted in a negative effect compared to *good differentiation* and *intermediate differentiation*. A positive ER status, on the other hand, was associated with a larger metastasis-free survival in all approaches. Regarding the coefficients of the 76 genes, results from our approach to boost the C-index were highly correlated to the ones

Table 2. Simulation results for different values of the smoothing parameter.

setting			smoothing parameter				
n	p^*	cens.	$\sigma=0.5$	$\sigma=0.25$	$\sigma=0.1$	$\sigma=0.075$	$\sigma=0.05$
100	5	50%	0.738 (0.06)	0.757 (0.05)	0.764 (0.04)	0.763 (0.04)	0.762 (0.04)
100	10	50%	0.728 (0.06)	0.744 (0.06)	0.746 (0.06)	0.746 (0.06)	0.741 (0.05)
100	30	50%	0.700 (0.06)	0.702 (0.07)	0.689 (0.07)	0.683 (0.07)	0.666 (0.07)
100	5	30%	0.802 (0.03)	0.815 (0.02)	0.820 (0.02)	0.821 (0.02)	0.822 (0.02)
100	5	70%	0.665 (0.10)	0.667 (0.10)	0.668 (0.10)	0.665 (0.10)	0.661 (0.10)
50	5	50%	0.719 (0.07)	0.737 (0.07)	0.741 (0.07)	0.740 (0.06)	0.725 (0.06)
200	5	50%	0.743 (0.05)	0.768 (0.03)	0.774 (0.02)	0.775 (0.02)	0.778 (0.02)
500	5	50%	0.723 (0.05)	0.769 (0.02)	0.778 (0.03)	0.779 (0.03)	0.781 (0.03)

Comparison of the discriminatory power resulting from the gradient boosting approach when applying different values of the smoothing parameter σ. Numbers refer to the median value and interquartile range (in parentheses) of the final $\hat{C}_{Uno}$ on 100 simulation runs. The amount of pre-selected genes is denoted as p^*, n is the size of the training samples and cens. refers to the censoring rate. We recommend to use the value $\sigma=0.1$, which is also the default value of the new Cindex family for the R add-on package mboost.
doi:10.1371/journal.pone.0084483.t002

of the other four boosting approaches (which rely on the same base-learners) – absolute correlation coefficients computed from the 100 subsamples ranged from 0.77 to 0.99. Also coefficients resulting from the popular ridge-penalized Cox regression were highly correlated with the ones from our approach – absolute correlation coefficients ranged from 0.47 to 0.84.

Breast cancer data by van de Vijver et al.

The second data set consists of 144 lymph node positive breast cancer patients that was collected by the Netherlands Cancer Institute [2]. The data set, which is publicly available as part of the R add-on package penalized [43], was used by van de Vijver et al. [2] to validate a 70-gene signature for metastasis-free survival after surgery developed by van't Veer et al. [47]. In addition to the expression levels of the 70 genes, the data set contains five clinical

predictor variables (tumor diameter, number of affected lymph nodes, ER status, grade of the tumor and patient age). Observed metastasis-free survival times ranged from 0.055 months to 17.660 months, with 67% of the survival times being censored.

Resulting values of the C-index of the new approach and the six considered competitors are presented in Figure 3. The improvement from applying the proposed unified framework compared to boosting the Cox proportional hazard model or applying ridge-penalized Cox regression was much less pronounced than in the previous data set. However, on average, boosting the C-index still led to the best combination of markers regarding the discriminatory power (median $\hat{C}_{Uno}=0.662$, range $= 0.257$–0.836). Interestingly, as in the previous data set, the lasso penalized Cox regression was clearly outperformed by the ridge-penalized competitor (which has been suggested for this specific data set
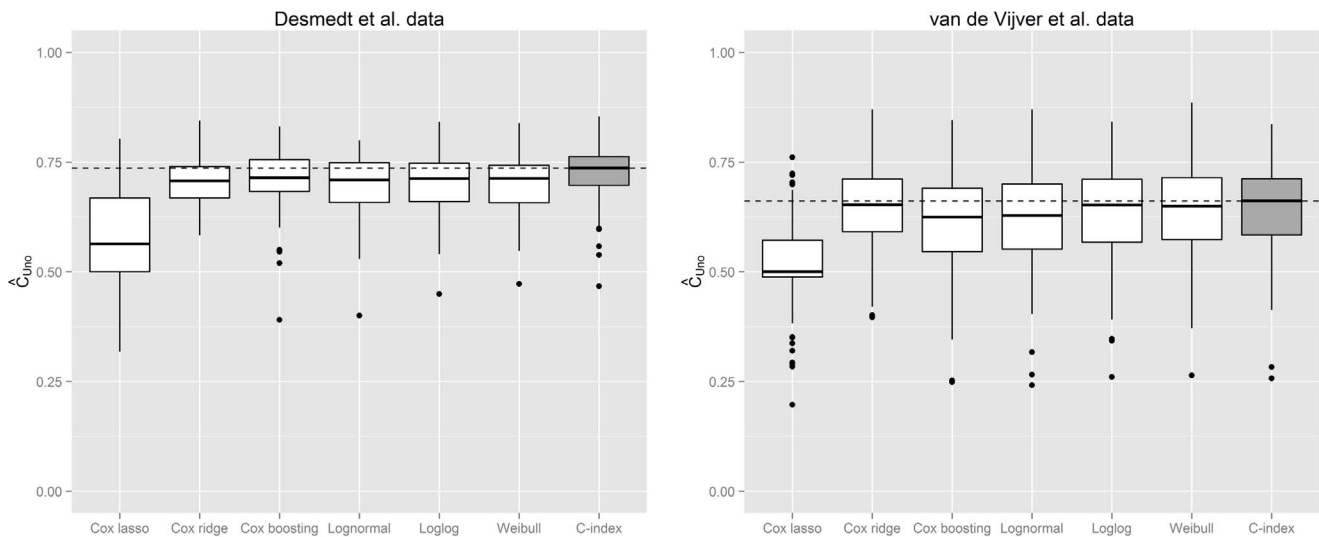

Figure 3. Comparing the discriminatory power of biomarker combinations to predict the time to distant metastases resulting from the proposed C-index boosting approach with competing estimation schemes. The plot on the left refers to the Desmedt et al. data, whereas the plot on the right presents results from the van de Vijver et al. data. All biomarker combinations were optimized via the corresponding algorithms based on the same 100 learning samples. Boxplots represent the empirical distribution of the resulting $\hat{C}_{Uno}$ on corresponding test samples. The dotted line corresponds to the median C-index resulting from the new approach.
doi:10.1371/journal.pone.0084483.g003

by van Houwelingen et al. [48]). Furthermore, the ridge-penalized approach performed at least as good as the considered boosting approaches (except the new approach to boost the C-index). As in the previous data set, we again additionally evaluated the C-index on the training sample in order to assess the resulting over-optimism. As expected, the estimated C-index on the training sample was extremely biased (median $\hat{C}_{\mathrm{Uno}} = 0.973$).

The resulting coefficients for the clinical predictors were again comparable for the seven different approaches. A positive ER status was associated with a larger metastasis-free survival for all seven approaches, the same also holds true for the age of the patient. On the other hand, the size of the tumor, the number of affected lymph nodes and a poor tumor grade resulted for all approaches in a negative effect on the survival time. Regarding the coefficients of the 79 genes, the highest correlation could again be observed for the boosting algorithms: Absolute correlation coefficients obtained from the 100 subsamples ranged from 0.64 to 0.95. Correlation between coefficients resulting from our approach to boost the C-index and the ones from ridge-penalized Cox regression was slightly lower, it ranged from 0.30 to 0.82.

Discussion

In this article we have proposed a framework for the development of survival prediction rules that is based on the concordance index for time-to-event data (C-index). As the C-index is an easy-to-interpret measure of the accuracy of survival predictions (based on clinical or molecular data), it has become an important tool in medical decision making. Generally, the focus of the C-index is on measuring the "discriminatory power" of a prediction rule: It quantifies how well the rankings of the survival times and the values of a biomarker (or marker combinations) in a sample agree. In particular, the C-index is methodologically different from measures that evaluate how well a prediction rule is "calibrated" (i.e., from measures that quantify "how closely the predicted probabilities agree numerically with the actual outcomes" [49]). Specifically, prediction rules that are well calibrated do not necessarily have a high discriminatory power (and vice versa).

While several authors have proposed the use of the C-index for feature selection [9] and the evaluation of molecular signatures [21,22], the main contribution of this paper is a new approach for the *derivation of marker combinations* that is based directly on the C-index. Consequently, when using the proposed method, it is no longer necessary to rely on traditional methods such as Cox regression – which focus on the derivation of well-calibrated prediction rules instead of well-discriminating prediction rules and may therefore be suboptimal when the optimization of the discriminatory power is of main interest.

Conceptually, our approach is in direct line with recent articles by Ma and Huang [34], Wang and Chang [50] and Schmid et al. [51] who developed a set of algorithms for the optimization of discrimination measures for binary outcomes (such as the area under the curve (AUC) and the partial area under the curve and (PAUC)). Because the C-index is in fact a summary measure of a correspondingly defined AUC measure for time-to-event data [26], our optimization technique relies on similar methodological concepts, such as the application of boosting methods and the use of smoothed indicator functions.

A possible future extension of our approach might be to include the task of selecting the most influential genes in the proposed boosting algorithm. While our simulation study and the breast-cancer examples were based on the pre-selection of genes, the proposed boosting method could also be applied directly to high-dimensional molecular data, so that Tasks 1 and 2 are effectively combined. This can be accomplished by optimizing the stopping iteration so that only a (low-dimensional) subset of the candidate genes is incorporated in the resulting marker combination ("early stopping", cf. [38]). Further research is warranted on the issues of early stopping and automated feature selection in the case of boosting the concordance index for survival data.

The results of our empirical analysis suggest that the new approach is competitive with state-of-the-art methods for the derivation of marker combinations. As demonstrated in the Numerical Results section, the resulting marker combinations are not only easy to compute and have a meaningful interpretation but can also lead to a higher discriminatory power of the resulting gene signatures.

Supporting Information

Text S1 This document provides technical details on boosting the concordance index. It contains the implementation of the new `Cindex()` family and a short description of the necessary `R` Code to apply the algorithm in practice.
(PDF)

Acknowledgments

Author Contributions

Conceived and designed the experiments: AM MS. Analyzed the data: AM MS. Wrote the paper: AM MS.

References

1. Desmedt C, Piette F, Loi S, Wang Y, Lallemand F, et al. (2007) Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independentvalidation series. Clinical Cancer Research 13: 3207–3214.

2. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AAM, et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine 347: 1999–2009.

3. Kok M, Linn SC, Laar RKV, Jansen MPHM, van den Berg TM, et al. (2009) Comparison of gene expression profiles predicting progression in breast cancer patients treated with tamoxifen. Breast Cancer Research and Treatment 13: 275–283.

4. Li H, Gui J (2004) Partial Cox regression analysis for high-dimensional microarray gene expression data. Bioinformatics 20: 208–215.

5. Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, et al. (2004) Gene expression signature of fibroblast serum response predicts human cancer progression: Similarities between tumors and wounds. PLoS Biology 2.

6. Gilad S, Meiri E, Yogev Y, Benjamin S, Lebanony D, et al. (2008) Serum micrornas are promising novel biomarkers. PLoS ONE 3: e3148.

7. Wistuba II, Gelovani JG, Jacoby JJ, Davis SE, Herbst RS (2011) Methodological and practical challenges for personalized cancer therapies. Nature Reviews Clinical Oncology 8: 135–141.

8. Witten DM, Tibshirani R (2010) Survival analysis with high-dimensional covariates. Statistical Methods in Medical Research 19: 29–51.

9. Ma S, Song X (2011) Ranking prognosis markers in cancer genomic studies. Brie_ngs in Bioinformatics 12: 33–40.

10. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365: 671-679.

11. Tibshirani R (1997) The lasso method for variable selection in the Cox model. Statistics in Medicine 16: 385–395.

12. Goeman JJ (2010) L_1 penalized estimation in the Cox proportional hazards model. Biometrical Journal 52: 70–84.

13. Li H, Luan Y (2002) Kernel cox regression models for linking gene expression profiles to censored survival data. In: Pacific Symposium on Biocomputing. volume 8, p. 65.
14. Gui J, Li H (2005) Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics 21: 3001–3008.
15. Schmid M, Potapov S (2012) A comparison of estimators to evaluate the discriminatory power of time-to-event models. Statistics in Medicine 31: 2588–2609.
16. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA (1982) Evaluating the yield of medical tests. Journal of the American Medical Association 247: 2543–2546.
17. Harrell FE, Lee KL, Califf, R M (1984) Regression modeling strategies for improved prognostic prediction. Statistics in Medicine 3: 143–152.
18. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ (2011) On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine 30: 1105–1117.
19. Kaderali L, Zander T, Faigle U,Wolf J, Schultze JL, et al. (2006) CASPAR: A hierarchical bayesian approach to predict survival times in cancer from gene expression data. Bioinformatics 22: 1495–1502.
20. Pepe MS, Zheng Y, Jin Y (2008) Evaluating the ROC performance of markers for future events. Lifetime Data Analysis 14: 86–113.
21. Zhang H, Xia W, Lu X, Sun R, Wang L, et al. (2013) A novel statistical prognostic score model that includes serum CXCL5 levels and clinical classification predicts risk of disease progression and survival of nasopharyngeal carcinoma patients. PLOS ONE 8: e57830.
22. Zhao X, Rdland EA, Srlie T, Naume B, Langerd A, et al. (2011) Combining gene signatures improves prediction of breast cancer survival. PLoS ONE 6: e17845.
23. Friedman JH (1997) On bias, variance, 0/1-loss, and the curse-of-dimensionality. Data Mining and Knowledge Discovery 1: 55-77.
24. B?uhlmann P, Hothorn T (2007) Boosting algorithms: Regularization, prediction and model fitting (with discussion). Statistical Science 22: 477–522.
25. Harrell F, Lee KL, Mark DB (1996) Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine 15: 361–387.
26. Heagerty PJ, Zheng Y (2005) Survival model predictive accuracy and ROC curves. Biometrics 61: 92–105.
27. PubMed (2013). US national library of medicine national institutes of health. URL http://www.ncbi.nlm.nih.gov/pubmed/.
28. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, et al. (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. Journal of the National Cancer Institute 81: 1879–1886.
29. Tice J, Cummings S, Ziv E, Kerlikowske K (2005) Mammographic breast density and the gail model for breast cancer risk prediction in a screening population. Breast Cancer Research and Treatment 94: 115-122.
30. Antolini L, Boracchi P, Biganzoli E (2005) A time-dependent discrimination index for survival data. Statistics in Medicine 24: 3927–3944.
31. van der Laan MJ, Robins JM (2003) Unified Methods for Censored Longitudinal Data and Causality. New York: Springer.
32. Schmid M, Kestler HA, Potapov S (2013) On the validity of time-dependent AUC estimators. Briefings in Bioinformatics .
33. Friedman JH, Hastie T, Tibshirani R (2000) Additive logistic regression: A statistical view of boosting (with discussion). The Annals of Statistics 28: 337–407.
34. Ma S, Huang J (2005) Regularized ROC method for disease classification and biomarker selection with microarray data. Bioinformatics 21: 4356–4362.
35. Buehlmann P (2006) Boosting for high-dimensional linear models. The Annals of Statistics : 559–583.
36. Schmid M, Hothorn T (2008) Boosting additive models using component-wise P-splines. Computational Statistics & Data Analysis 53: 298–311.
37. Hothorn T, Bühlmann P, Kneib T, Schmid M, Hofner B (2013) mboost: Model-Based Boosting. URL http://CRAN.R-project.org/package = mboost. R package version 2.2-3.
38. Mayr A, Hofner B, Schmid M (2012) The importance of knowing when to stop. A sequential stopping rule for component-wise gradient boosting. Methods of Information in Medicine 51: 178–186.
39. R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. ISBN 3-900051-07-0.
40. Lecocke M, Hess K (2006) An empirical study of univariate and genetic algorithm-based feature selection in binary classification with microarray data. Cancer Informatics 2: 313–327.
41. Klein JP, Moeschberger ML (2003) Survival Analysis - Techniques for Censored and Truncated Data. New York: Springer, 2nd edition.
42. Mayr A, Fenske N, Hofner B, Kneib T, Schmid M (2012) Generalized additive models for location, scale and shape for high-dimensional data – a exible approach based on boosting. Journal of the Royal Statistical Society: Series C (Applied Statistics) 61: 403–427.
43. Goeman J (2012) penalized: L1 (Lasso) and L2 (Ridge) Penalized Estimation in GLMs and in the Cox Model. R package version 0.9-42.
44. Potapov S, Adler W, Schmid M (2012) survAUC: Estimators of Prediction Accuracy for Time-to-Event Data. R package version 1.0-5. http://cran.rproject. org/web/packages/survAUC/index.html.
45. Hothorn T, B?uhlmann P, Dudoit S, Molinaro A, Van Der Laan M (2006) Survival ensembles. Biostatistics 7: 355–373.
46. Schmid M, Hothorn T (2008) Flexible boosting of accelerated failure time models. BMC Bioinformatics 9: 269.
47. van't Veer LJ, Dai HY, van de Vijver MJ, He YDD, Hart AAM, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–536.
48. van Houwelingen HC, Bruinsma T, Hart AA, van't Veer LJ, Wessels LF (2006) Cross-validated cox regression on microarray gene expression data. Statistics in Medicine 25: 3201–3216.
49. Pencina MJ, D'Agostino RB (2004) Overall C as a measure of discrimination in survival analysis: Model speci_c population value and confidence interval estimation. Statistics in Medicine 23: 2109–2123.
50. Wang Z, Chang Y (2011) Marker selection via maximizing the partial area under the ROC curve of linear risk scores. Biostatistics 12: 369–385.
51. Schmid M, Hothorn T, Krause F, Rabe C (2012) A PAUC-based estimation technique for disease classification and biomarker selection. Statistical Applications in Genetics and Molecular Biology 11, Article 3.