JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 23, Number 4, 2016
Mary Ann Liebert, Inc.
Pp. 279–290
DOI: 10.1089/cmb.2015.0192

Unbiased Prediction and Feature Selection in High-Dimensional Survival Regression

MICHAEL LAIMIGHOFER,[1,2] JAN KRUMSIEK,[1,3]
FLORIAN BUETTNER,[1,4] and FABIAN J. THEIS[1,2]

ABSTRACT

With widespread availability of omics profiling techniques, the analysis and interpretation of high-dimensional omics data, for example, for biomarkers, is becoming an increasingly important part of clinical medicine because such datasets constitute a promising resource for predicting survival outcomes. However, early experience has shown that biomarkers often generalize poorly. Thus, it is crucial that models are not overfitted and give accurate results with new data. In addition, reliable detection of multivariate biomarkers with high predictive power (feature selection) is of particular interest in clinical settings. We present an approach that addresses both aspects in high-dimensional survival models. Within a nested cross-validation (CV), we fit a survival model, evaluate a dataset in an unbiased fashion, and select features with the best predictive power by applying a weighted combination of CV runs. We evaluate our approach using simulated toy data, as well as three breast cancer datasets, to predict the survival of breast cancer patients after treatment. In all datasets, we achieve more reliable estimation of predictive power for unseen cases and better predictive performance compared to the standard CoxLasso model. Taken together, we present a comprehensive and flexible framework for survival models, including performance estimation, final feature selection, and final model construction. The proposed algorithm is implemented in an open source R package (SurvRank) available on CRAN.

Key words: feature selection, high-dimensional survival regression, repeated nested cross validation.

1. INTRODUCTION

In past years, new experimental technologies that allow measurement of tens of thousands of SNPs, transcripts, peptides, and metabolites in a cost-effective, high-throughput fashion have been developed. Consequently, omics measurements in patient samples are increasingly becoming part of clinical trials (McShane et al., 2013), because they promise to serve diagnostic purposes and accurately model patient

[1]Institute of Computational Biology, Helmholtz-Zentrum München, Neuherberg, Germany.
[2]Department of Mathematics, TU München, Garching, Germany.
[3]German Center for Diabetes Research (DZD), München-Neuherberg, Germany.
[4]European Bioinformatics Institute, European Molecular Biology Laboratory Hinxton, Cambridge, United Kingdom.

survival times. However, for such survival models to be adopted in clinical practice and diagnosis, it is crucial to accurately estimate the generalizability of these models (i.e., how well they perform with new patient cohorts). In addition, identification of a small set of highly predictive features in a high-dimensional survival setting is of particular clinical interest as it can facilitate large-scale screening of large patient cohorts. Example applications include identification of genetic marker sets to predict survival times after surgery in cancer research (Desmedt et al., 2007; van de Vijver et al., 2002) and the prediction of time to diabetes onset (Abbasi et al., 2012).

In high-dimensional medical datasets, the number of features p usually far exceeds the number of observations n ($n << p$). Several previous studies have addressed the $n << p$ problem in survival settings using regularization or feature selection approaches. Some authors have combined test statistics from univariate analyses into risk scores, for example, for lung cancer (Beer et al., 2002) and colorectal cancer (Eschrich et al., 2005). A drawback of these approaches is that each feature is individually associated with survival; however, joint information across features is not used. With polygenic risk scores or multivariate biomarkers, interest in full multivariable models has increased. As standard regression-based models are prone to overfitting in the $n << p$ scenario, shrinkage-based models, which regularize the effect estimates, are commonly used (Gui and Li, 2005; Wu et al., 2011; Gong et al., 2014; Datta et al., 2007). Alternatively, dimensionality reduction (e.g., PCA or clustering) can be performed prior to survival modeling (Alizadeh et al., 2000; Takamizawa et al., 2004; Zhao et al., 2005).

Here, we propose an approach that tackles two major challenges for predictive survival models in a single unified algorithm. TASK 1: A predictor must show good generalizability, that is it must correctly predict an outcome using unseen observations. Here, we aim to obtain unbiased predictions using only training data, that is in the absence of a validation dataset. The generalizability of this type of prediction model can be quantified using measures such as the concordance index (C-index) within a cross-validation (CV) framework for survival data (Harrell et al., 1982). For applicability in clinical settings, it is crucial to estimate this predictive power for new, unseen patients in an unbiased fashion. TASK 2: We aim to select a reduced set of informative features that retains high predictive accuracy. While different approaches to address these tasks in binary classification settings exist, to the best of our knowledge, there is no unified framework for high-dimensional survival settings.

We use a repeated nested CV strategy to tackle both tasks (Fig. 1). Specifically, we use a feature ranking-based approach to perform model selection followed by determination of the optimal number of features in the inner CV loop. The outer CV is used to estimate the prediction accuracy with the C-index with unseen data. By repeating the entire procedure, we quantify the intrinsic variation in the prediction accuracy. As

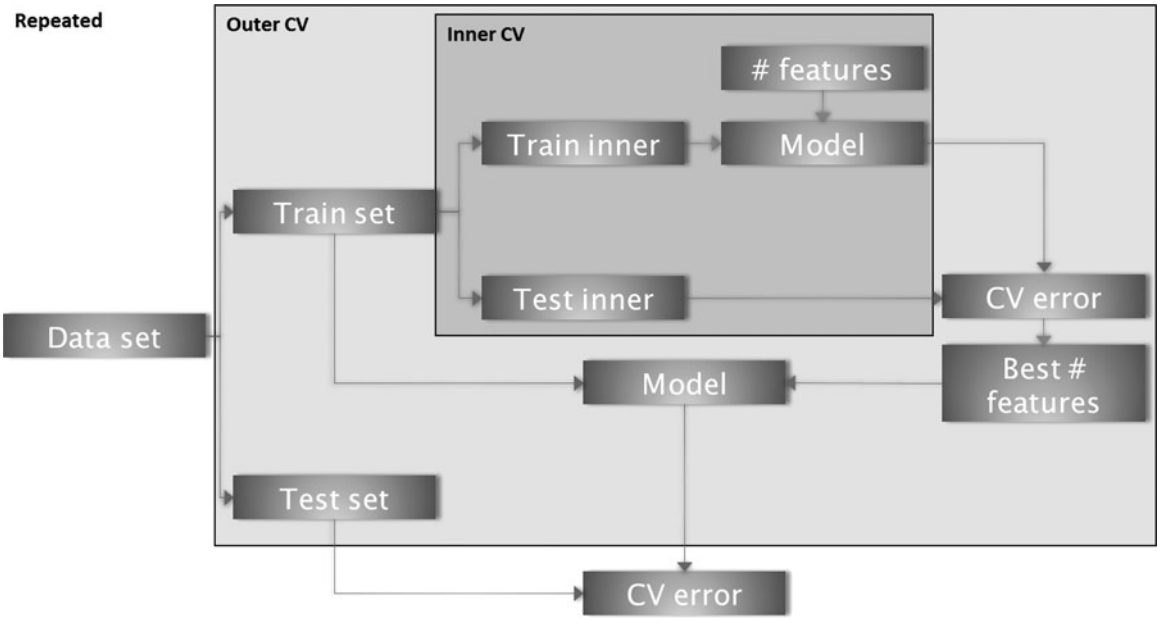

FIG. 1. Overview of the repeated nested Cross-validation scheme. In the inner CV, the optimal number of parameters is determined. The outer CV loop estimates unbiased prediction accuracy. The variance of the prediction accuracy is estimated by repeating the entire procedure.

different CV folds will produce different lists of feature rankings, we propose an algorithm to combine results. We weight the features according to their performance in the CV. TASK 1 is addressed by our method due to the strict separation of the training and test sets. We solve TASK 2 using our proposed approach to aggregate CV information into a final set of features.

We evaluate our approach with simulated data with a fixed set of features and show that existing methods (a regularized survival Cox model) exhibit strong bias. In addition, we test performance with three publicly available breast cancer datasets. These microarray-based datasets contain gene expression data from patients with lymph node-negative breast cancer after surgery or radiotherapy. Our pipeline is available as an R package (R Core, Team, 2014) SurvRank online.

2. METHODS

A survival dataset is defined by the triple $(T_i, \delta_i, \mathbf{x}_i)$ $i = 1,\ldots,n$ subjects, where T_i is the observed time (either failure time or censoring time), $\delta_i \in \{0, 1\}$ denotes the censoring indicator for a failure event (e.g., $\delta_i = 1$ in the case of relapse or death) or censoring information ($\delta_i = 0$), and the p-dimensional vector $\mathbf{x}_i$ defines the observed covariates of subject i. A subject is at risk if it undergoes an event or is censored. With $t_1 < \ldots < t_m$ being the ordered unique event times (with $\delta_i = 1$), at time t_j, all at-risk individuals constitute the risk set $R(t_j)$, which is defined as the set of all observations with longer observation time $T_i > t_j$.

In order to relate survival and observed covariates in our algorithm, we use the Cox proportional hazards model (Cox, 1972). In this model, the hazard for subject i is defined in semi-parametric form:

$$h_i(t|\mathbf{x_i}) = h_0(t) \exp\left(\sum_{k=1}^{p} x_{i,k}\beta_k\right) \tag{2.1}$$

where h_0 is a common baseline hazard and $\boldsymbol{\beta}$ is a vector of regression coefficients of length p. Inference on $\boldsymbol{\beta}$ is performed by maximizing the partial likelihood, defined as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp\left(\sum_{k=1}^{p} x_{i,k}\beta_k\right)}{\sum_{j \in R_i} \exp\left(\sum_{k=1}^{p} x_{j,k}\beta_k\right)} \tag{2.2}$$

where the baseline hazard $h_0(t)$ cancels out. The estimated risk score per subject is summarized by $\hat{\eta}_i = \sum_{k=1}^{p} x_{k,i}\hat{\beta}_k$, which expresses the linear combination of covariates with an estimated vector of coefficients $\hat{\boldsymbol{\beta}}$.

Investigation of the prediction accuracy and feature selection in our framework is performed with the C-index definition (Uno et al., 2011), denoted as C_{Uno}. The C-index of Uno for a prespecified point in time τ is defined as follows:

$$C_{Uno,\tau} = \frac{\sum_{j,k} \hat{G}(T_j)^{-2} I(T_j < T_k) I(\hat{\eta}_j < \hat{\eta}_k)\delta_j}{\sum_{j,k} \hat{G}(T_j)^{-2} I(T_j < T_k)\delta_j} \in [0, 1] \tag{2.3}$$

where $I()$ is an indicator function. Here, $\hat{G}(T_j)$ is estimated from the training data and is defined as the Kaplan–Meier estimator of the unconditional survival function:

$$\hat{G}(t) = \prod_{t_j \leq t} 1 - \frac{d_j}{R(t_j)} \tag{2.4}$$

with d_i being the number of events at t_j. The C_{Uno} index is estimated nonparametrically, thereby adjusting for the censoring bias via inverse probability weighting. A risk score η_i is estimated for the selected features with new data $\mathbf{x}_{test}$ for each individual in the test set. This score is used as input for the C_{Uno} function. To obtain the variation in C_{Uno} with an independent test set, we calculated prediction performance with different random subsamples (of 90%).

An advantage of the C_{Uno} approach compared to other C-index definitions (Heagerty and Zheng, 2005; Antolini et al., 2005) lies in its independence of the Cox proportional hazard assumption. The C-index can be interpreted as the probability of concordance between the predicted and observed survival times over all pairs of observations at a given time point. Similar to the standard binary AUC, a value of 0.5 indicates that the marker is not better than random guessing and a value of 1 represents perfect separation. In contrast to the standard area under the ROC curve, models with C-index of relatively low values (between 0.6 and 0.7) are often considered as

having satisfactory predictive power. For example, a C-index of 0.67 was achieved (Tice et al., 2005) in a model predicting breast cancer based on genetic information, known as the Gail model (Gail et al., 1989). In cancer research, the absolute discrimination power is often not required; however, separation and classification of patients into groups of high and low risk is the primary goal. Therefore, this C-index is a favorable choice.

2.1. SurvRank

A schematic overview of the algorithm is shown in Figure 1, and further details are given in Algorithm 1. To fit a survival model and estimate generalizability, a repeated nested CV approach is used to first estimate the best number of features within an inner CV loop and then to estimate the performance of the model containing these features in an outer CV loop. Note that the identification of important features within the CV is based on different ranking methods.

2.1.1. Feature ranking methods. Three approaches to generate ranked output lists of features were considered, that is, an approach based on the log-rank statistic (survCox), a Lasso-based approach for survival data (survLasso), and a randomized Cox model (survRand).

Cox score ranking - survCox The Cox-based ranking approach sorts covariates according to their association with the survival response based on the Wald score test. For each feature, a univariate Cox model is fitted, and the obtained log-rank statistic is used as the ranking criterion (Moeschberger and Klein, 2003). A high test statistic indicates stronger association with the outcome. Note that this Cox score ranking is univariate in contrast to the other two approaches.

L_1 ***norm (Lasso) ranking - survLasso*** In this approach, ranking is generated using a penalized L_1 Cox regression (Tibshirani and others, 1997). Briefly, the L_1 penalty (Lasso) in the Cox regression case seeks to find a solution for the following:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\frac{2}{n} \left(\sum_{i=1}^{m} \mathbf{x}_{j(i)}^T \boldsymbol{\beta} - \log \left(\sum_{j \in R_i} \exp\left(\mathbf{x}_j^T \boldsymbol{\beta}\right) \right) - \lambda \sum_{i=1}^{p} |\beta_i| \right) \right). \tag{2.5}$$

An efficient implementation of the regularization path has been demonstrated (Simon et al., 2011). The complexity parameter λ determines the amount of shrinkage. The ranks of features are then obtained according to their appearance in the regularization path. All covariates not selected in the model obtain a rank that corresponds to the number of features p.

Randomized Cox ranking - survRand This ranking method consists of a two-step procedure. In the first step, L_1 penalization is used to preselect a smaller number of features ($p_{\text{pre}} < p$). The cut-off criterion in the Lasso is defined such that at least 95% of the deviance is explained at the end of the lambda sequence. In the second step, a sub-sampling approach (500 times) randomly chooses without replacement a smaller number of features and estimates a multivariate Cox model. To avoid convergence issues in the fitting procedure of the multivariate Cox model, the number of features in each subsampling step n_{sub} is limited to the number of observations ($n_{sub} = n/3$). Each feature in one subsampled Cox model yields a Z-statistic. The number of selections per feature is controlled by p_{pre}, thereby leading to $p_{pre}/3$ number of Z-statistics for each feature on average. Finally, by calculating the mean over all Z-score subsamples, a final feature score is derived and used for ranking in survRand.

2.1.2. Nested CV for estimating generalizability—TASK 1. The full dataset $D := D_i$ with $D_i := (T_i, \delta_i, x_i)$ is split into training set $D^{-cv_{out}}$ and test set $D^{cv_{out}}$ with index set cv_{out} and its complement.

Inner CV Inner CV is applied to only the training set $D^{-cv_{out}}$ by performing a second CV stratification, thereby yielding inner training $D^{(-cv_{out}, -cv_{in})}$ and inner test set $D^{(-cv_{out}, cv_{in})}$. Then, one of the described ranking functions is applied to the inner training set. By adding one feature at a time (following the ranking), a Cox model is estimated using the inner training data and evaluated with C_{Uno} for the inner test data. This procedure is performed until a predefined maximum number of features is achieved and is repeated for all inner CV folds. The best number of features is determined by averaging over all inner CVs and selecting the number of features that corresponds to the maximum mean C_{Uno}.

Outer CV For the outer CV, one feature ranking is performed for the whole training set. Then, using the best number of features derived in the inner CV, a Cox model is estimated using the training set, thereby yielding effect estimates for the selected features. Using these estimates, the unbiased prediction performance with the unseen test set is quantified by C_{Uno}, corresponding to TASK 1. Note that the entire procedure (including the inner CV) is applied to all outer CV folds.

Repeated CV To obtain an estimate of the variance of prediction accuracy, this approach is repeated *t_times* for different splits of the dataset.

2.1.3. Final model—TASK 2. The repeated nested CV combined with stepwise feature selection based on the ranking function yields a ranked set of features for each CV run. In addition, the performance on the test set for each run is recorded (number of runs $K = cv_{out} \times t\_times$). As these ranked lists of selected features are not necessarily the same, it is not clear how to aggregate them to a final set of features that can be used for predicting risk scores for new patients. Here, we propose an approach that leverages the information from all individual CV runs to determine a final set of features for which a final model can be fit.

Our weighted approach uses information from the outer CV performance corresponding to each run, thereby addressing TASK 2. The weight of run i is defined as follows:

$$w_i = \begin{cases} \frac{1}{K}\exp\left(\log(2)\frac{devAUC_i}{0.1}\right), & \text{if } C_{Uno,i} \geq 0.5 \\ 0, & \text{if } C_{Uno,i} < 0.5 \end{cases} . \tag{2.6}$$

Here, $devAUC_i$ denotes the relative C_{Uno} of an individual CV run compared to the average performance of all runs. The weights w_i' are further normalized to sum to one ($w_i' = w_i / \Sigma w_i$). The final set of predictors is determined by majority voting as follows:

$$I(p_j) = \begin{cases} 1 & \text{if } p_j > 0.5 \\ & \text{with } p_j = \sum_{i=1}^{K} I(j,i) w_i' \\ 0, & \text{if } p_j \leq 0.5 \end{cases} \tag{2.7}$$

where $I(j, i)$ is 1 if the feature p_j was selected in run i.

Algorithm 1: SurvRank algorithm with repeated nested CV

Data: survival data set (T_j, δ_j, $\mathbf{x_j}$);
parameters of rep CV: repetition *t_times*, outer CV *cv_out*, inner CV *cv_in*;
maximum number of features *max_var*;
ranking_fct (survLasso, survCox, survRand);
coxmodel function estimates $\hat{\beta}$ on a data set;
final_feature_fct (weighted);
Result: final set of selected features of the nested CV approach

```
for t = 1 : t_times do
  for j = 1 : cv_out do
    train_outer ← (T_j^(−cv_out), δ_j^(−cv_out), X_j^(−cv_out));
    test_outer ← (T_j^(cv_out), δ_j^(cv_out), X_j^(cv_out));
    for k = 1 : cv_in do
      train_inner ← (T_j^(−cv_out, −cv_in), δ_j^(−cv_out, −cv_in), X_j^(−cv_out, −cv_in));
      test_inner ← (T_j^(−cv_out, cv_in), δ_j^(−cv_out, cv_in), X_j^(−cv_out, cv_in));
      ranking_in ← ranking_fct(train_inner);
      for i = 1 : max_var do
        coxmodel_in ← coxmodel(ranking_in[1 : i], train_inner);
        surv_in[i, k] ← Cindex(coxmodel_in, test_inner);
      end
    end
    meanCurve ← mean(surv_in,k);
    maxFeature ← which_max(meanCurve);
    ranking_out ← ranking_fct(train_outer)[1:maxFeature];
    coxmodel_out ← coxmodel(ranking_out, train_outer);
    surv_out[j,t] ← Cindex(coxmodel_out, test_outer);
  end
end
sel_features ← final_feature_fct(surv_out, ranking_out);
final_model ← coxmodel(sel_features, (T_j, δ_j, x_j));
```

Finally, one survival model can be calculated with the selected features using the entire dataset, thereby leading to effect estimates $\hat{\boldsymbol{\beta}}_{train}$ and risk scores for each subject. This is used to predict survival probabilities with unseen data with similar predictive power as estimated in the nested CV.

2.2. Comparison method

To compare this approach to existing methods, a commonly used regularized survival model based on Cox-Lasso was selected (coxLasso). For coxLasso, the same unbiased approach was performed to estimate the prediction accuracy with CV by applying the same repeated CV parameters. One CV step consists of separation into different folds and optimizing the penalization parameter $\hat{\lambda}$ by the inner CV of one fold. This optimized $\hat{\lambda}$ was used to predict the unseen test fold, thereby measuring performance with C_{Uno}. For coxLasso, the final selection of covariates, which are used for prediction with the test set, was estimated by applying CV to the entire training dataset once. By optimizing the partial likelihood in the Cox regression, the number of features was obtained with cross-validated minimum deviance for coxLasso.

3. RESULTS

3.1. Simulation and validation setup

To evaluate our algorithm, we generated a high-dimensional, multivariate normally distributed dataset with $n=100$ observations and $p=500$ features. The survival times T_i followed an exponential distribution with mean $\eta_i = 1/(\lambda_T \sum_{i_1}^{4} x_i \beta_i)$, which we set to $\lambda_t = 0.5$ and $\beta_1 = 1.5$, $\beta_2 = -1.5$, $\beta_3 = -1$, and $\beta_4 = 1$ for our framework. An independent random censoring time T_{cens} was simulated such that it followed an exponential distribution, which we fixed to mean 2. The observed survival times T_{obs} are expressed by $T_{obs} = min(T_{cens}, T_i)$, which leads to independent censoring of approximately 50%. The maximum number of features was set to 30. Partitioning into training and test sets was applied in all configurations with the same parameters ($cv_{in}=10$, $cv_{out}=10$, $t\_times=10$). To calculate C_{Uno}, we fixed τ at the last observed survival time.

We first used the simulated data to estimate generalizability accurately, which is directly related to TASK 1. By applying a final model fit on the training set and estimating the performance with 10 simulated test sets, we retrieved the performance of our model selection with new unseen data. Ideally, the performance difference between the training and test data should be small. Otherwise, we would have a classical overfitting situation with the training data, where generalization accuracy to new unseen test data is not fulfilled. This procedure was repeated for 100 different training datasets. Furthermore, we calculated the true C_{Uno} for the training set and the test sets using only the true effects $\beta_1,\ldots,\beta_4$.

We then attempted to retrieve the correct set of features, thereby addressing TASK 2 (feature selection). To achieve this, we calculated the F_1 score, which is defined as the harmonic mean of precision and recall, that is, $F_1 = 2 \times \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$. Here, the F_1 score was calculated to compare the selected features with respect to the four true features.

We then compared our approach with a commonly used regularized survival model. Here, we estimated a penalized survival Cox model with Lasso (coxLasso based on the R package glmnet).

3.2. Simulated dataset results

We observed good performance with the test data and comparable results for accuracy with the training set compared to the test sets (Fig. 2), thereby addressing TASK 1. The coxLasso approach performed similarly with the training data compared to survLasso from our package; however, as expected, prediction with unseen new data shows substantial overfitting. The survRand ranking function demonstrated higher variance of C_{Uno} with the training set. survCox ranking performed worse with the training data; however, the final feature selection results showed comparable prediction accuracy with new test data. The overall worse performance of survCox illustrates the advantage of the multivariate ranking function of survLasso and survRand compared to survCox with univariate ranking.

Compared to standard coxLasso, the sparser set of selected features represents an advantage of our ranking and final feature selection approach (Fig. 3A), thereby addressing TASK 2. This illustrates that selecting features according to the data fit (deviance), as used in the standard coxLasso approach, produces too many selected features. In addition, we investigated whether the correct covariates were selected. We

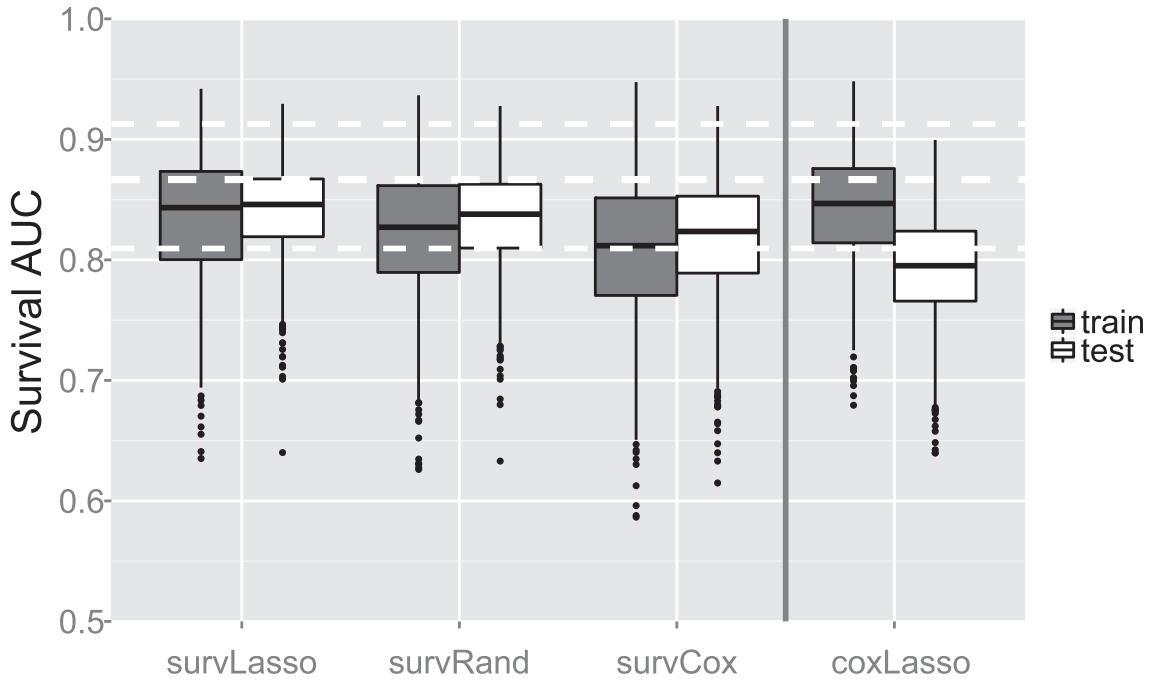

FIG. 2. Prediction performance with simulated data. A total of 100 training datasets were simulated, and unbiased C_{Uno}s were obtained for each repetition of the nested CV. For each of the 100 training datasets, 10 test sets were created to test prediction performance with new data. White dashed lines indicate the average of the true C_{Uno} with the simulated datasets with an empirical 95% quantile range.

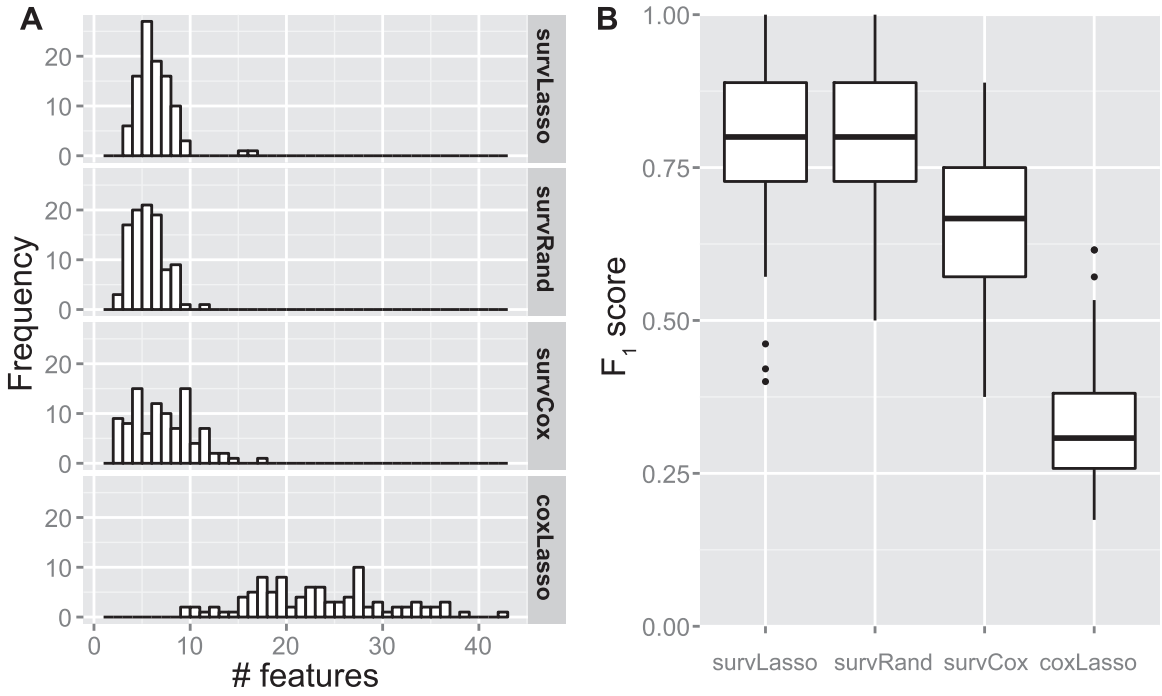

FIG. 3. **(A)** Number of selected features across simulated training datasets in the weighted approach. **(B)** F_1 scores for selected features to compare the selected features with the set of four true features.

Table 1. Computation Time in Minutes for Different $p \times n$ Setups

p	100	200	200	500
n	100	100	200	100
survLasso	9.00	9.43	20.33	19.00
survCox	10.03	16.43	16.48	27.58
survRand	77.07	82.70	186.57	86.28
coxLasso	3.70	3.98	14.38	5.42

Parameters set to $t\_times = 10$, $cv_{out} = 10$, and $cv_{in} = 10$.

observed higher F_1 scores (Fig. 3B) with our approach compared to coxLasso. These results illustrate the overfitting of the coxLasso approach, that is, it selects several random, noninformative features (resulting in a high FPR) and considerably overestimates predictive power with training sets (reduction of C_{Uno} on average of 0.05 or 6% from training to test).

3.2.1. Runtime evaluation. An important aspect for nested CV approaches is the required computation time. The SurvRank package inherently supports parallelization across multiple cores on the same machine. Table 1 shows the runtimes for different variable settings for the three ranking functions using a single core of an Intel Core i5 2.6 GHz CPU. Here, we observed that the number of features p scaled approximately linearly with computation time for survLasso, survRand, and coxLasso. survLasso was slower than coxLasso in the first two settings by a factor of approximately 2.5, taking the additional stepwise selection into account. Doubling the number of observations n increased computation time by a factor of 2.2 for survLasso and survRand and by a factor of 3.6 for coxLasso. In contrast, the computation time of survCox scaled approximately linearly with the number of features due to the univariate ranking procedure. For survCox, an increasing sample size increased computation time only slightly.

3.3. Application to three breast cancer gene expression datasets

To evaluate our approach with real clinical data, we applied the pipeline to microarray datasets from breast cancer patients with survival information (relapse time) after surgery (mastectomy) or radiotherapy. We used two independent datasets to estimate the prediction accuracy with unseen data to assess how well our method performs with TASK 1. To identify a predictive subset of features, we used our approach with different ranking functions, thereby addressing TASK 2. In addition, we compared the performance of our approach to a standard CoxLasso model and a set of 76 marker genes identified in the primary publication (referred to as geneMarker). This geneMarker was derived by ranking the features according to an averaged Cox score (using bootstrap samples).

The first dataset contained 286 patients with lymph node-negative breast cancer. For each patient, information about estrogen receptor status positive (ER+) and estrogen receptor status negative (ER-) was recorded, assuming that disease progression differs for these subgroups. This first dataset served as the training set [accession number GSE2034 (Wang et al., 2005)]. Wang et al. identified a predictive set of 76 genes (geneMarker) composed of 60 genes for the ER+ group and 16 genes for the ER- group. We attempted to obtain an alternative sparse set of genes with better generalizability to evaluate the performance of our approach with two independent validation sets, that is, accession numbers GSE7390 (Desmedt et al., 2007) and GSE1456 (Pawitan et al., 2005). There was an overlap of 18,842 features across the three datasets. In the training data, there were 209 patient samples in the ER+ group and 77 observations with ER- status. The first test dataset (test set 1) consisted of 134 samples in the ER+ group and 64 in the ER- group. The second test set (test set 2) contained 125 subjects in the ER+ group and 27 in the ER- group. Due to the larger number of observations, we focused on the ER+ subgroup for our evaluation.

We applied our different ranking algorithms to the dedicated training set and obtained a final marker. Furthermore, the selected genes were evaluated with the new and unseen test sets. The parameters of the repeated nested CV were determined as $t\_times = 20$, $cv_{out} = 10$, and $cv_{in} = 10$. The maximum number of features was set to 75, and τ in C_{Uno} was set to 10 years.

The geneMarker and the coxLasso approach served as comparison models for our ranking algorithms. The results of geneMarker were calculated by applying ridge regression to the training data and then

evaluating performance with the two test sets. For coxLasso, we repeated the final feature selection ten times to determine the optimal penalization parameter, because coxLasso depends on the sampling of CV folds.

3.4. Breast cancer data results

For our approach, performance with the unseen test dataset showed similar prediction accuracy compared to the training data (Fig. 4). This indicates that our nested CV strategy was able to estimate the generalizability of the predictor correctly, thereby solving TASK 1. The number of selected features varied slightly between the three approaches of our package (24, 19, and 29 for survLasso, survRand, and survCox, respectively), thereby addressing TASK 2. survLasso and survCox showed larger overlap of selected genes compared to survRand (Fig. 5). As in the simulation study, survLasso performed considerably better than survCox (on average C_{Uno} decreased by 0.03 or 5%), again illustrating the advantages of a multivariate ranking approach compared to univariate ranking. Similar to the results of the simulation study, coxLasso selected 53 features with too many false positives, resulting in a reduced performance with the test data sets. geneMarker resulted in clear overfitting of this marker set with the training dataset (as expected), where geneMarker was derived. Therefore, these results can be interpreted as training performance. Consequently, the predictive power decreased strongly with the test sets. Comparing the geneMarker set with the selected markers in survLasso, survRand, and survCox yielded a small overlap, that is, survLasso 2 genes, survRand 0, and survCox 5 (details in Supplementary Fig. S1, available online at www.liebertpub.com/cmb).

4. DISCUSSION

We have proposed a new framework to reliably estimate prediction accuracy and generalizability and to select the most predictive features in a high-dimensional survival prediction setting. To avoid overfitting

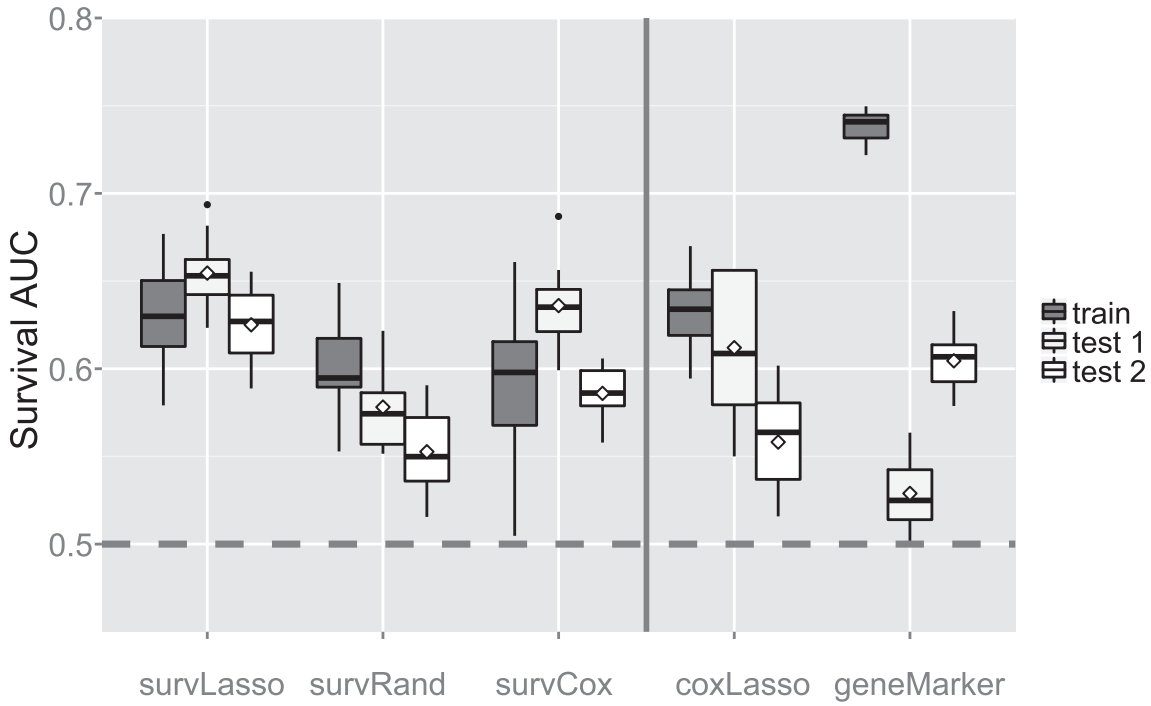

FIG. 4. Prediction accuracy with three breast cancer data sets. The performance of the training data set was compared to two independent test sets for the ER+ group. Feature selection was based on the weighted approach. Diamonds show performance with the whole test set, whereas variation in the boxplots was obtained by subsampling the test data sets.

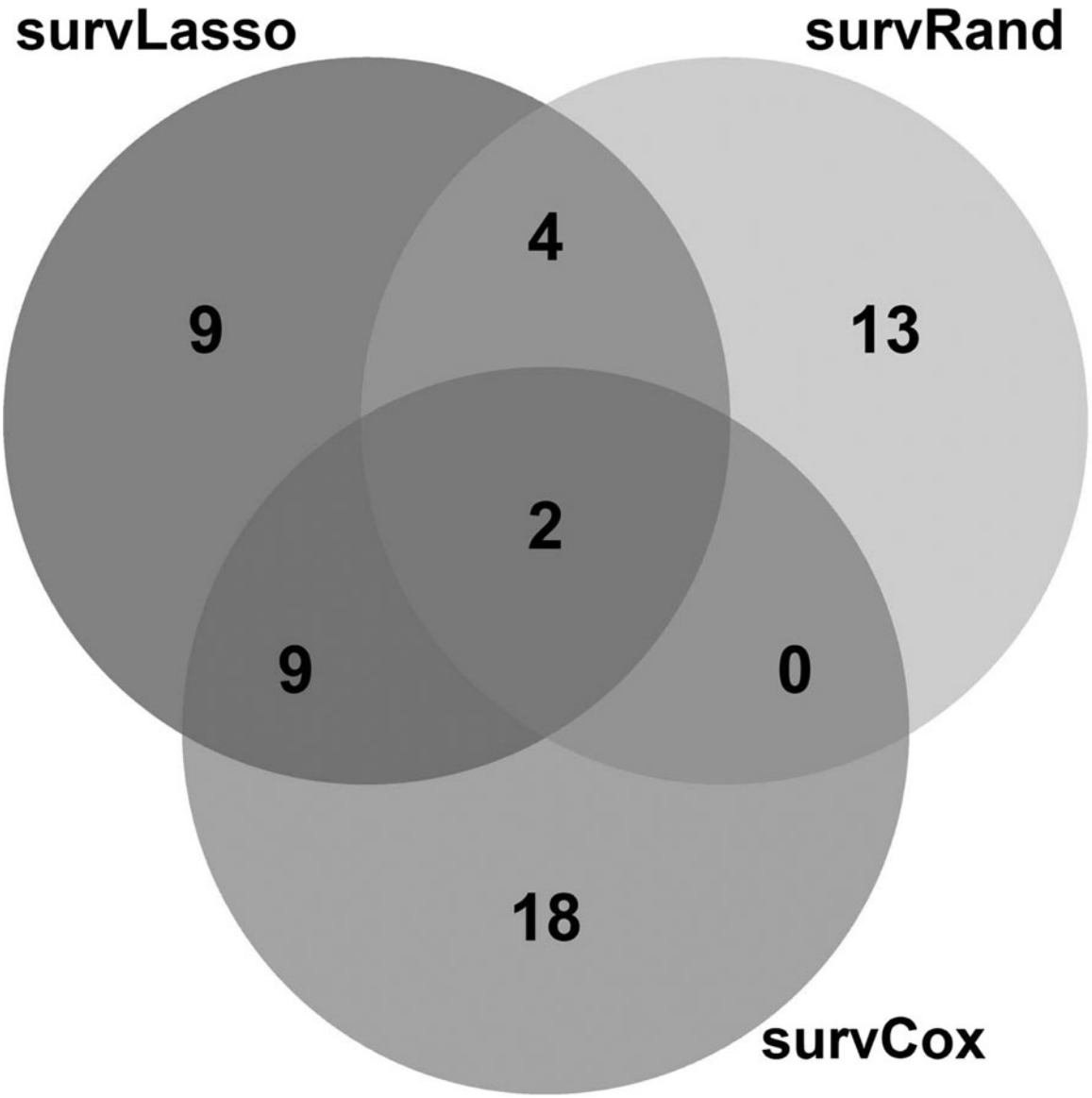

FIG. 5. Overlap of selected genes of the different ranking functions.

while selecting features with high predictive power, the proposed approach estimates accuracy and performs feature selection using repeated nested CV with novel feature combination heuristics.

Our approach differs from standard approaches, such as the CoxLasso approach, in two ways. First, the selection of features is determined by the best predictive feature combination (using C_{Uno}) rather than the best data fitting combination, thereby reducing the risk of overfitting. Second, for final feature selection, our approach leverages information from different CV runs. The CoxLasso approach uses the minimum cross-validated deviance of the whole dataset, while the proposed approach aggregates the results of different CV runs and applies a weighting scheme to select only predictive features. This combination of aggregating CV runs by weighting results in sparser feature selection with more accurate estimation of predictive power.

Using simulated data, we demonstrated that the proposed method can identify true features and can correctly estimate prediction accuracy with new data without overfitting. By comparing the results of different methods in this simulation setup, we observed that survLasso dominates survCox with training and test data. This effect can be explained by the multivariable ranking procedure of survLasso (considering all features) in contrast to the univariate ranking of survCox, which treats features independently.

With breast cancer data, our pipeline based on two of our ranking approaches was able to estimate similar prediction performance with the test datasets compared to the training data. However, the survRand

approach showed a drop in prediction performance with the breast cancer test data. This effect is illustrated in Figure 5, where we observe that this ranking approach has only small overlap compared to survLasso and survCox. The 19 selected features in this approach lead to lower prediction performance. By comparing coxLasso and survRand, we observed an overlap of six features that are only picked by these methods (Supplementary Fig. S1), thereby introducing noise to the model. In addition, the sampling strategy of survRand might introduce some noise to the selection process. This again confirms the robust performance of survLasso compared to the other ranking methods.

Our approach can be extended in several directions. (1) In clinical applications, variables such as age, gender, height, and BMI are collected routinely. Therefore, it would be desirable to force such features into the model and evaluate the additional benefit of omics data. (2) Our framework uses the Cox proportional hazards model. Extending the approach to accelerated failure time models or frailty models may improve the baseline hazard estimation, such as time-varying hazards or random effects. (3) Applying repeated nested CV to classification tasks may also be an interesting extension.

Importantly, our approach as a biomarker discovery method focuses on identifying a predictive biomarker combination and does not provide functional interpretation of the selected features (e.g., genes and transcripts). Therefore, we recommend using the SurvRank package with the survLasso approach and weighted final feature selection, due to the low computational demands and best results from both the simulation study and the clinical data.

In summary, we provide a flexible, ready-to-use toolbox for survival data that allows for unbiased estimation of prediction accuracy for survival models and extracts the most predictive features from high-dimensional survival datasets.

ACKNOWLEDGMENTS

This work was funded in part by grants from the German Federal Ministry of Education and Research (BMBF), grant No. 01ZX1313C (project e:Athero-MED) and 01ZX1314G (project IntegraMent), and from the European Union's Seventh Framework Programme [FP7-Health-F5-2012] under grant agreement number 305280 (MIMOmics). F.B. was supported by a UK Medical Research Council Career Development Award (Biostatistics).

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Abbasi, A., Peelen, L.M., Corpeleijn, E., et al. 2012. Prediction models for risk of developing type 2 diabetes: Systematic literature search and independent external validation study. *BMJ* 345, e5900.

Alizadeh, A.A., Eisen, M.B., Davis, R.E., et al. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511.

Antolini, L., Boracchi, P., and Biganzoli, E. 2005. A time-dependent discrimination index for survival data. *Stat. Med.* 24, 3927–3944.

Beer, D.G., Kardia, S.L., Huang, C.-C., et al. 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* 8, 816–824.

Cox, D.R. 1972. Regression models and life-tables. *J. R. Stat. Soc. B* 34, 187–220.

Datta, S., Le-Rademacher, J., and Datta, S. 2007. Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics* 63, 259–271.

Desmedt, C., Piette, F., Loi, S., et al. 2007. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res.* 13, 3207–3214.

Eschrich, S., Yang, I., Bloom, G., et al. 2005. Molecular staging for survival prediction of colorectal cancer patients. *J. Clin. Oncol.* 23, 3526–3535.

Gail, M.H., Brinton, L.A., Byar, D.P., et al. 1989. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* 81, 1879–1886.

Gong, H., Wu, T.T., and Clarke, E.M. 2014. Pathway-gene identification for pancreatic cancer survival via doubly regularized Cox regression. *BMC Syst. Biol.* 8, 1–9.

Gui, J., and 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.

Harrell, F.E., Califf, R.M., Pryor, D.B., et al. 1982. Evaluating the yield of medical tests. *JAMA* 247, 2543–2546.

Heagerty, P.J., and Zheng, Y. 2005. Survival model predictive accuracy and ROC curves. *Biometrics* 61, 92–105.

McShane, L.M., Cavenagh, M.M., Lively, T.G., et al. 2013. Criteria for the use of omics-based predictors in clinical trials. *Nature* 502, 317–320.

Moeschberger, M.L., and Klein, J. 2003. *Survival Analysis: Techniques for Censored and Truncated Data: Statistics for Biology and Health*. Springer, New York.

Pawitan, Y., Bjhle, J., Amler, L., et al. 2005. Gene expression profiling spares early breast cancer patients from adjuvant therapy: Derived and validated in two population-based cohorts. *Breast Cancer Res.* 7, R953.

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

Simon, N., Friedman, J.H., Hastie, T., and Tibshirani, R. 2011. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* 39, 1–13.

Takamizawa, J., Konishi, H., Yanagisawa, K., et al. 2004. Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res.* 64, 3753–3756.

Tibshirani, R., et al. 1997. The lasso method for variable selection in the Cox model. *Stat. Med.* 16, 385–395.

Tice, J.A., Cummings, S.R., Ziv, E., and Kerlikowske, K. 2005. Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population. *Breast Cancer Res. Treat.* 94, 115–122.

Uno, H., Cai, T., Pencina, M.J., et al. 2011. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* 30, 1105–1117.

van de Vijver, M.J., He, Y.D., van't Veer, L.J., et al. 2002. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* 347, 1999–2009.

Wang, Y., Klijn, J.G.M., Zhang, Y., et al. 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 365, 671–679.

Wu, T.T., Gong, H., and Clarke, E.M. 2011. A transcriptome analysis by lasso penalized Cox regression for pancreatic cancer survival. *J. Bioinform. Comput. Biol.* 9 Suppl 1, 63–73.

Zhao, H., Ljungberg, B., Grankvist, K., et al. 2005. Gene expression profiling predicts survival in conventional renal cell carcinoma. *PLoS Med* 3, e13.

Address correspondence to:
Florian Buettner or Fabian Theis
Institute of Computational Biology
Helmholtz-Zentrum München
Ingolstädter Landstraße 1
85764 Neuherberg
Germany

E-mail: buettner@ebi.ac.uk
or
fabian.theis@helmholtz-muenchen.de