Madjar and Rahnenführer
BMC Medical Informatics and Decision Making (2021) 21:342
https://doi.org/10.1186/s12911-021-01698-1

BMC Medical Informatics and Decision Making

TECHNICAL ADVANCE

Open Access

Weighted Cox regression for the prediction of heterogeneous patient subgroups

Katrin Madjar* and Jörg Rahnenführer

Abstract

Background: An important task in clinical medicine is the construction of risk prediction models for specific subgroups of patients based on high-dimensional molecular measurements such as gene expression data. Major objectives in modeling high-dimensional data are good prediction performance and feature selection to find a subset of predictors that are truly associated with a clinical outcome such as a time-to-event endpoint. In clinical practice, this task is challenging since patient cohorts are typically small and can be heterogeneous with regard to their relationship between predictors and outcome. When data of several subgroups of patients with the same or similar disease are available, it is tempting to combine them to increase sample size, such as in multicenter studies. However, heterogeneity between subgroups can lead to biased results and subgroup-specific effects may remain undetected.

Methods: For this situation, we propose a penalized Cox regression model with a weighted version of the Cox partial likelihood that includes patients of all subgroups but assigns them individual weights based on their subgroup affiliation. The weights are estimated from the data such that patients who are likely to belong to the subgroup of interest obtain higher weights in the subgroup-specific model.

Results: Our proposed approach is evaluated through simulations and application to real lung cancer cohorts, and compared to existing approaches. Simulation results demonstrate that our proposed model is superior to standard approaches in terms of prediction performance and variable selection accuracy when the sample size is small.

Conclusions: The results suggest that sharing information between subgroups by incorporating appropriate weights into the likelihood can increase power to identify the prognostic covariates and improve risk prediction.

Keywords: Cox proportional hazards model, Heterogeneous cohorts, High-dimensional data, Subgroup analysis, Weighted regression

Background

Survival analysis is an important field of biomedical research, particularly cancer research. The main objectives are the prediction of a patient's risk and the identification of new prognostic biomarkers to improve patients' prognosis. In recent years, molecular data such as gene expression data have increasingly gained importance in diagnosis and prediction of disease outcome. Technologies for the measurement of gene expression have made rapid progress and the use of high-throughput technologies allows simultaneous measurements of genome-wide data for patients, resulting in a vast amount of data.

A typical characteristic of this kind of high-dimensional data is that the number of genomic predictors greatly exceeds the number of patients ($p >> n$). In this situation, the number of genes associated with a clinical outcome, here time-to-event endpoint, is typically small. Important objectives in modeling high-dimensional data are good prediction performance and finding a subset of predictors that are truly relevant to the outcome. A sparse model solution may reduce noise in estimation and increase interpretability of the results. Another

*Correspondence: madjar@statistik.tu-dortmund.de
Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany

problem with high-dimensional data is that standard approaches for parameter estimation in regression models cannot handle such a large number of predictors; conventional regression techniques may not provide a unique solution to maximum likelihood problems or may result in an overfitted model. During the last years, different approaches have been proposed for handling the $p >> n$ situation, often implying automatic variable selection, such as regularization [33, 38, 45] or boosting algorithms [4, 20, 21, 35].

In clinical practice, patient cohorts are typically small. However, when data of several patient cohorts or subgroups with the same or similar disease are available it can be reasonable to use this information and appropriately combine the data. In multicenter studies, patients of all subgroups are often simply pooled. When subgroups are heterogeneous with regard to their relationship between predictors and outcome, this combined analysis may suffer from biased results and averaging of subgroup-specific effects. Standard subgroup analysis, on the other hand, includes only patients of the subgroup of interest and may lead to a loss of power when the sample size is small.

We aim at providing a separate prediction model for each subgroup that allows for identifying common as well as subgroup-specific effects and has improved prediction accuracy over both standard approaches. Therefore, we propose a Cox proportional hazards model that allows sharing information between subgroups to increase power when this is supported by data. We use a lasso penalty for variable selection and a weighted version of the Cox partial likelihood that includes patients of all subgroups but assigns them individual weights based on their subgroup affiliation. Patients who are likely to belong to the subgroup of interest obtain higher weights in the subgroup-specific model. We estimate individual weights for each patient from the training data following the idea of Bickel et al. [3].

We assume subgroups are pre-known and determined by multiple cancer studies or cohorts. However, our approach can be applied to any other type of subgroups, for example, defined by clinical covariates. Our proposed model is evaluated through simulations and application to real lung cancer cohorts, and compared to the standard subgroup model and the standard combined model.

Related work

Different approaches have been published recently suggesting the use of weights in regression models to consider subgroups. Weyer and Binder [39] aim at improving stability and prediction quality of a Cox model for a specific subgroup by including one additional weighted subgroup. The authors use a weighted and stratified Cox regression model based on componentwise boosting for automatic variable selection. They study the effects of a set of different fixed weights $w \in (0, 1)$ for the additional subgroup, while all observations in the subgroup of interest obtain a weight of 1 in the stratum-/subgroup-specific likelihood. In this paper, we compare a set of different fixed weights as suggested by Weyer and Binder [39] to our more flexible approach with individual weights for each patient from each subgroup estimated from the (training) data. However, we assume the same baseline hazard rate across all subgroups in contrast to the stratified Cox model by Weyer and Binder [39].

Alternatively, subgroup weights can be considered as a tuning parameter in model-based optimization (MBO) to improve prediction performance in the Cox model. This approach by Richter et al. [29] is more flexible than the previously mentioned one by Weyer and Binder [39] since it allows different fixed weights for different subgroups in each subgroup model. However, it also makes the restriction that all weights for patients from the same subgroup must be the same, which is quite different in terms of spirit from our proposed approach with individual weights for each patient from each subgroup. The major different idea of the MBO method is to quickly find a good set of fixed weights for the other subgroups in terms of prediction performance. Despite its difference in spirit, this alternative procedure could be an interesting outlook for further comparison studies.

Bayesian approaches for the estimation of subgroup weights were proposed by Bogojeska and Lengauer [7] and Simon [31]. However, they are not designed for our high-dimensional situation since they do not perform variable selection.

Weighted regression models are also used in local regression, however without predefined groups. For each individual, a local regression model is fitted based on its neighboring observations. The latter are weighted by their distances from the observation of interest. Penalized localized regression approaches for dealing with high-dimensional data exist [5, 34]. Instead of using distance in covariate space, our proposed weights correspond to the relationship between covariates and subgroup membership. A drawback of localized regression is that it does not provide global regression parameters, making interpretation difficult. Furthermore, only a small number of observations is used for each local fit in contrast to our approach, where the weighted likelihood is based on all training data.

We define subgroups by multiple cancer studies or cohorts and aim at appropriately combining them to increase power and simultaneously, considering heterogeneity among the subgroups. This idea of combining data from different data sources is similar to integrative

analysis. In high-dimensional settings with genomic predictors, different publications suggest the use of specific penalties in regularized regression for parameter estimation and variable selection across multiple data types. For example, Liu et al. [24] and Liu et al. [25] propose composite penalties with two-level gene selection. In the first selection level represented by an outer penalty, the association of a specific gene in at least one study is determined. In the second level, inner penalties of ridge or lasso type are used to allow the selection of either the same set of genes or different sets of genes in all studies. Instead of aggregating multiple studies with the same type of (omics) data, Boulesteix et al. [8] perform an integrative analysis of multiple omics data types available for the same patient cohort. The authors use a lasso penalty with different penalty parameters for the different data types. Bergersen et al. [2] integrate external information provided by another genomic data type by using a weighted lasso that penalizes each covariate individually with weights inversely proportional to the external information. Gade et al. [15] use a bipartite graph to integrate miRNA and gene expression data from the same patient cohort into one prediction model to find a combined signature that improves the prediction. This graph is built by combining correlations between both data types and external information on target predictions.

Methods

Cox proportional hazards model

Assume the observed data of patient i consists of the tuple (t_i, δ_i), the covariate vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})' \in \mathbb{R}^p$, and the subgroup membership $s_i \in \{1, \ldots, S\}$ with S the number of subgroups in the complete data set, and $i = 1, \ldots, n$. $t_i = \min(T_i, C_i)$ denotes the observed time of patient i, with T_i the event time and C_i the censoring time. $\delta_i = \mathbb{1}(T_i \leq C_i)$ indicates whether a patient experienced an event ($\delta_i = 1$) or was (right-)censored ($\delta_i = 0$).

The most popular regression model in survival analysis is the Cox proportional hazards model [12]. It models the hazard rate $h(t|\boldsymbol{x}_i)$ of an individual at time t as

$$h(t|\boldsymbol{x}_i) = h_0(t) \cdot \exp(\boldsymbol{\beta}'\boldsymbol{x}_i) = h_0(t) \cdot \exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right),$$

where $h_0(t)$ is the baseline hazard rate, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the unknown parameter vector. The regression coefficients β_j are estimated by maximizing a partial likelihood without having to specify the baseline hazard rate.

Penalized Cox regression model

We consider high-dimensional settings where the number of covariates p exceeds the sample size n. In this situation, the solution maximizing the Cox partial likelihood is not unique. One possibility to deal with this problem is to introduce a penalty term into the partial log-likelihood $l(\boldsymbol{\beta})$, referred to as regularization. This approach is also reasonable in $p < n$ settings since it considers collinearity among the predictors and helps to prevent overfitting. We use a lasso penalty [32, 33] that performs variable selection and yields a sparse model solution. The resulting maximization problem of the penalized partial log-likelihood is given by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ l(\boldsymbol{\beta}) - \lambda \cdot \sum_{j=1}^{p} |\beta_j| \right\}.$$

The parameter λ controls the strength of penalization and is optimized by tenfold cross-validation. For parameter estimation, we use the implementation in the R package `glmnet` [14].

Weighted Cox partial likelihood

In the standard unweighted partial likelihood, all patients contribute to the same extent to the estimation of the regression coefficients. This might not be desirable when the cohort is heterogeneous due to known subgroups that are associated with different prognosis. In this situation, it is reasonable to fit a separate Cox model for each subgroup. This can be done by using only the data from the subgroup of interest or by including information from the other subgroups. We include patients from all subgroups in the likelihood for one specific subgroup but assign them individual weights $w_i \geq 0, i = 1, \ldots, n$ to account for the heterogeneity in the data. The size of each weight determines to which extent the corresponding patient contributes to the estimation.

In accordance with Weyer and Binder [39], the weighted version of the partial log-likelihood is defined as

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i w_i \left[\boldsymbol{\beta}'\boldsymbol{x}_i - \log\left(\sum_{k=1}^{n} \mathbb{1}(t_i \leq t_k) w_k \exp\left(\boldsymbol{\beta}'\boldsymbol{x}_k\right) \right) \right].$$

Weyer and Binder [39] propose the use of fixed weights. The idea is to focus on a specific subgroup s of patients and assign each of these patients a weight of 1, while all other patients are down-weighted with a fixed weight $w \in (0, 1)$:

$$w_i = \begin{cases} 1, & \text{if } s_i = s \\ w, & \text{else.} \end{cases}$$

Standard subgroup analysis is based only on the patients in the subgroup of interest s, which corresponds to $w = 0$ for all patients not belonging to s. A combined model that

pools patients from all subgroups corresponds to $w = 1$ for all patients. Alternatively to the idea of Weyer and Binder [39], we propose to estimate individual weights for each patient from the training data. This approach is described in the following section.

Estimation of weights

Individual weights for each patient in each subgroup-specific likelihood can be estimated from the training data following the idea of Bickel et al. [3]. The weights match the joint distribution of all subgroups to the target distribution of a specific subgroup s, such that a patient who is likely to belong to the subgroup of interest receives a higher weight in the subgroup-specific model.

Assume the entire training data from all subgroups are summarized in the covariates $\boldsymbol{x}$ and a response $\boldsymbol{y}$. In time-to-event settings, the response y_i corresponds to the tuple (t_i, δ_i), with t_i the observed time until an event or censoring and δ_i the event indicator. Let $\ell(\boldsymbol{y}, f_s(\boldsymbol{x}))$ be an arbitrary loss function and $f_s(\boldsymbol{x})$ the predicted response based on the observed covariates in subgroup s. $f_s(\boldsymbol{x})$ should correctly predict the true response and thus minimize the expected loss with respect to the unknown joint distribution $p(\boldsymbol{y}, \boldsymbol{x}|s)$ for each subgroup s, given by $\mathrm{E}_{p(\boldsymbol{y},\boldsymbol{x}|s)}[\ell(\boldsymbol{y}, f_s(\boldsymbol{x}))]$. The following equation shows that this expected loss for each subgroup equals the expected weighted loss with respect to the joint distribution of the pooled data from all subgroups $p(\boldsymbol{y}, \boldsymbol{x})$

$$\begin{aligned}
\mathrm{E}_{p(\boldsymbol{y},\boldsymbol{x}|s)}[\ell(\boldsymbol{y}, f_s(\boldsymbol{x}))] &= \int p(\boldsymbol{y}, \boldsymbol{x}|s)\ell(\boldsymbol{y}, f_s(\boldsymbol{x}))\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x} \\
&= \int \frac{p(\boldsymbol{y}, \boldsymbol{x}|s)}{p(\boldsymbol{y}, \boldsymbol{x})} p(\boldsymbol{y}, \boldsymbol{x})\ell(\boldsymbol{y}, f_s(\boldsymbol{x}))\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x} \\
&= \mathrm{E}_{p(\boldsymbol{y},\boldsymbol{x})}\left[\frac{p(\boldsymbol{y}, \boldsymbol{x}|s)}{p(\boldsymbol{y}, \boldsymbol{x})}\ell(\boldsymbol{y}, f_s(\boldsymbol{x}))\right] \\
&= \mathrm{E}_{p(\boldsymbol{y},\boldsymbol{x})}[w_s(\boldsymbol{y}, \boldsymbol{x})\ell(\boldsymbol{y}, f_s(\boldsymbol{x}))].
\end{aligned}$$

The subgroup-specific weights for each patient are defined as

$$w_s(\boldsymbol{y}, \boldsymbol{x}) = \frac{p(\boldsymbol{y}, \boldsymbol{x}|s)}{p(\boldsymbol{y}, \boldsymbol{x})} = \frac{p(s|\boldsymbol{y}, \boldsymbol{x})}{p(s)}, \quad p(s) > 0.$$

The last equation shows that the weights can be expressed in terms of $p(s)$ and $p(s|\boldsymbol{y}, \boldsymbol{x})$. $p(s)$ can be estimated by the relative frequency of subgroup s in the overall training cohort, and $p(s|\boldsymbol{y}, \boldsymbol{x})$ can be considered as a multi-class classification problem [3]. We estimate $p(s|\boldsymbol{y}, \boldsymbol{x})$ by multinomial logistic regression or by random forest, using the implementation in the R packages `glmnet` [14] and `ranger` [43], respectively. Unlike Bickel et al. [3], we use tenfold cross-validation to estimate $p(s|\boldsymbol{y}, \boldsymbol{x})$ from the training data to prevent overfitting. As a result, for each subgroup, we obtain an n-dimensional vector of estimated individual weights.

Unlike the fixed weights by Weyer and Binder [39], our proposed estimated weights are not constrained to (0, 1) as the ratio $\frac{p(s|\boldsymbol{y},\boldsymbol{x})}{p(s)}$ can take values larger than 1. The R package `glmnet`, which we use to fit the weighted penalized Cox model, internally rescales the weights so that they add up to the sample size (see the vignette "An Introduction to glmnet"). However, normalizing the weights to range from 0 to 1 is not necessary as all individuals contribute to the likelihood with a certain weight and rescaling all weights in the likelihood would not change the estimated Cox model.

Prediction performance

Prediction performance of all Cox models is evaluated by Harrell's C-(concordance) index [17], implemented in the R package `Hmisc` [18]. The C-index is a measure of predictive discrimination and defined as the proportion of all usable pairs of patients with concordant predicted and observed survival times. For a concordant pair of patients, the survival time of the patient with larger risk score is known to be shorter than the survival time of the patient with lower risk score, such that the risk measure and the survival time lead to the same ordering of patients.

Let t_i, t_{i^*} be the observed survival times of patients i and i^*, and $\hat{r}(\boldsymbol{x}_i) = \hat{\boldsymbol{\beta}}' \boldsymbol{x}_i$, $\hat{r}(\boldsymbol{x}_{i^*}) = \hat{\boldsymbol{\beta}}' \boldsymbol{x}_{i^*}$ the corresponding risk scores (with $t_i, t_{i^*}, \boldsymbol{x}_i, \boldsymbol{x}_{i^*}$ corresponding to the test data and $\hat{\boldsymbol{\beta}}$ estimated from the training data). A pair (i, i^*) is considered concordant if $t_i \lessgtr t_{i^*} \Leftrightarrow \hat{r}(\boldsymbol{x}_i) \gtrless \hat{r}(\boldsymbol{x}_{i^*})$. The C-index is defined as

$$CI = \frac{1}{n_c} \sum_{\{i:\, \delta_i = 1\}} \sum_{\{i^*:\, t_{i^*} > t_i\}} \Big(\mathbb{1}(\hat{r}(\boldsymbol{x}_{i^*}) < \hat{r}(\boldsymbol{x}_i)) + \frac{1}{2}\mathbb{1}(\hat{r}(\boldsymbol{x}_{i^*}) = \hat{r}(\boldsymbol{x}_i)) \Big),$$

where n_c is the number of comparable pairs (i, i^*) that standardizes CI to [0, 1]. A patient pair is considered unusable, if both patients die at the same time, or both patients are censored, or if one is censored before the other one dies. $CI \approx 1$ stands for a very good prediction and values around 0.5 suggest a random prediction.

While Harrell's C-index is an easy to interpret and compute approach for quantifying the accuracy of prognostic survival models, it depends on the censoring distribution. To overcome this shortcoming, Uno et al. [37] introduce inverse probability censoring weights to the C-index to adjust for right censoring. Instead of evaluating the "overall" prediction accuracy, it can be of interest to quantify the discriminative ability at each time point

under consideration. In this situation, time-dependent ROC analysis can be used to distinguish at each time point $t > 0$ between patients having an event at or up to t and those having an event after t. The corresponding area under the time-dependent ROC curve provides an estimator of incidence/dynamic or cumulative/dynamic AUC for right-censored time-to-event data [19, 36].

Model fitting and evaluation

We compare our weighted approach with the standard (unweighted) models, i.e. the combined model and the subgroup model, as well as a weighted Cox model with fixed weights as proposed by Weyer and Binder [39]. In the latter, patients belonging to a certain subgroup are assigned a weight of 1 in the subgroup-specific likelihood, while all other observations are down-weighted with a constant weight $w \in (0, 1)$. For our proposed approach we compare three different classification methods for weights estimation with respect to prediction performance: Multinomial logistic regression with lasso (*lasso*) or ridge (*ridge*) penalty, and random forest (*rf*). All Cox models include a lasso penalty for variable selection. We compare the following Cox models concerning prediction performance. The italic expressions in parentheses denote the abbreviations of the models in the following analyses:

- Weighted model with estimated weights (*lasso*, *ridge*, *rf*)
- Weighted model with fixed weights ($w = 0.1, 0.2, \ldots, 0.9$)
- Standard subgroup model (*sub*), using only patients of a specific subgroup
- Standard combined model (*all*), using patients of all subgroups. The subgroup indicator is included as additional covariate.

Figure 1 provides a schematic representation of the analysis pipeline. First, we randomly generate training data sets for model fitting and test data sets for model evaluation and repeat this procedure 100 times. In the application example, we repeatedly randomly split the complete data into training (with proportion 0.632) and test sets. We perform subsampling stratified by subgroup and event indicator, to take different subgroup sizes and

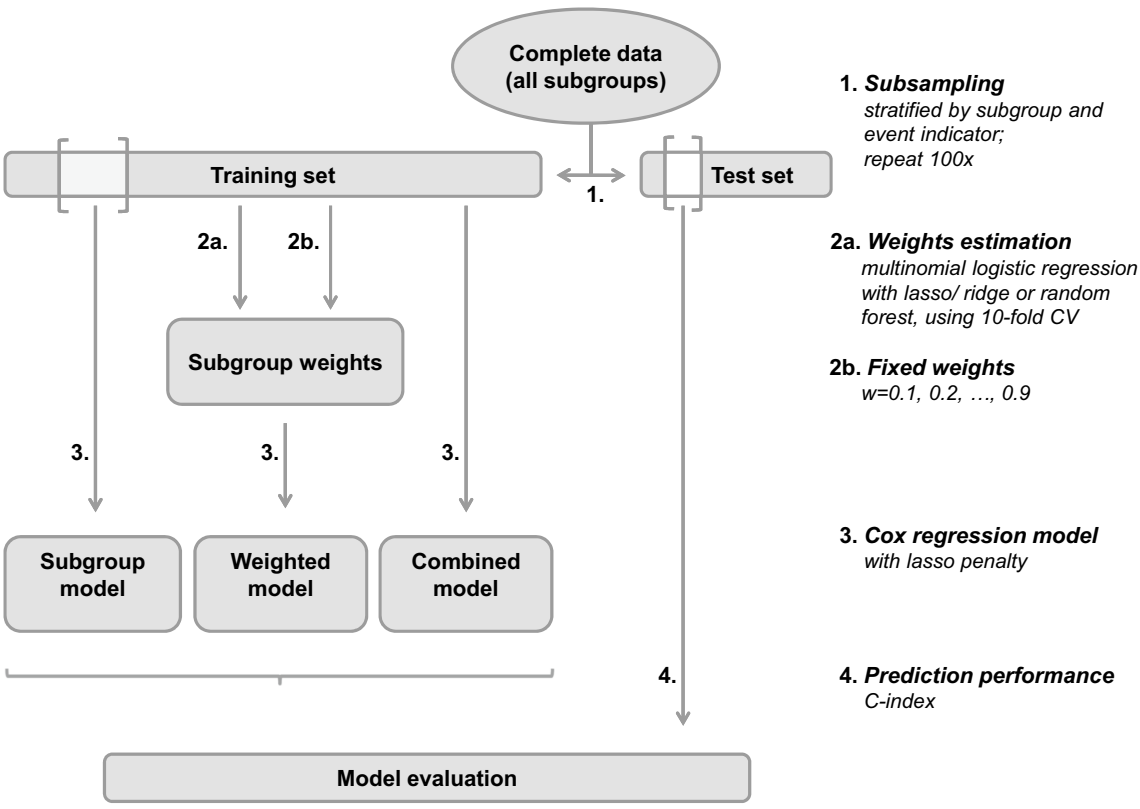

Fig. 1 Simulation set-up. Analysis pipeline for the simulation study; Brighter regions in the training and test set indicate the observations of the subgroup

censoring proportions into account. In the simulation study, we repeatedly randomly generate independent training and test sets of the same size and with the same distribution parameters. Second, we estimate individual subgroup weights from the training data using different classification methods and 10-fold cross-validation (CV). Next, we fit the combined and weighted Cox models based on the training data of all subgroups, while the standard subgroup model is based on the training data of the respective subgroup only. Finally, we evaluate the prediction performance of the estimated Cox models with respect to a certain subgroup using only the test data of this particular subgroup. The R package `batchtools` [23] is used for parallelization and the R package `mlr` [6] is used as a framework for weights estimation, Cox model fitting and evaluation by the C-index.

Results of the simulation study

Simulated data

We simulate four subgroups (1A, 1B, 2A, 2B) of equal size n from two differently distributed groups denoted by the index $g^* = 1, 2$: group 1 including subgroups 1A and 1B, and group 2 including subgroups 2A and 2B. Within each group we use the same parameters for the simulation of the data. We simulate the survival data from a Weibull distribution according to Bender et al. [1], with scale parameter η_{g^*} and shape parameter κ_{g^*} estimated from two independent lung cancer cohorts (GSE37745 and GSE50081). For this purpose, we compute survival probabilities at 3 and 5 years using the Kaplan-Meier estimator for both lung cohorts separately. The corresponding probabilities are 57% and 75% for 3-years survival, and 42% and 62% for 5-years survival, respectively. Individual event times in group g^* are simulated as

$$T_{g^*} \sim \left(-\frac{\log(U)}{\eta_{g^*} \exp(\boldsymbol{x}_{g^*} \boldsymbol{\beta}_{g^*})} \right)^{1/\kappa_{g^*}}, \quad U \sim \mathcal{U}[0,1],$$

with true effects $\boldsymbol{\beta}_{g^*} \in \mathbb{R}^p$, $g^* = 1, 2$. We randomly draw noninformative censoring times C_{g^*} from a Weibull distribution with the same parameters as for the event times, resulting in approximately 50% censoring rates in both groups. The individual observed event indicators and times until an event or censoring are defined as $\delta_{g^*} = \mathbb{1}(T_{g^*} \leq C_{g^*})$ and $t_{g^*} = \min(T_{g^*}, C_{g^*})$.

For each subgroup we simulate p uncorrelated (genetic) covariates $\boldsymbol{x}_{g^*}$ from a multivariate normal distribution with mean vector $\boldsymbol{\mu}_{g^*}$ and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{I}_{p \times p}$. In previous simulation studies we compared the results of different covariance structures, including realistic dependence structures estimated from real gene expression data, but found no remarkable differences [26]. Elements of $\boldsymbol{\mu}_{g^*}$ are defined by a linear function with parameter $\epsilon \in [0, 1]$ that reflects the degree of similarity between the two groups. We assign $\mu = 4 + 4 \cdot \epsilon$ to genes with a strong effect on the outcome ($|\beta| = 1$), $\mu = 4 + 2 \cdot \epsilon$ corresponds to genes with a moderate effect ($|\beta| = 0.5, 0.75$), and $\mu = 4$ to genes with a weak or no effect ($|\beta| = 0, 0.25$). This choice relies on the assumption that prognostic genes have a higher expression level than noise genes. The magnitude of μ is chosen following real gene expression data, where the expression values typically range from 4 to 12 after transformation to $\log_2$ scale.

In all simulated scenarios, we assume the first 12 genes to be prognostic in at least one of the two groups, with corresponding effects given in Table 1. We include subgroup-specific effects (genes 1 to 4), opposite effects (genes 5 and 6), effects in the same direction but of different size (genes 7 and 8), and joint effects of varying sizes (genes 9 to 12). We choose these effects with alternate signs so that they sum up to zero, resulting in reasonable simulated survival times. In settings with $p > 12$, we assume all remaining genes to represent noise and being unrelated to the survival times in both groups ($\beta_{13} = \cdots = \beta_p = 0$).

In our simulation study we focus on high-dimensional settings where the sample size n is small compared to the number of covariates (genes) p, a typical characteristic of gene expression data. Table 2 shows all parameters tested

Table 2 Parameter combinations in the simulation study

Parameter	Values (per subgroup)
n	20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 500, 1000
p	12, 100, 200
ϵ	0, 0.1, 0.2, 0.3, 0.4, 0.5, 1

All parameters tested in the simulation study with their respective values, resulting in 252 different combinations in total

Table 1 Effects in the simulation study

Gene	1	2	3	4	5	6	7	8	9	10	11	12
β_1	1	1	0	0	−0.5	0.5	0.75	0.25	−1	−1	−0.75	−0.25
β_2	0	0	1	1	0.5	−0.5	0.25	0.75	−1	−1	−0.75	−0.25

Effects of the first 12 genes for the simulation of survival outcome

in the simulation study with their respective values, resulting in 252 different combinations in total.

Weights estimation

Our proposed subgroup model uses patients from all subgroups for training but assigns them individual weights in the Cox partial likelihood based on their subgroup membership. Weights for a specific subgroup are estimated by the individual predicted probabilities of belonging to this subgroup, obtained by classification, divided by the subgroup proportion. Thus, a patient who is likely to belong to the subgroup of interest receives a higher weight in the subgroup-specific likelihood. We compare three different classification methods that are appropriate for multi-class problems and high-dimensional covariates with respect to their predictive quality and their ability to discriminate between differing subgroups.

Figure 2 displays boxplots of the estimated weights for subgroup 1A across all training sets in two selected simulation scenarios with $\epsilon = 0$ and $\epsilon = 0.5$. The x-axis represents the true subgroup membership of each observation, and the y-axis the individual weights estimated by random forest (*rf*) for subgroup 1A. Results of all three classification methods (*lasso*, *ridge*, *rf*) are relatively similar, altough *rf* tends to perform best for small ϵ and n, whereas for large sample size the discriminative ability of *lasso* and *ridge* is slightly better. The largest difference in results is obtained for different values of ϵ. When all subgroups are very similar ($\epsilon = 0$), multi-class classification fails to distinguish the two differing groups. All observations are assigned a weight of approximately 1 in all subgroup models, similar to the standard combined Cox model. The corresponding area under the ROC curve (AUC) for the distinction between group 1 and 2 (computed based on test data and cross-validated training data) is approximately 0.5, indicating that prediction performance is not much better than random (see Additional file 1: Figure S1). Increasing values of ϵ, meaning larger differences between the two groups, lead to improved prediction performance (see Additional file 1: Figure S1), and for $\epsilon = 0.5$ classification succeeds in providing an almost perfect separation between both groups with AUC ≈ 1. Larger sample size n and smaller number of covariates p also result in better prediction performance.

Parameter estimation and prediction performance

Weighted Cox models, including fixed or estimated weights (with different classification methods for weights estimation), are compared to the standard combined and subgroup model, first by estimated regression coefficients and second by prediction performance.

Figure 3 shows scatterplots of the mean estimated regression coefficients of the first 12 prognostic genes in group 1 (mean across all training sets and subgroups 1A and 1B) for simulated data with $n = 50$, $p = 12,100$ and $\epsilon = 0, 0.5$. For $\epsilon = 0$, the combined and weighted model with estimated weights provide very similar results, as expected. They identify joint effects better than the subgroup model when the sample size is small ($n \leq p$) and otherwise equally well. However, the subgroup model estimates subgroup-specific effects better, especially for increasing sample size, whereas the other two model approaches tend to average effects across all subgroups. For larger values of ϵ the estimated weights model detects subgroup-specific effects increasingly better than the

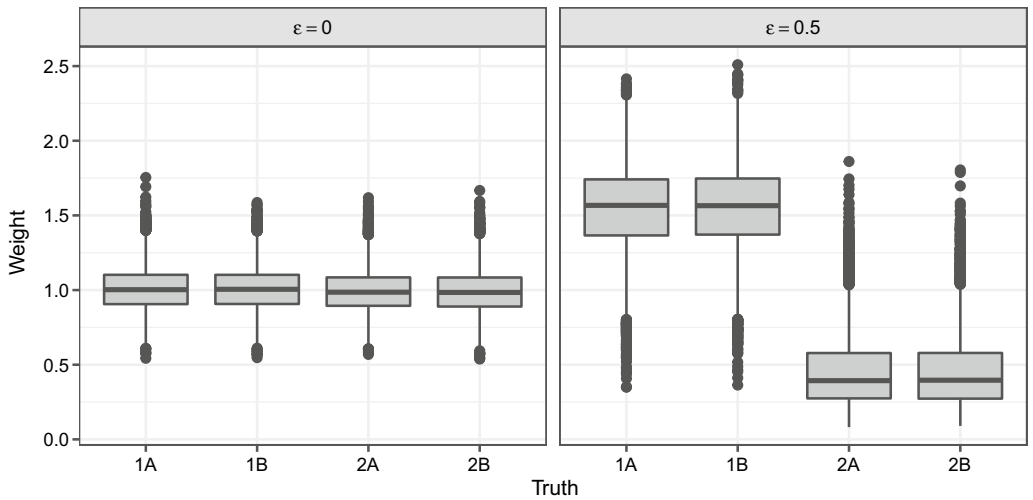

Fig. 2 Estimated weights in the simulation. Estimated weights for subgroup 1A obtained by random forest based on simulated training data with $p = n = 100$ and $\epsilon = 0, 0.5$

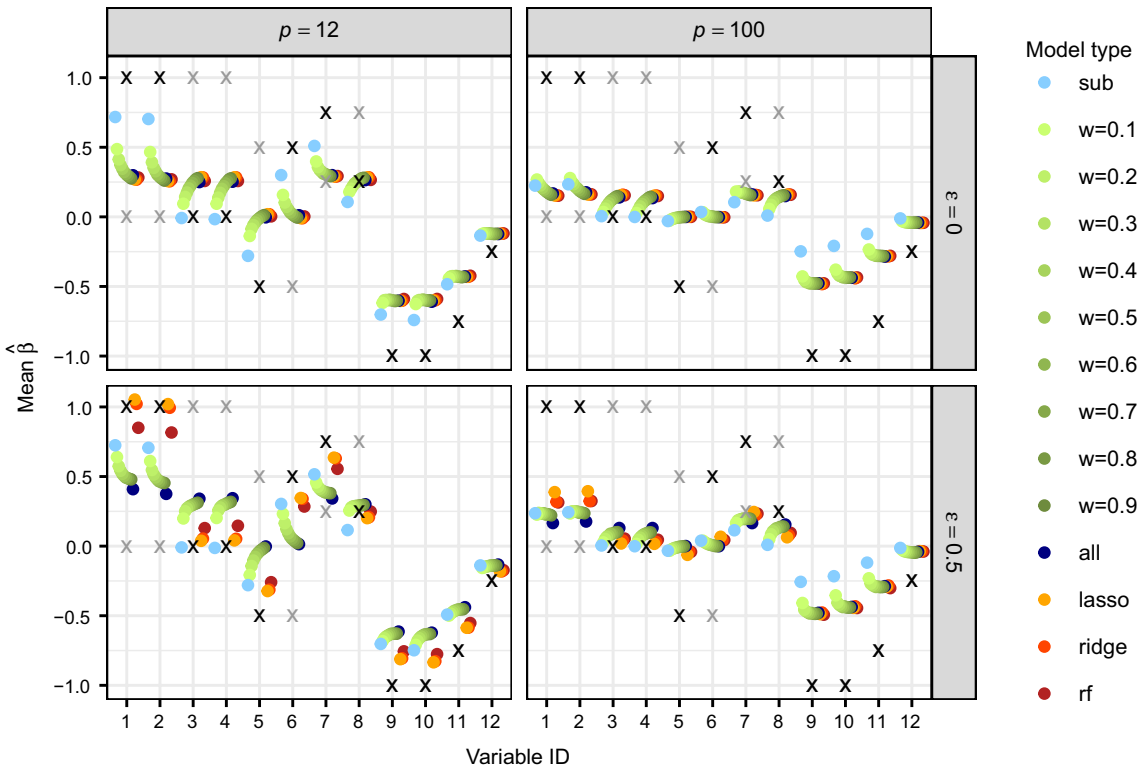

Fig. 3 Estimated effects in the simulation. Mean estimated regression coefficients of the first 12 prognostic genes in all Cox models for group 1 (averaged across all simulated training data sets, regardless of whether a gene is selected in individual splits or not, and subgroups 1A and 1B) for $n = 50$ and $\epsilon = 0, 0.5$. The symbol 'x' represents the true simulated effects β of both groups: in black the group of interest (here group 1) and in grey the other group

combined model, and similarly well or even better than the standard subgroup model when sample size is small. Results for fixed weights lie between the subgroup model and the combined model.

These findings agree with the corresponding mean inclusion frequencies (MIFs), defined as the proportion of training data sets in which a specific covariate j is included in the model ($\hat{\beta}_j \neq 0$). For small sample size, the MIFs of the standard combined model and the estimated weights approach are larger than the MIFs of the standard subgroup model. This has a positive impact on the detection of joint effects, but subgroup-specific effects that are present in only one group may be more often erroneously selected in the other group. For increasing sample size the MIFs of all models also increase. For larger values of ϵ, the MIFs of the estimated weights model move closer to the MIFs of the subgroup model regarding subgroup-specific effects and are still similar to the combined model for joint effects.

Finally, we assess the prediction performance of all Cox models in terms of the C-index. High values of the C-index (close to 1) indicate a good predictive performance, whereas 0.5 corresponds to random prediction. Figure 4 displays the mean C-index (averaged across all test sets and subgroups). For $\epsilon = 0$ the combined model and the weighted model with estimated weights exhibit a very similar predictive ability, that is better compared to the subgroup model when sample size is small. However, when the sample size increases the subgroup model outperforms the other methods. For larger values of ϵ, the estimated weights approach performs best when the sample size is small and otherwise equally well as the subgroup model. Estimated weights by *lasso* and *ridge* improve in comparison to *rf* (random forest) for larger n. Unsurprisingly, the prediction performance of fixed weights lies between the standard combined model and the subgroup model. Mean C-index values for all 252 simulation scenarios and all 14 Cox model types can be found in Additional file 1: Table S2.

Results of the application to NSCLC cohorts

We apply all methods presented in the previous section to the following four non-small cell lung cancer (NSCLC) cohorts comprising in total $n = 635$ patients with available overall survival endpoint and Affymetrix microarray gene expression data: GSE29013 ($n = 55$, 18 events),

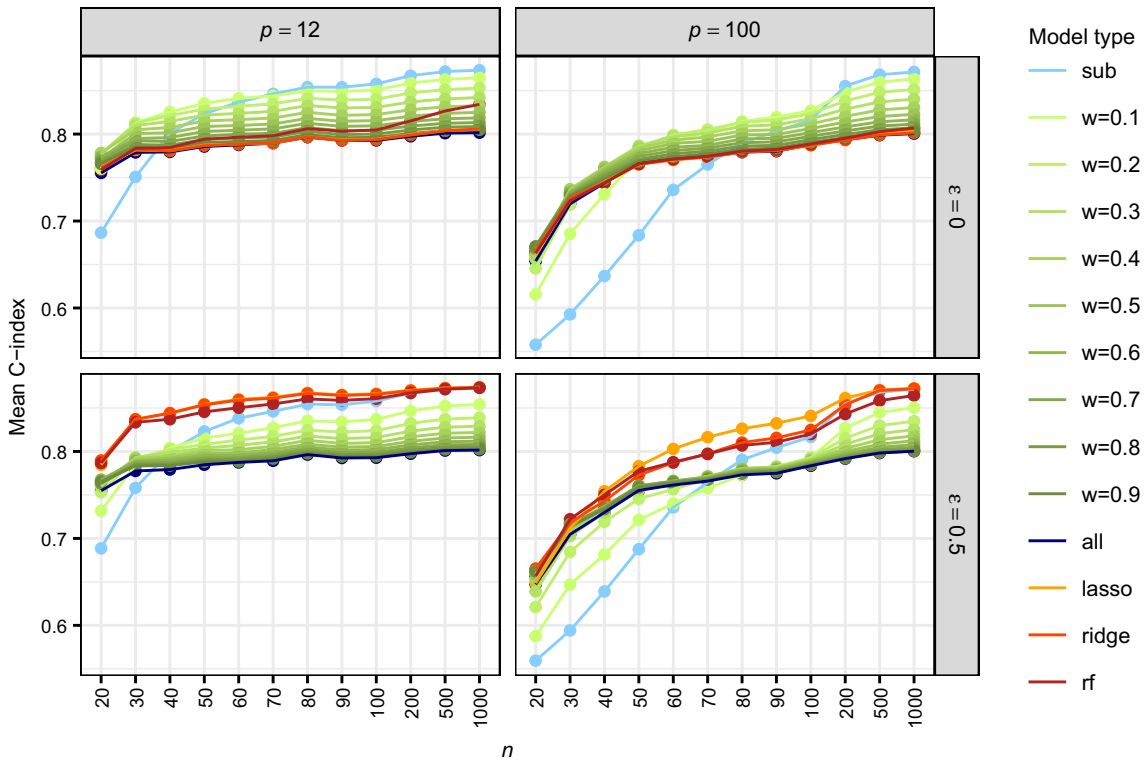

Fig. 4 Prediction performance in the simulation. Mean C-index, averaged across all test data sets and subgroups, for $\epsilon = 0, 0.5$

GSE31210 ($n = 226$, 35 events), GSE37745 ($n = 194$, 143 events), and GSE50081 ($n = 160$, 65 events). For the analysis, we use the total number of $p = 54,675$ genetic covariates measured in each cohort, as well as two pre-selected reduced gene sets. One gene filter is defined by the $p = 1000$ features with the highest variance in gene expression values across all four cohorts, referred to as top-1000-variance genes. The second gene filter is a literature-based selection of $p = 3429$ prognostic genes. More details on the data description and preprocessing can be found in Additional file 1.

Weights estimation

In the following, we consider four lung cancer cohorts as subgroups. We compare the estimated weights using three classification methods (*lasso*, *ridge*, *rf*) and three different pre-specified sets of genes (gene filters): all available genes ($p = 54,675$), top-1000-variance genes ($p = 1000$), and a literature-based selection of prognostic genes ($p = 3429$). Since all results are very similar, we only show them exemplary for the top-1000-variance genes and *rf* in Fig. 5. Boxplots of the estimated weights suggest that subgroups are very different from each other. Patients belonging to the subgroup of interest receive a relatively large weight in the respective subgroup-specific model, while the contribution of all other subgroups is close to zero. This resembles the standard subgroup model.

The estimated weights for patients from GSE29013 are the highest in the corresponding subgroup model for GSE29013, and much higher compared to the other cohorts. The reason is that GSE29013 is by far the smallest subgroup and when the estimated probabilities of belonging to $s = \text{GSE29013}$ $\hat{p}(s|\boldsymbol{y}, \boldsymbol{x})$ are divided by the very small relative frequency $\hat{p}(s)$, the resulting probability ratio corresponding to the weights gets very large.

Parameter estimation and prediction performance

All analyses are based on probe set level of gene expression data, but for the illustration of the parameter estimates in the Cox models, probe set IDs are translated into gene symbols using the R/Bioconductor annotation packages `hgu133plus2.db` [10] and `AnnotationDbi` [28]. In case of missing gene symbols, original probe set IDs are retained. Corresponding gene annotation is retrieved from the Ensembl website [44] to obtain gene-specific information on encoded proteins, related pathways, Gene Ontology (GO) annotations, associated diseases, and related articles in PubMed. This

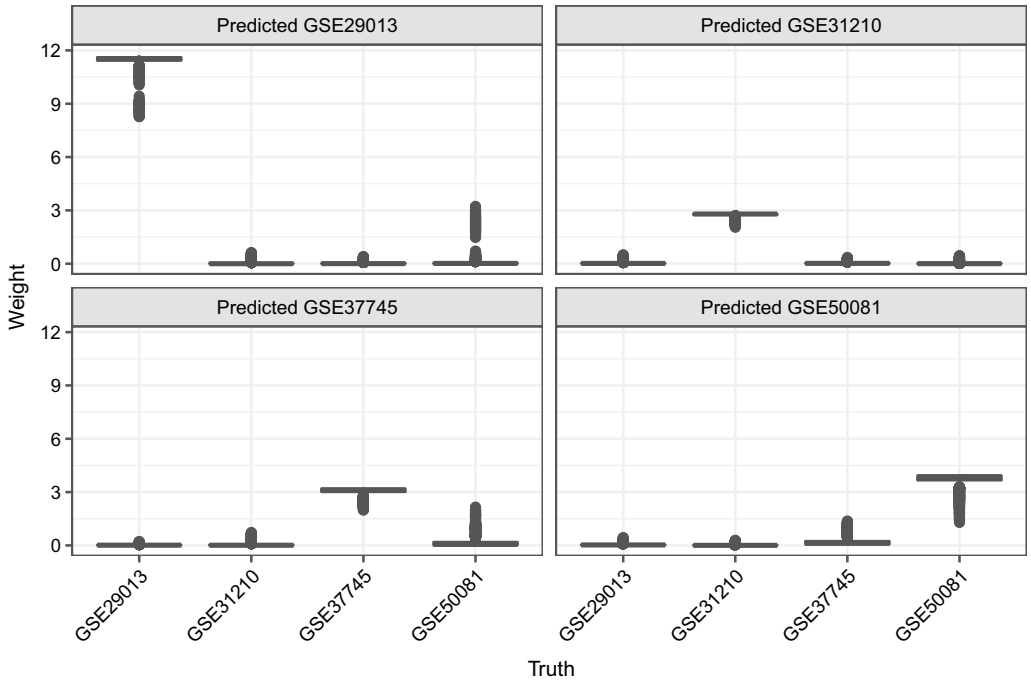

Fig. 5 Estimated weights in the application. Estimated weights for all lung cancer cohorts using random forest and the top-1000-variance genes as gene filter

information is retrieved from the NCBI Gene [9] and GeneCards [16] databases.

Figure 6 shows, separately for each subgroup, the mean estimated regression coefficients of the most frequently selected top-1000-variance genes (genes with a mean inclusion frequency (MIF) larger or equal than 0.5 in any model type). Eight genes are in the overlap of all subgroups, among them immune-related genes (DEFB1, AOC1, JCHAIN) as well as genes (215780_s_at/SET, SPP1) that were reported in the literature to be associated with different types of cancer. Often they are most frequently selected by the combined model and the weighted model with large fixed weights. Subgroup-specific genes with strong effects on overall survival and high MIFs in the proposed weighted model involve the following cancer-related genes: ADH1C, BMP5, LCN2 and PLOD2 in GSE31210, CST1 in GSE37745, as well as AREG and COL4A3 in GSE29013.

For the other two gene filters (prognostic genes and all genes), parameter estimates of the most stable genes in all Cox models are displayed in Additional file 1: Figures S2 and S3. Cox models including all genes identify fewer genes compared to the other gene filters which is likely caused by the large number of noise genes. There are two cancer-related genes most frequently selected across all subgroups by the combined model and the weighted model with large fixed weights: CPNE8 and SPP1 MIFs and estimated regression coefficients of the subgroup model and the proposed weighted model are mainly close to zero, except for PTGER3 in GSE31210. PTGER3 induces tumor progression in different cancer types including adenocarcinoma of the lung. This may explain the specific association with GSE31210 being the only subgroup comprising exclusively adenocarcinoma.

Interestingly, almost all selected genes are either in the overlap of all subgroups or specific for only one subgroup. There are hardly any genes selected by two or three subgroups, which may be due to the fact that these lung cancer studies are heterogeneous (see Additional file 1: Figure S4). There is one gene (SPP1) that is in the overlap of all four subgroups and all three gene sets. SPP1—also known as Osteopontin (OPN)—is involved in inflammatory response, osteoblast differentiation for bone formation and attachment of osteoclasts to the mineralized bone matrix for bone resorption. Further, SPP1 is associated with several malignant diseases and prognosis in NSCLC.

Finally, all Cox models are compared with regard to prediction performance. In Fig. 7 results of the C-index across all test sets are shown for the top-1000-variance genes. The combined model and fixed weights of increasing size tend to have the highest predictive accuracy, while the estimated weights approach and the standard subgroup model perform similarly. Particularly in the

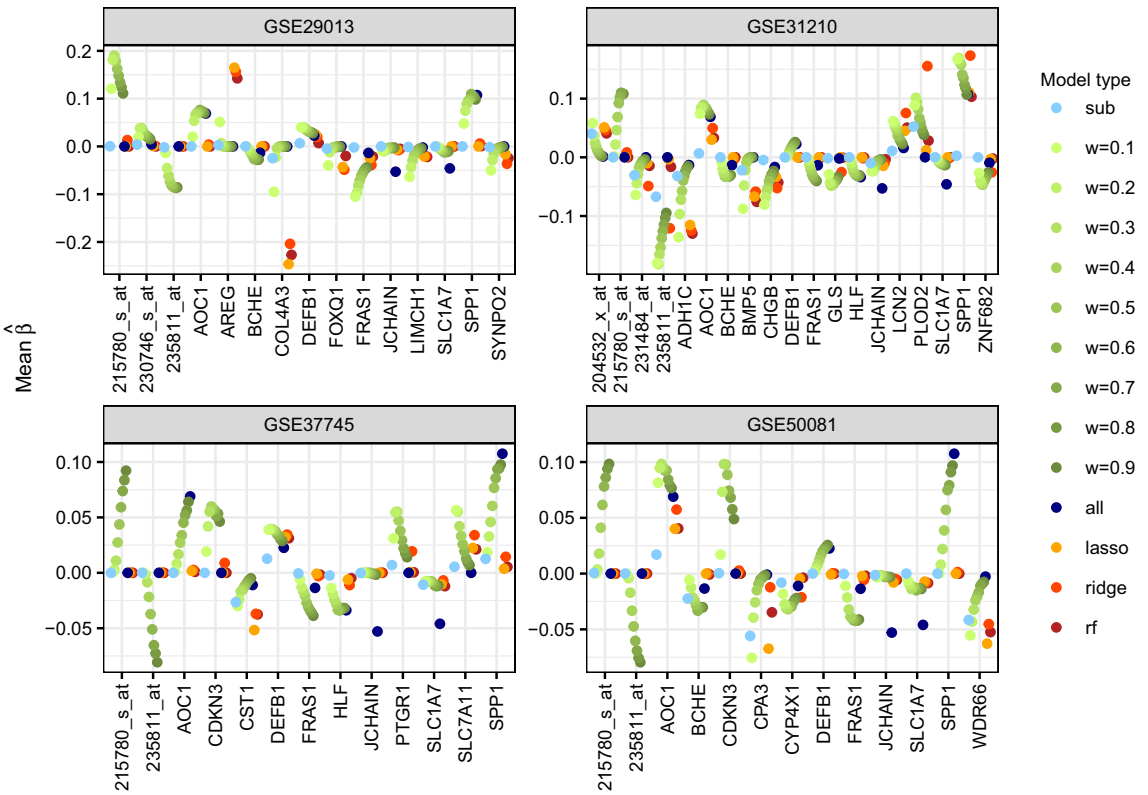

Fig. 6 Estimated effects in the application. Different types of Cox models including the top-1000-variance genes as covariates. Mean estimated regression coefficients of selected genes (averaged across all training sets regardless of whether a gene is selected in individual splits or not). For each subgroup genes with a mean inclusion frequency larger or equal than 0.5 in any model type are selected

subgroup model for GSE29013 the performance of the estimated weights differs from the fixed weights because the estimated weights for GSE29013 are much higher compared to those for all other subgroups, which is similar to the standard subgroup model. The corresponding boxplots of the C-index for the prognostic gene filter and all genes are shown in Additional file 1: Figures S5 and S6. Random forest tends to be the best classification method in combination with prognostic genes and all genes, whereas ridge tends to perform slightly better than the other classification methods along with top-1000-variance genes. However, overall prediction performance is mostly moderate and not much better than random.

Discussion

We have focused on three major objectives: prediction of a patient's survival, selection of important covariates, and consideration of heterogeneity in data due to pre-known subgroups of patients. Specifically, we have aimed at estimating a separate risk prediction model for each subgroup using patient-level training data from all available subgroups and individually weighting patients according to their similarity to the subgroup of interest. Our approach should correctly identify common as well as subgroup-specific effects and have improved prediction accuracy over standard approaches. As standard approaches, we consider standard subgroup analysis, including only patients from the subgroup of interest, and standard combined analysis that simply pools patients from all subgroups.

We have proposed a Cox model with lasso penalty for variable selection and a weighted version of the partial likelihood that includes patients from all subgroups but with individual weights. This allows sharing information between subgroups to increase power when this is supported by the data, meaning that subgroups are similar in their covariates and survival outcome. Weights for a specific subgroup are estimated by classification and cross-validation on the training data from all subgroups, such that they represent the probability of belonging to that subgroup given the observed covariates and survival outcome. These predicted conditional probabilities are divided by the a priori probability of the respective subgroup to obtain the subgroup-specific weights for each patient. Patients who fit well into the subgroup of interest receive higher weights in the subgroup-specific model.

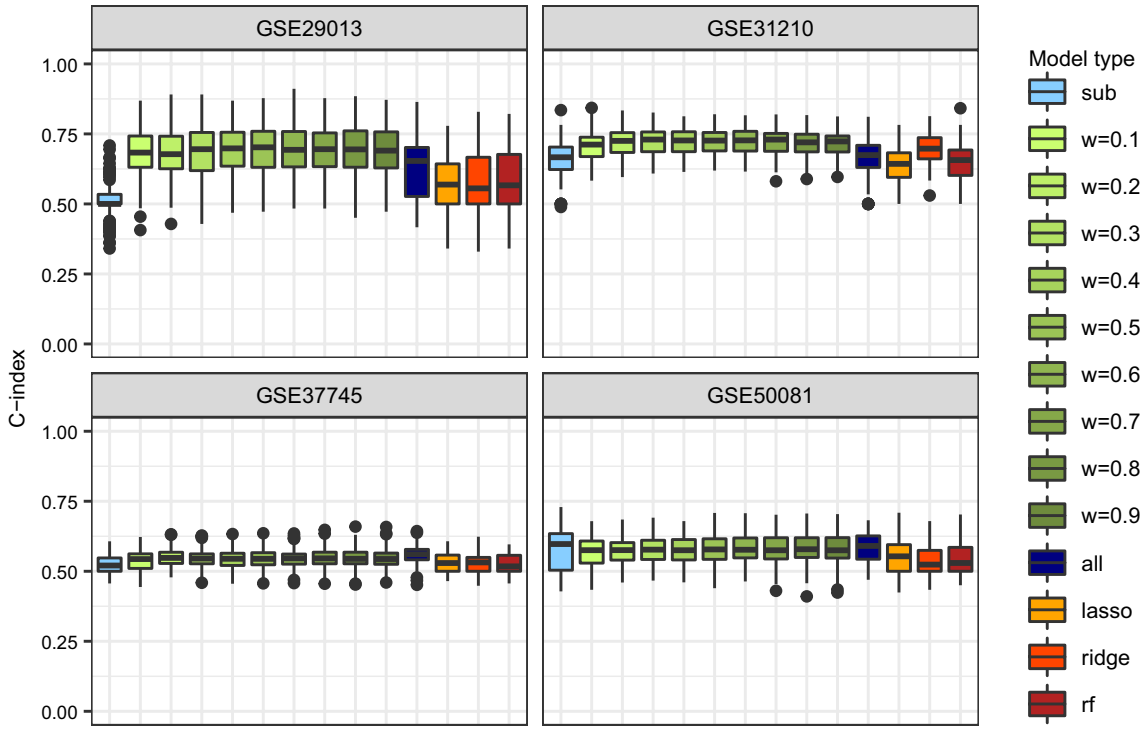

Fig. 7 Prediction performance in the application. Different types of Cox models including the top-1000-variance genes as covariates. Boxplots of C-index based on all test sets for the prediction of each subgroup

The estimated subgroup-specific model can then be applied to the test data from the corresponding subgroup to obtain predictions for that subgroup. Alternatively to our individual weights, one could restrict the model to the case where all weights for the patients from a subgroup must be the same [29, 39].

We have considered three different classification methods for weights estimation (multinomial logistic regression with lasso or ridge penalty and random forest), and, based on simulated data and on real data, we have compared our proposed weighted Cox model to both standard Cox models (combined and subgroup), as well as a weighted Cox model with different fixed weights as proposed by Weyer and Binder [39]. Observations belonging to a certain subgroup were assigned a weight of 1 in the subgroup-specific likelihood, while all other observations were down-weighted with a constant weight $w \in \{0.1, 0.2, \ldots, 0.9\}$.

Simulation results have shown that when subgroups were very similar and hardly distinguishable from each other in terms of their covariate values and only had a few different subgroup-specific effects, classification methods failed to discriminate between distinct subgroups and all observations were assigned a weight around one corresponding to the standard combined model. In this situation, results of the combined model and the proposed weighted model were very similar as intended. Both models had better prediction performance and larger power to correctly identify joint effects than the standard subgroup model when the sample size was small ($n \leq p$). The potential bias introduced in the estimation of subgroup-specific effects (tendency to average subgroup-specific effects across subgroups) is, however, not very likely in the situation of very similar subgroups. For increasing sample size, the standard subgroup model outperformed the other models regarding prediction and selection accuracy, in particular in terms of unbiased estimation of subgroup-specific effects.

When differences between subgroups became larger, classification succeeded in discriminating between different subgroups, and our proposed weighted model improved over the combined model in correctly identifying subgroup-specific effects and resulted in higher prediction accuracy. It clearly outperformed the standard subgroup model when the sample size was low, and otherwise performed similarly well. Results with fixed weights, as expected, always lay between the standard subgroup model and the combined model. However, they cannot flexibly adapt to different degrees of heterogeneity between subgroups as our proposed estimated weights do.

In the application example, we considered four lung cancer studies as subgroups comprising overall survival outcome, and gene expression data as covariates. Three different gene filters were used: all available genes, top-1000-variance genes, and a literature-based selection of prognostic genes. The real data application demonstrated the case of strongly differing subgroups where adding data from other subgroups is not appropriate as reflected by the small estimated weights. Our proposed weighted approach resembled the standard subgroup model, where only the subgroup of interest is assigned a high weight and all other subgroups have weights close to zero. The results of all three classification methods were similar. Prediction performance of Cox models indicated that logistic regression with ridge penalty and top-1000-variance genes outperformed the other two classification methods, while random forest tended to perform best in combination with all genes and with prognostic genes. However, the prediction performance of all Cox models was mainly moderate and not much better than random prediction. The combined model and the weighted model with fixed weights of increasing size tended to have slightly higher predictive accuracy, while the estimated weights approach and the standard subgroup model performed similarly. Genes identified most frequently by the former models were often present in all subgroups and some of them were reported in the literature to be associated with prognosis in various cancers. However, the corresponding estimated regression coefficients were often relatively small suggesting weak effects on survival outcome. Few candidate genes with reported cancer relation and relatively strong subgroup-specific effects were selected most frequently by either the subgroup model or the proposed weighted model.

A major reason for the overall moderate prediction accuracy in the application example may be that the present lung cancer studies are too heterogeneous. On the one hand, they comprise different histological subtypes that are known to be associated with a different prognosis. One could think of using only patients belonging to the same histological subtype such as adenocarcinoma. However, this would make the sizes of the patient subgroups even smaller. On the other hand, tissue processing and RNA extraction for generating gene expression data as well as patient inclusion criteria vary between studies. In GSE29013 genome-wide expression profiling was based on formalin-fixed paraffin-embedded (FFPE) tissues rather than fresh frozen tissues like in GSE37745 and GSE50081, which might influence expression levels. GSE31210 and GSE50081 include only patients with stage I and II, and GSE31210 is additionally restricted to lung adenocarcinomas.

In Madjar [26] we studied the influence of further parameters for weights estimation on prediction performance: the inclusion of interactions between genomic covariates and survival time in the classification model, as well as replacement of the survival time by the Nelson–Aalen estimator of the cumulative hazard rate in the set of covariates in the classification model. The latter was proposed by White and Royston [40] in the context of multiple imputation. We also considered a simulation with uneven sample sizes across subgroups and compared standard classification without sampling techniques with two oversampling techniques (random oversampling and synthetic minority oversampling technique). Oversampling increases the sample size of the small subgroup so that it is balanced with respect to the other subgroups. However, we found no considerable influence of the further parameters for weights estimation on prediction performance and also oversampling seemed to have no effect. Simulations with uneven sample sizes showed that the predicted probabilities of belonging to a specific subgroup $\hat{p}(s|\boldsymbol{y},\boldsymbol{x})$ were smaller for the subgroup with smaller sample size compared to the other subgroups having the same large sample size. However, this effect was compensated for when $\hat{p}(s|\boldsymbol{y},\boldsymbol{x})$ was divided by the relative frequency of each subgroup $\hat{p}(s)$ to obtain the weights ratio. This resulted in similar prediction accuracies for all subgroups, whereas the standard subgroup model clearly showed a worse prediction performance for the small subgroup.

We make the important assumption that subgroups are pre-known with the subgroup affiliation of each patient being unique and fixed, which is generally the case when patients from different clinical centers are considered. However, in situations with unknown subgroups the latent subgroup structure would first need to be determined using methods such as clustering. A wide variety of approaches have been proposed for the clustering of molecular data [13, 27, 42] with extensions to sparse clustering [30, 41] and integrative clustering of multiple omics data types [11, 22].

Conclusions

Predicting cancer survival risk based on high-dimensional molecular measurements for patients combined from heterogenous subgroups/cohorts is an important problem. The central motivation and idea of our proposed approach is to improve the prediction for a specific selected subgroup when also data from other subgroups are available, however, when it is not a priori clear which other subgroups can help to improve the prediction for the subgroup of interest. By adding data from other subgroups in a penalized weighted Cox model we aim at increasing the power through larger sample size compared to the classical subgroup analysis that ignores the information from all other individuals. Weights are based

on the probability of belonging to the subgroup of interest and are estimated from the (training) data instead of having to determine them a priori. In the situation of small sample sizes, simulation results clearly demonstrated the benefit of our proposed approach, suggesting that incorporating information from other subgroups in the estimation of a subgroup-specific risk model can improve the prediction performance and variable selection accuracy over standard approaches.

Abbreviations

MBO: Model-based optimization; C-index: Concordance index; CV: Cross-validation; AUC: Area under the ROC curve; MIF: Mean inclusion frequency; NSCLC: Non-small cell lung cancer.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-021-01698-1.

Additional file 1: Additional supporting information referenced in the Results Sections: Description of the NSCLC data preprocessing, Supplementary Figures 1-6, Supplementary Tables 1-2).

Acknowledgements

Not applicable.

Author's contributions

KM implemented the analyses, generated the results and wrote the manuscript. KM and JR contributed to the study design and the interpretation of results. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The preprocessed lung cancer data analyzed in this paper and the R code implementing our method are publicly available on GitHub, https://github.com/KatrinMadjar/WeightedCoxRegression.git.

Declarations

Ethics approval and consent to participate

Only published data that are publicly available online were used in this paper.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 27 January 2021 Accepted: 23 November 2021
Published online: 07 December 2021

References

1. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Stat Med. 2005;24(11):1713–23. https://doi.org/10.1002/sim.2059.
2. Bergersen LC, Glad IK, Lyng H. Weighted lasso with data integration. Stat Appl Genet Mol Biol. 2011. https://doi.org/10.2202/1544-6115.1703.
3. Bickel S, Bogojeska J, Lengauer T, Scheffer T. Multi-task learning for HIV therapy screening. In: Proceedings of the 25th international conference on machine learning. ICML '08, pp. 56–63. ACM, New York, USA (2008). https://doi.org/10.1145/1390156.1390164.
4. Binder H, Schumacher M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC Bioinform. 2008;9:14. https://doi.org/10.1186/1471-2105-9-14.
5. Binder H, Müller T, Schwender H, Golka K, Steffens M, Hengstler JG, Ickstadt K, Schumacher M. Cluster-localized sparse logistic regression for SNP data. Stat Appl Genet Mol Biol. 2012. https://doi.org/10.1515/1544-6115.1694.
6. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, Casalicchio G, Jones ZM. mlr: machine learning in r. J Mach Learn Res. 2016;17(170):1–5.
7. Bogojeska J, Lengauer T. Hierarchical Bayes model for predicting effectiveness of HIV combination therapies. Stat Appl Genet Mol Biol. 2012. https://doi.org/10.1515/1544-6115.1769.
8. Boulesteix A-L, De Bin R, Jiang X, Fuchs M. IPF-LASSO: integrative l_1-penalized regression with penalty factors for prediction based on multi-omics data. Comput Math Med. 2017; 2017: Article ID 7691937. https://doi.org/10.1155/2017/7691937.
9. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, Tolstoy I, Tatusova T, Pruitt KD, Maglott DR, Murphy TD. Gene: a gene-centered information resource at NCBI. Nucl Acids Res. 2015;43(D1):36–42. https://doi.org/10.1093/nar/gku1055.
10. Carlson M. Hgu133plus2.db: affymetrix human genome U133 Plus 2.0 array annotation data (chip Hgu133plus2). 2016. R package version 3.2.3
11. Chalise P, Koestler DC, Bimali M, Yu Q, Fridley BL. Integrative clustering methods for high-dimensional molecular data. Transl Cancer Res. 2014;3(3):202–16. https://doi.org/10.3978/j.issn.2218-676X.2014.06.03.
12. Cox DR. Regression models and life-tables. J R Stat Soc Ser B (Methodol). 1972;34(2):187–220.
13. de Souto MC, Costa IG, de Araujo DS, Ludermir TB, Schliep A. Clustering cancer gene expression data: a comparative study. BMC Bioinform. 2008;9:497. https://doi.org/10.1186/1471-2105-9-497.
14. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw, 2010;33(1), 1–22. Vignette "An Introduction to glmnet": https://glmnet.stanford.edu/articles/glmnet.html (accessed September 2021)
15. Gade S, Porzelius C, Fälth M, Brase JC, Wuttig D, Kuner R, Binder H, Sültmann H, Beißbarth T. Graph based fusion of miRNA and mRNA expression data improves clinical outcome prediction in prostate cancer. BMC Bioinform. 2011;12:488. https://doi.org/10.1186/1471-2105-12-488.
16. GeneCards: GeneCards®: the human gene database. https://www.genecards.org. Accessed: June 2018.
17. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361–87.
18. Harrell Jr FE, with contributions from Charles Dupont, many others. Hmisc: Harrell Miscellaneous. 2018. R package version 4.1-1. https://CRAN.R-project.org/package=Hmisc
19. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics. 2005;61(1):92–105. https://doi.org/10.1111/j.0006-341X.2005.030814.x.
20. Hothorn T, Bühlmann P. Model-based boosting in high dimensions. Bioinformatics. 2006;22(22):2828–9. https://doi.org/10.1093/bioinformatics/btl462.
21. Hothorn T, Bühlmann P, Dudoit S, Molinaro A, Van Der Laan MJ. Survival ensembles. Biostatistics. 2006;7(3):355–73. https://doi.org/10.1093/biostatistics/kxj011.
22. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics. 2012;28(24):3290–7. https://doi.org/10.1093/bioinformatics/bts595.
23. Lang M, Bischl B, Surmann D. batchtools: tools for r to work on batch systems. J Open Source Softw. 2017. https://doi.org/10.21105/joss.00135
24. Liu J, Huang J, Ma S. Integrative analysis of cancer diagnosis studies with composite penalization. Scand J Stat Theory Appl. 2014;41(1):87–103. https://doi.org/10.1111/j.1467-9469.2012.00816.x.
25. Liu J, Huang J, Zhang Y, Lan Q, Rothman N, Zheng T, Ma S. Integrative analysis of prognosis data on multiple cancer subtypes. Biometrics. 2014;70(3):480–8. https://doi.org/10.1111/biom.12177.

26. Madjar K. Survival models with selection of genomic covariates in heterogeneous cancer studies. Dissertation. Faculty of Statistics, TU Dortmund University (2018). https://doi.org/10.17877/DE290R-19140
27. Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghiren E, Ameh F, Achas M, Adebiyi E. Clustering algorithms: their application to gene expression data. Bioinform Biol Insights. 2016;10:38316. https://doi.org/10.4137/BBI.S38316.
28. Pagès H, Carlson M, Falcon S, Li N. AnnotationDbi: manipulation of SQLite-based annotations in bioconductor. 2019. R package version 1.46.0. https://bioconductor.org/packages/AnnotationDbi
29. Richter J, Madjar K, Rahnenführer J. Model-based optimization of subgroup weights for survival analysis. Bioinformatics. 2019;35(14):484–91. https://doi.org/10.1093/bioinformatics/btz361.
30. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–12. https://doi.org/10.1093/bioinformatics/btp543.
31. Simon R. Bayesian subset analysis: application to studying treatment-by-gender interactions. Stat Med. 2002;21(19):2909–16. https://doi.org/10.1002/sim.1295.
32. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B (Methodol). 1996;58(1):267–88.
33. Tibshirani R. The Lasso method for variable selection in the Cox model. Stat Med. 1997;16(4):385–95.
34. Tutz G, Binder H. Localized classification. Stat Comput. 2005;15(3):155–66. https://doi.org/10.1007/s11222-005-1305-x.
35. Tutz G, Binder H. Generalized additive modeling with implicit variable selection by likelihood-based boosting. Biometrics. 2006;62(4):961–71. https://doi.org/10.1111/j.1541-0420.2006.00578.x.
36. Uno H, Cai T, Tian L, Wei LJ. Evaluating prediction rules for t-year survivors with censored regression models. J Am Stat Assoc. 2007;102(478):527–37.
37. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30(10):1105–17. https://doi.org/10.1002/sim.4154.
38. Verweij PJ, Van Houwelingen HC. Penalized likelihood in Cox regression. Stat Med. 1994;13(23–24):2427–36.
39. Weyer V, Binder H. A weighting approach for judging the effect of patient strata on high-dimensional risk prediction signatures. BMC Bioinform. 2015;16:294. https://doi.org/10.1186/s12859-015-0716-8.
40. White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med. 2009;28(15):1982–98. https://doi.org/10.1002/sim.3618.
41. Witten DM, Tibshirani R. A framework for feature selection in clustering. J Am Stat Assoc. 2010;105(490):713–26. https://doi.org/10.1198/jasa.2010.tm09415.
42. Wiwie C, Baumbach J, Röttger R. Comparing the performance of biomedical clustering methods. Nat Methods. 2015;12(11):1033–8. https://doi.org/10.1038/nmeth.3583.
43. Wright M.N, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw. 2017;77(1):1–17. https://doi.org/10.18637/jss.v077.i01.
44. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J. 51 other authors: Ensembl 2018. Nucl Acids Res. 2018;46(D1):754–61. https://doi.org/10.1093/nar/gkx1098.
45. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Stat Methodol). 2005;67(2):301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Publisher's Note