Bioinformatics, 35, 2019, i484–i491
doi: 10.1093/bioinformatics/btz361
ISMB/ECCB 2019

Model-based optimization of subgroup weights for survival analysis

Jakob Richter*, Katrin Madjar and Jörg Rahnenführer

Department of Statistics, TU Dortmund University, Dortmund, Germany

*To whom correspondence should be addressed.

Abstract

Motivation: To obtain a reliable prediction model for a specific cancer subgroup or cohort is often difficult due to limited sample size and, in survival analysis, due to potentially high censoring rates. Sometimes similar data from other patient subgroups are available, e.g. from other clinical centers. Simple pooling of all subgroups can decrease the variance of the predicted parameters of the prediction models, but also increase the bias due to heterogeneity between the cohorts. A promising compromise is to identify those subgroups with a similar relationship between covariates and target variable and then include only these for model building.

Results: We propose a subgroup-based weighted likelihood approach for survival prediction with high-dimensional genetic covariates. When predicting survival for a specific subgroup, for every other subgroup an individual weight determines the strength with which its observations enter into model building. MBO (model-based optimization) can be used to quickly find a good prediction model in the presence of a large number of hyperparameters. We use MBO to identify the best model for survival prediction of a specific subgroup by optimizing the weights for additional subgroups for a Cox model. The approach is evaluated on a set of lung cancer cohorts with gene expression measurements. The resulting models have competitive prediction quality, and they reflect the similarity of the corresponding cancer subgroups, with both weights close to 0 and close to 1 and medium weights.

Availability and implementation: `mlrMBO` is implemented as an R-package and is freely available at http://github.com/mlr-org/mlrMBO.

Contact: jakob.richter@tu-dortmund.de

1 Introduction

Survival analysis is a central aspect in cancer research with the aim of predicting a patient's risk based on genomic and/or clinical covariates. In clinical practice, this is often challenging because patient cohorts are typically small and can be heterogeneous with regard to their relationship between covariates and survival outcome. One standard approach in multicenter studies is to simply pool different patient cohorts (here cohorts from different clinical centers) to increase sample size. However, this can lead to biased results especially when the cohorts are heterogeneous. In standard subgroup analysis, only the patients of the subgroup of interest s^* are included in the subgroup-specific model. This can lead to unstable results, especially for smaller subgroups.

We aim at improving the prediction performance for a specific subgroup by adaptively adding data from the other subgroups, in order to benefit from the larger sample size, but at the same time taking into account heterogeneity. Our proposed model potentially uses all subgroups but assigns them subgroup-dependent weights. If the inclusion of another subgroup increases the predictive performance, this subgroup enters with a higher weight into the model building process.

This idea extends the work of Weyer and Binder (2015) who aim at improving stability and prediction quality of a model for a specific subgroup by including one additional weighted subgroup. The authors study the effects of a set of different fixed weights for the additional subgroup in a stratified Cox model, with respect to both, model performance and parameter stability.

In our approach, we use multiple additional subgroups and efficiently optimize respective subgroup-specific weight parameters to improve the prediction quality of a Cox model. The optimal subgroup weights are determined by optimizing the cross-validated Concordance index (C-index) through Bayesian optimization (Jones *et al.*, 1998). In an adapted version of classical cross-validation, only the patients of the subgroup s^* of interest are included in the test set, while all patients from all subgroups can potentially be used for training. The idea is to assign large weights exactly to those subgroups that improve the prediction performance of the model for subgroup s^*. Those subgroups that deteriorate the predictive performance (mainly due to a different relationship between covariates and survival outcome) are assigned lower weights.

We show that with our subgroup weights optimization approach, the predictive quality can be improved compared to the two naïve approaches to either fully include or fully exclude all other subgroups. As an application example, we use 10 non-small-cell lung cancer (NSCLC) studies as subgroups and optimize the prediction quality for each subgroup, respectively, using all other subgroups with optimized weights.

1.1 Related work

The approach of Weyer and Binder (2015) uses the same fixed weight for all other subgroups. Alternatively, individual weights for each patient can be estimated from the training data as proposed by Bickel *et al.* (2008). The idea is that weights match the joint distribution of the complete data to the distribution in each subgroup, such that a patient who is likely to belong to the subgroup of interest receives a higher weight in the subgroup-specific model. Weights correspond to the conditional probability of belonging to the target subgroup s^* given the observed covariates and outcome divided by the prior probability for s^*. The former is estimated from the training data by multiclass classification and the latter by the relative frequency of s^*.

Bayesian approaches for estimating subgroup weights were proposed by Bogojeska and Lengauer (2012) and Simon (2002). Bogojeska and Lengauer (2012) use a hierarchical weighted logistic regression model with prior distributions for the subgroup weights to predict the binary treatment response of an HIV combination therapy. Simon (2002) considers a Cox model including a binary treatment effect, a binary subgroup indicator and the corresponding treatment-by-subgroup interaction with prior distribution for the regression coefficients. The author shows that the components of the posterior mean are linear combinations of the estimated treatment effects in different subgroups and extracting the respective scalars yields the subgroup-specific weights.

Integrative analysis combines data from different data sources such as multiple studies. In the context of high-dimensional genomic predictors, Liu *et al.* (2014a, b) suggest regularized regression with composite penalties for parameter estimation and gene selection. These penalties allow to select 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.* (2017) perform integrative analysis of multiple omics data types available for the same patient cohort. They use a lasso penalty with different penalty parameters for the different data types. Bergersen *et al.* (2011) 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.

Instead of sharing information between subgroups by integrating external information into variable selection, Huang *et al.* (2011) propose a weighted approach for combining positive predictive value (PPV) and negative predictive value (NPV) across populations when the assumption of common classification accuracy is justified. ROC curve estimation is used to evaluate the ability of a risk prediction marker in discriminating diseased from non-diseased. The estimates of PPV and NPV are based on a weighted average of the ROC curves from a target and an auxiliary population.

Local regression uses weighted regression models but without predefined groups. A separate model is fitted to each observation based on its neighboring observations. Weights in the likelihood of the local regression model represent the distance from the observation of interest and determine to which extent the neighboring observations influence the estimation. All single local regression models together form the local weighted regression based on all observations (Hastie *et al.*, 2009, chapters 2.8.2 and 6).

Instead of using distance in covariate space, our proposed weights are optimized with respect to prediction performance. 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. This makes estimation in high-dimensional settings even more complicated. To deal with this problem, Tutz and Binder (2005) developed a penalized localized classification approach and Binder *et al.* (2012) propose a cluster-localized logistic regression with weighted component-wise likelihood-based boosting for automatic variable selection and a special clustering for SNP data.

Above-mentioned approaches have in common that the goal is to achieve a high predictive accuracy on a specific subgroup by combining data from different subgroups. If multiple subgroups are available, another aim can be the identification of models that not only work well on the data they have been trained on but also on the other subgroups. Bernau *et al.* (2014) point out that classical cross-validation tends to be too optimistic, with respect to accuracy on data of unseen subgroups. They propose to include all available subgroups for the validation of a single model. Zhao *et al.* (2014) compare different regression methods for gene expression data and survival outcome with respect to accuracy. They use the largest subgroup for training and investigate which regression approach performs well on the remaining subgroups.

2 Gene expression data

Ten lung cancer cohorts with overall survival and censoring information, Affymetrix microarray gene expression data of the tumor material, and several clinicopathologic information, were downloaded from the Gene Expression Omnibus (GEO) data repository (Edgar *et al.*, 2002) and manually curated as follows. Raw gene expression data (CEL files), measured on the Affymetrix HG-U133 Plus 2.0 and HGU-133A array, were normalized using frozen robust multiarray analysis (fRMA) (McCall *et al.*, 2010), except for GSE3141 and GSE4573, where only MAS5-normalized data were available. All cohorts were checked for duplicates by looking at correlations of the expression value vectors. Duplicates, small cell cancer samples and normal (nontumorous) samples, as well as samples with missing survival endpoint were removed. More details on the data curation process can be found in Hellwig *et al.* (2016).

The resulting 10 NSCLC cohorts comprise $n = 1779$ patients with available overall survival endpoint and gene expression data as covariates which are used for analysis. The total number of measured genetic covariates (probe sets that represent genes) in each cohort is 22 283 or 54 675 depending on the Affymetrix array. We restricted the analysis to the 22 277 probe sets in the overlap of both Affymetrix arrays. The majority of these probe sets are noise and do not contain relevant information regarding survival outcome. This makes the identification of the prognostic genes more difficult and slows down computation time. Therefore, we use a reduced gene set for analysis that is based on the 1000 features (probe sets) with the highest variance in gene expression values across all 10 cohorts within the optimization dataset and a small number of literature-based prognostic genes. The selection of the top-1000-variance features is based on the assumption that important prognostic genes imply systematic changes in their expression values and thus, a larger variance in contrast to irrelevant noise genes. The following 30 prognostic features (probe sets) belong to 13 of the 14 prognostic lung cancer genes from Kratz *et al.* (2012) with matches on the Affymetrix HG-U133 Plus 2.0 and HGU-133A array. Gene symbols (provided in brackets) were translated into corresponding probe set IDs using the R/Bioconductor annotation package hgu133plus2.db (version 3.2.3):

`202387_at` (BAG1), `211475_s_at` (BAG1), `229720_at` (BAG1), `204531_s_at` (BRCA1), `211851_x_at` (BRCA1), `203967_at` (CDC6), `203968_s_at` (CDC6), `201938_at` (CDK2AP1), `1563252_at` (ERBB3), `1563253_s_at` (ERBB3), `202454_s_at` (ERBB3), `215638_at` (ERBB3), `226213_at` (ERBB3), `214088_s_at` (FUT3), `216010_x_at` (FUT3), `206924_at` (IL11), `206926_s_at` (IL11), `204890_s_at` (LCK), `204891_s_at` (LCK), `212724_at` (RND3), `204979_s_at` (SH3BGR), `209009_at` (ESD), `215095_at` (ESD), `215096_s_at` (ESD), `228162_at` (ESD), `240808_at` (ESD), `203135_at` (TBP), `213342_at` (YAP1), `224894_at` (YAP1), `224895_at` (YAP1).

For each cohort, the Kaplan–Meier estimator of the survival function is plotted in Figure 1. The Kaplan–Meier plot shows that patients in cohort `GSE31210` have the best prognosis with a 10-year overall survival probability of about 75%, while `GSE37745` exhibits a poor prognosis with a 10-year overall survival of 25%. `GSE29013` has the shortest maximum follow-up time with about 7 years, and `GSE30219` the longest maximum follow-up time with about 20 years. A summary of the clinicopathologic variables is provided in Table 1.

3 Weighted Cox model

Assume the observed data of patient i consists of the tuples (t_i, δ_i), the covariate vectors $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})' \in \mathbb{R}^p$, and the subgroup membership $s_i \in \{1, \ldots, S\}$ with S the number of subgroups in the complete dataset, 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. $\delta_i = \mathbb{1}(T_i \le C_i)$ 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 (Cox, 1972). 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 parameters are estimated by maximizing the partial log-likelihood (Klein and Moeschberger, 2003, chapter 8.3).In order to take subgroups into account, a weighted version of the partial log-likelihood as in Weyer and Binder (2015) is used:

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

In the subgroup-specific model for subgroup s^*, the individual weights are given by

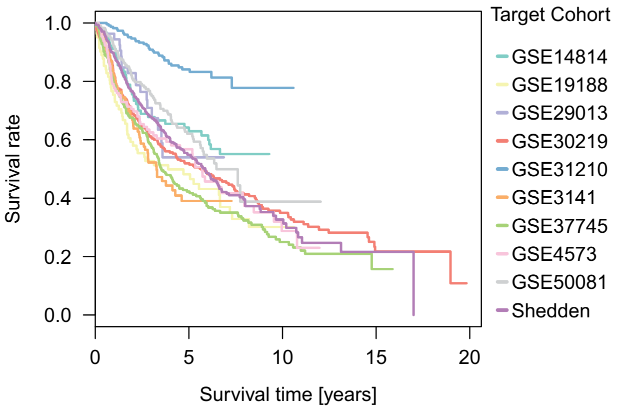

Fig. 1. Kaplan–Meier plots of the estimated survival functions for all 10 lung cancer cohorts

$$w_i = \begin{cases} 1, & \text{if } s_i = s^* \\ w^{(g)}, & \text{if } s_i = g,\ g \in \{1, \ldots, S\} \setminus s^* \end{cases} \qquad (2)$$

where $w^{(g)} \in [0, 1]$ is the specific weight for subgroup g. Standard subgroup analysis is based only on the patients in the subgroup of interest (target subgroup 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.

In high-dimensional settings where the number of covariates p is typically much larger than the sample size n, standard maximum likelihood cannot be used for parameter estimation. Therefore, we add a lasso penalty (Tibshirani, 1996, 1997) to the partial log-likelihood. Lasso regression 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 10-fold cross-validation.

4 Model-based optimization

Sequential model-based optimization (MBO) (Jones *et al.*, 1998) (also known as Bayesian optimization) is a state-of-the-art (Shahriari *et al.*, 2016) technique for expensive black-box optimization problems. In comparison to other black-box optimization methods, like genetic algorithms or simulated annealing, MBO is especially suitable when evaluating a configuration (e.g. fitting and evaluating a model with specific hyperparameters, here denoted by θ) is very time consuming, as it becomes infeasible to evaluate the black box for thousands of configurations. MBO solves the optimization problem within a bounded search space θ:

$$\theta^* := \operatorname{argmin}_{\theta \in \Theta} f(\theta),$$

where $f(\theta)$ denotes the evaluation of the black box with the input configuration θ. To reduce the number of evaluations on f the key idea of MBO is to only evaluate values of θ that are expected to lead to a small value of $f(\theta)$. The estimate $\hat{f}(\theta)$ is generated by a so-called *surrogate model*. Typically, this is a regression model that predicts the outcome of f based on previous evaluations of f. First, an initial design of already evaluated configurations is needed. Then,

Table 1. Overview of clinical variables for each lung cancer cohort in the complete NSCLC dataset

Variable	Values	GSE14814	GSE19188	GSE29013	GSE30219	GSE31210	GSE3141	GSE37745	GSE4573	GSE50081	Shedden
Sample size		90	82	55	269	226	110	194	130	181	442
	Min.	38		32	15	30		39	42	40	33
Age (years)	Mean	62		64	61	60		64	67	68	64
	Max.	81		76	84	76		84	91	87	87
	Male	67	59	38	228	105	0	105	82	98	223
Sex	Female	23	23	17	40	121	0	89	47	83	219
	NA	0	0	0	1	0	110	0	1	0	0
	I	45	0	24	183	168	0	128	73	127	0
	II	45	0	14	35	58	0	35	34	54	0
pTNM stage	III	0	0	17	42	0	0	27	23	0	0
	IV	0	0	0	4	0	0	4	0	0	0
	NA	0	82	0	5	0	110	0	0	0	442
	SQC	52	24	25	61	0	52	64	130	43	0
	ADC	28	40	30	85	226	58	106	0	127	442
Histology	LCC	10	18	0	55	0	0	24	0	7	0
	other NSCLC	0	0	0	68	0	0	0	0	4	0
	NA	0	0	0	0	0	0	0	0	0	0
	Never smoker	0	0	2	0	115	0	15	0	24	49
Smoking status	Current /ex-smoker	0	0	53	0	111	0	179	123	136	300
	NA	90	82	0	269	0	110	0	7	21	93
Survival status	Censoring	52	32	37	99	191	52	51	63	106	206
	Event	38	50	18	170	35	58	143	67	75	236

iteratively, the MBO algorithm fits the surrogate on the previous evaluations, proposes a new configuration θ and evaluates it on f.

A so-called infill criterion guides the proposal of new configurations θ based on $\hat{f}$. It balances between exploration of not yet evaluated regions in Θ and exploitation, i.e. the search in regions that promise best outcomes. As infill criterion, we use the augmented expected improvement (Huang *et al.*, 2006) that is well suited for noisy functions. The steps are repeated until a budget is exhausted. For non-noisy optimization, we would choose the configuration θ^* that has led to the best outcome of f to be returned as the tuning result. If the function outcome is noisy, the best observed outcome is likely to be distorted by noise and not located at the true posterior mean. Therefore, we employ the surrogate to estimate the posterior mean for each evaluated configuration to cancel out the noise. The configuration for which the surrogate estimates the best outcome is then returned as the optimization result θ^*.

We apply Kriging (also called Gaussian process regression) to fit the surrogate model that predicts the outcome of f for unknown values of θ. We use the implementation in the R-package *DiceKriging* (Roustant *et al.*, 2012) and configure it to apply the Mattern 3/2 kernel with an estimated *nugget effect* to account for the noisy response of f.

In our study, we apply MBO to optimize the subgroup-specific weights $w^{(g)}$ in the weighted Cox model. For each weight configuration $\theta = (w^{(1)}, \dots, w^{(S-1)})$ (assuming $s^* = S$), we evaluate the weighted Cox model with a 10-fold cross-validation. As a result, we obtain 10 noisy outcomes for each θ. These are fed to the MBO. However, due to numerical instabilities in some situations, the maximum likelihood estimation of the covariance matrix of the Kriging surrogate model can fail which results in a constant mean prediction. This leads to randomly proposed points for the next MBO step. To avoid this case, we implemented a fallback model: If the prediction of the surrogate model is constant, all noisy response values $f(\theta)$ belonging to the same θ are aggregated by their means. These simplified data usually lead to models without constant predictions.

5 Evaluation and results

We apply the methods described above to obtain a separate predictive model for each of the 10 NSCLC cohorts. We use a weighted Cox model to predict the survival function of each patient in the respective target subgroup s^*. The unknown parameter vector $\boldsymbol{\beta}$ is estimated by maximizing the penalized weighted partial log-likelihood in (1). Subgroup-specific weights (2) are optimized using MBO, with a budget of 300 evaluations. Parameters to be optimized are the Cox model parameters and the weight vector. The initial design for MBO consists of $2 \cdot (S-1)$ randomly sampled subgroup weights and the following additional specific extreme cases: exactly one other subgroup has weight 1 and all others have weight 0; all other subgroups have weight 0 or all other subgroups have weight 1. The target subgroup s^* always has weight 1. The objective is to maximize the predictive performance by adapting the weights for all other subgroups.

The predictive performance of the weighted Cox model is evaluated using the C-index. To assess the performance of a weight configuration, the C-index is evaluated on each fold of 10-fold cross-validation. We use a modified version of the cross-validation to take into account that we are only interested in the predictions on the target subgroup: The target subgroup is divided into 10 chunks, and to obtain the prediction for one chunk all remaining 9 chunks plus all observations from the additional subgroups are combined to the training dataset. The C-index is calculated only on the chunk of the target subgroup that was not used for model building.

To avoid overfitting and to judge the stability, we conduct the optimization in a nested cross-validation setting. We use a 5-fold cross-validation for the outer validation and we use the same modification as described above for the inner cross-validation. The target subgroup is divided into five chunks and four of those chunks plus all observations from the additional subgroups are combined to the data that is used as *optimization dataset*, i.e. for the inner cross-validation. Accordingly, the optimization is carried out five times on slightly different data with different random samples of the initial design. For each outer cross-validation fold, the optimization returns an optimal weight vector. The Cox model is then trained on the complete optimization dataset and the final C-index is calculated on the remaining chunk of the target subgroup. Therefore, we obtain five different weight vectors and five different C-index values. This enables us to judge the stability of the optimization results.

Our subgroups are derived from the gene expression data introduced in Section 2. We only use gene expression data as covariates, and each cohort acts as a subgroup. The number of genes included in the analysis is initially reduced to the 1000 features with highest variance across all 10 subgroups within the optimization dataset and in addition the 30 mandatory prognostic features given in Section 2.

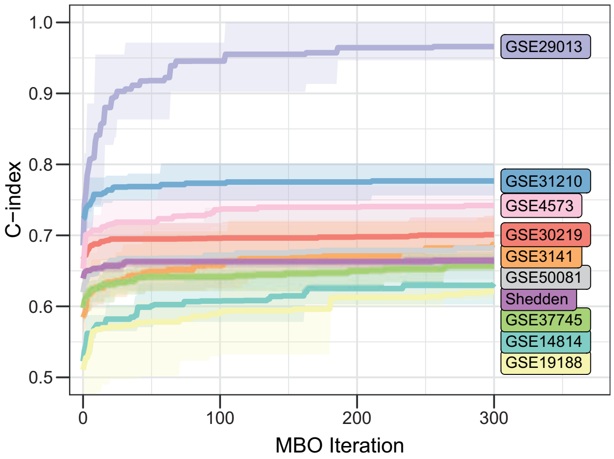

Fig. 2. Averaged progress of the MBO optimization runs for each target subgroup over all 300 optimization iterations

The underlying algorithms in this study are implemented in R, for the MBO the R-package mlrMBO (Bischl *et al.*, 2017) is used and survival analysis is performed using the R-package mlr (Bischl *et al.*, 2016).

We evaluate the effectiveness of the optimization by comparing the C-index resulting from three different strategies.

Subgroup uses only the observations of the target subgroup to train the Cox model (all weights 0, expect for target subgroup).

Pooled uses all subgroups to train the Cox model (all weights 1).

MBO uses the weight configuration that is proposed by the MBO for noisy black-box functions.

Figure 2 shows the averaged optimization curves from the five MBO optimization processes per target subgroup. It shows the averaged predictive performance of the so far best model at each optimization iteration for each target subgroup. For some cohorts, the predictive performance increases strongly over time, while for others no major improvement is observed. A strong increase can be seen especially for GSE29013, which is the smallest subgroup, with only 55 patients. Minor improvements can be seen for GSE4573, GSE3141, GSE14814 and GSE19188, although the last two suffer from a bad predictive performance even after optimization. For GSE30219, GSE31210 and Shedden, no major increase in performance can be observed. These are also the largest subgroups with 269, 226 and 442 patients. The performance reached at the end of the optimization is the *training error* and has the tendency to be too optimistic regarding the predictive performance on unseen data. Therefore, we use the C-index measured on the outer cross-validation in the following part.

Figure 3 compares the predictive performance of the weights obtained by *MBO* against the *Pooled* and *Subgroup* approach. Table 2 shows the median C-index values obtained by the outer 5-fold cross-validation and their ranks for each target subgroup. For five target subgroups *MBO* obtained the best median C-index, for four subgroups *Pooled* gave the best results and in one case *Subgroup* yielded the best results. Using only the target subgroup to train the Cox model yields the worst C-index for 6 out of 10 subgroups. Although the box plots in Figure 3 do not directly indicate a

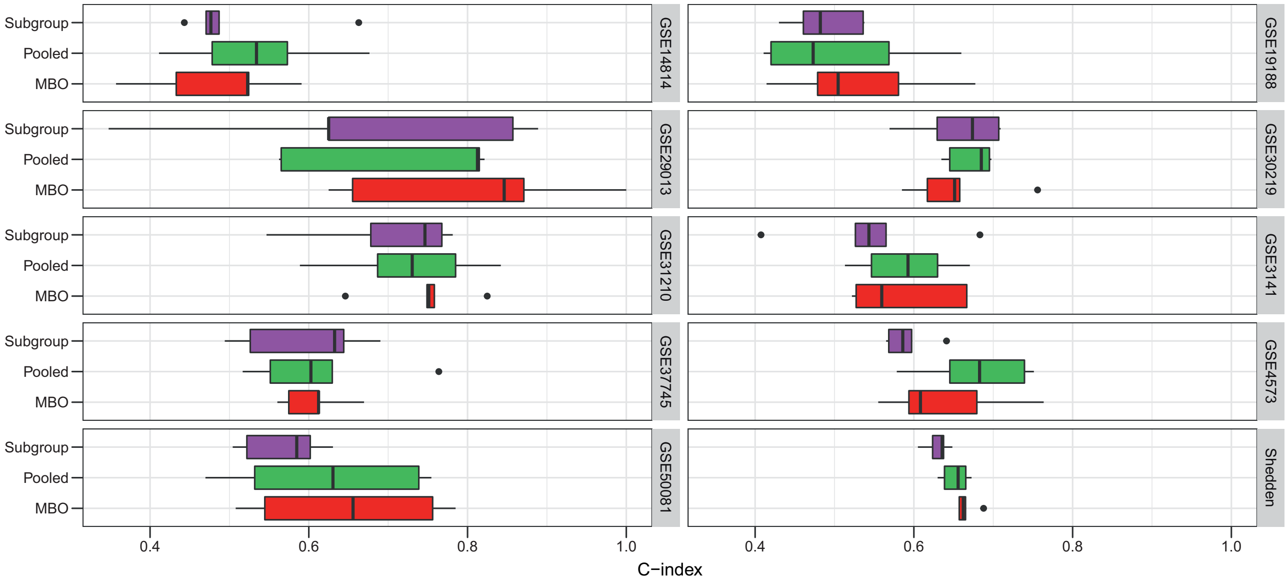

Fig. 3. The predictive performance of the best weight configuration according to different strategies. Each box plot includes the C-indices measured on the outer 5-fold cross-validation

superiority of *MBO* for each single target subgroup, the combined ranks in Table 2 show that on average using *MBO* is a promising strategy.

To verify whether there are statistically significant differences between the strategies, we perform a nonparametric test in accordance with Demšar (2006). The Friedman test is employed to test whether there are statistical differences between the given strategies at an α-level of 0.05. With a *P*-value of 0.12, the null hypothesis cannot be rejected.

Altogether, the results indicate that optimizing the subgroup weights is in most cases superior to the *Subgroup* strategy and competitive to the *Pooled* strategy. Moreover, we showed that *MBO* is capable of optimizing these weights.

5.1 Subgroup weights

Figure 4 shows the optimal weight vectors for each subgroup identified by *MBO*. Rows correspond to target subgroups and columns per plot indicate the subgroups to be used for model building. The line denotes the mean optimal weights averaged over the results of the five outer cross-validation folds. Overall, we see different patterns with weights close to 0 and close to 1, but sometimes also medium weights. For instance, for the target subgroup GSE31210, we see that there are consistent optimization results for the weights of the additional subgroups `GSE14814`, `GSE29013`, `GSE3141`, `GSE37745` and `GSE4573`. Interestingly, for the target subgroups `GSE19188`, `GSE29013` and `GSE4573`, no clear preferences for weights close to 0 or 1 are observable for any of the additional subgroups. However, this does not imply that *MBO* failed for those target subgroups. Looking at the results in Table 2, we see that *MBO* performs better or comparable then the *Pooled* approach for those cases.

Table 2 Median C-index for each target subgroup obtained on the outer 5-fold cross-validation by the different strategies

	Subgroup		Pooled		MBO	
Target subgroup	Med	Rank	Med	Rank	Med	Rank
`GSE14814`	0.48	(3)	**0.53**	(1)	0.52	(2)
`GSE19188`	0.48	(2)	0.47	(3)	**0.50**	(1)
`GSE29013`	0.62	(3)	0.81	(2)	**0.85**	(1)
`GSE30219`	0.67	(2)	**0.68**	(1)	0.65	(3)
`GSE31210`	0.75	(2)	0.73	(3)	**0.75**	(1)
`GSE3141`	0.54	(3)	**0.59**	(1)	0.56	(2)
`GSE37745`	**0.63**	(1)	0.60	(3)	0.61	(2)
`GSE4573`	0.59	(3)	**0.68**	(1)	0.61	(2)
`GSE50081`	0.58	(3)	0.63	(2)	**0.66**	(1)
`Shedden`	0.64	(3)	0.66	(2)	**0.66**	(1)
Average rank		2.50		1.90		1.60

Note: Ranks are given in brackets and averaged ranks across all target subgroups are given at the bottom.

An immediate question is whether weight values are bidirectional, meaning that an additional subgroup with a high weight for predicting the target subgroup also includes the latter with a high weight if it is the target subgroup itself. In Figure 4 we can especially notice that for some additional subgroups, a weight near 0 or 1 is clearly chosen by *MBO*. For example, `GSE14814` clearly benefits from `GSE30219` and vice versa. The same can be observed for the pairs `GSE37745`, `Shedden` as well as `GSE31210`, `GSE4573` and `GSE31210`, `GSE3141`. One can suspect that these datasets are similar in terms of which models achieve high performance values, and thus having a larger training dataset helps to increase the predictive performance.

A different scenario can be observed for the subgroup `GSE29013` (55 patients). As target subgroup, there is no clear preference for any additional subgroup. However, it is included with a high weight into the prediction of `Shedden` (442 patients). Also `GSE4573` (130 patients) does not show any clear preference toward additional subgroups but is included into the prediction of `GSE30219` (269 patients), `GSE31210` (226 patients) and `GSE37745` (194 patients) with a high weight. It appears that for smaller target subgroups, it can be advantageous to only include bigger additional subgroups with a slightly lower weight than 1 to obtain a high

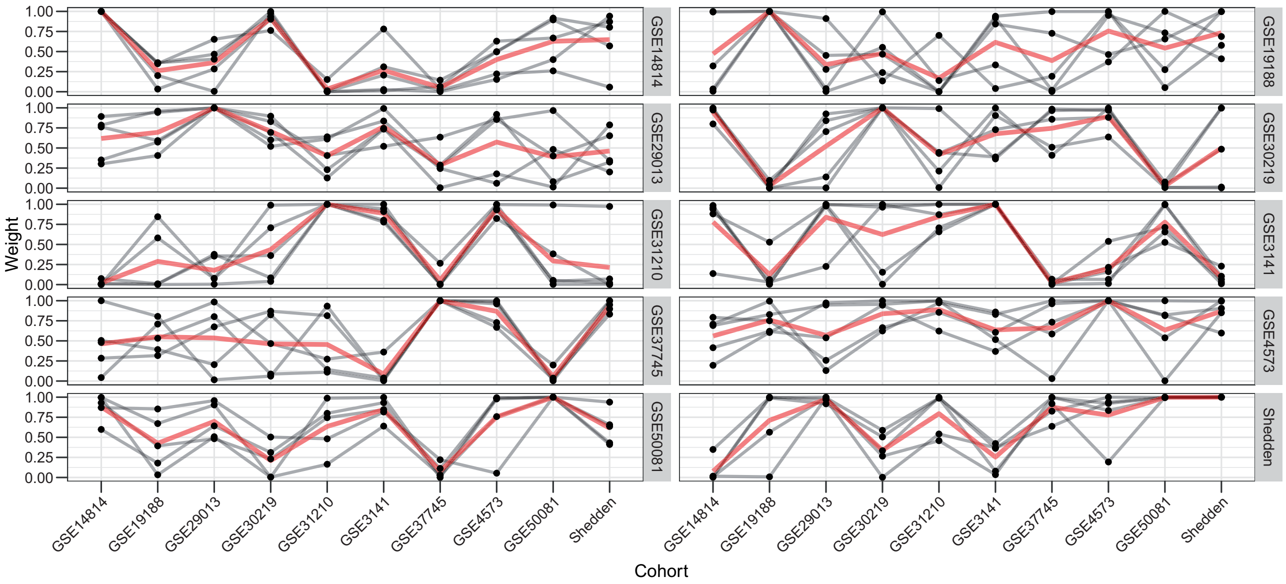

Fig. 4. Subgroup weights corresponding to the best predictive performance found by the model-based optimization. The row indicates the target cohort, the columns indicate the cohorts to be used for model building. Each dot represents the optimal weight for the respective subgroup obtained in one repetition of the optimization run. The red line denotes the mean over the five repetitions. If the dots per subgroup scatter heavily this indicates an unstable result

predictive accuracy. The other way around, for the bigger subgroups, the relatively few observations from the additional subgroups can be included with a high weight.

5.2 Deteriorated subgroups

To analyze how sensitive the different strategies react to subgroups that do not contain any information, we permuted the survival data of specific subgroups. For this analysis we used GSE3141 and GSE37745 as target subgroups. This has the effect that for these subgroups no useful information should be obtainable from the survival learner. Including them is expected to deteriorate the predictive performance for the target subgroup.

none uses the original data without permutations.

perm1 permutes the survival data of those subgroups to that MBO assigned weights below 0.5 in at least two runs. For the target subgroup GSE3141, we permute the survival data of GSE19188, GSE30219, GSE37745, GSE4573 and Shedden. For the target subgroup GSE37745, we permute the survival data of GSE14814, GSE19188, GSE29013, GSE30219, GSE31210, GSE3141 and GSE50081.

perm2 permutes the survival data of those subgroups to that MBO assigned the highest and the lowest weights on average. For the target subgroup GSE3141, we permute the survival data of the highly weighted GSE31210 and the down-weighted GSE37745 subgroup. For the target subgroup GSE37745, we permute the survival data of the highly weighted Shedden and the down-weighted GSE50081.

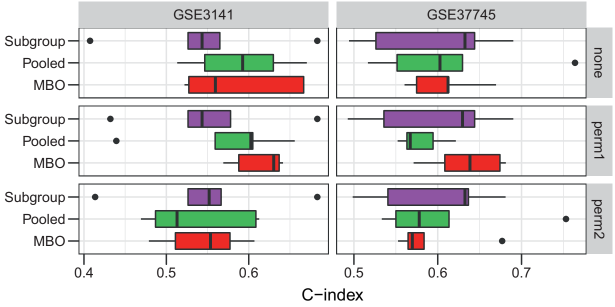

Fig. 5. Similar to Figure 3, the predictive performance is shown. For perm1 the subgroups with previously low predicted weights are additionally deteriorated and for perm2 the additional subgroup with the highest and the lowest previous weights are deteriorated

In Figure 5 we show how this permutation affects the predictive performance. As expected, the *Subgroup* strategy is not affected by distorting other subgroups. The results only vary slightly due to the randomness of the cross-validation. Interestingly, the *Pooled* strategy is not heavily affected by distorting many of the additional subgroups in perm1. The performance just slightly decreases. In contrast, the performance of *MBO* is capable of improving its performance. However, this cannot be easily explained by the corresponding weights. For perm2 the performance of the *Pooled* strategy and *MBO* drops noticeably. This indicates that the subgroups that were included with a high weight in the original setting were important to obtain the good prediction performance for both, *MBO* and the *Pooled* strategy.

Looking at the newly obtained weights in Figure 6, we can observe that for perm1 the assigned weights of the distorted subgroups are not closer to zero for GSE3141. However, for GSE37745 a slight tendency toward lower weights for deteriorated subgroups can be observed, expect for GSE3141 and GSE50081. The results are more conclusive for perm2: subgroups with a high weight in the original setting, now obtain low weights after permuting. Here *MBO* reliably detects the deterioration. We observe that deteriorating data from subgroups with small weights does not change their weights toward zero, different from what would be expected. It can be speculated that some of the additional subgroups have a small influence on the prediction on the target subgroup in the *Pooled* strategy and therefore, decreasing their weight does not strongly affect the predictive performance. If including a certain subgroup in the original data has a negative effect on the predictive performance, it has obtained a low weight by *MBO*. After permuting the survival data of this subgroup, this negative effect vanishes and thus, it is not as crucial anymore to assign low weights for the subgroup.

6 Summary

When multiple patient cohorts with a similar disease and treatment are available, it is tempting to pool the cohorts to one overall cohort to increase sample size and therefore, the stability of conclusions drawn from the data. However, heterogeneity between the cohorts can heavily distort these conclusions. We considered the situation in which one is interested in a good prediction model for one specific cohort out of a set of potentially similar cohorts. We analyzed a weighted likelihood strategy that is intended to only add those cohorts to the prediction model building process that represent a

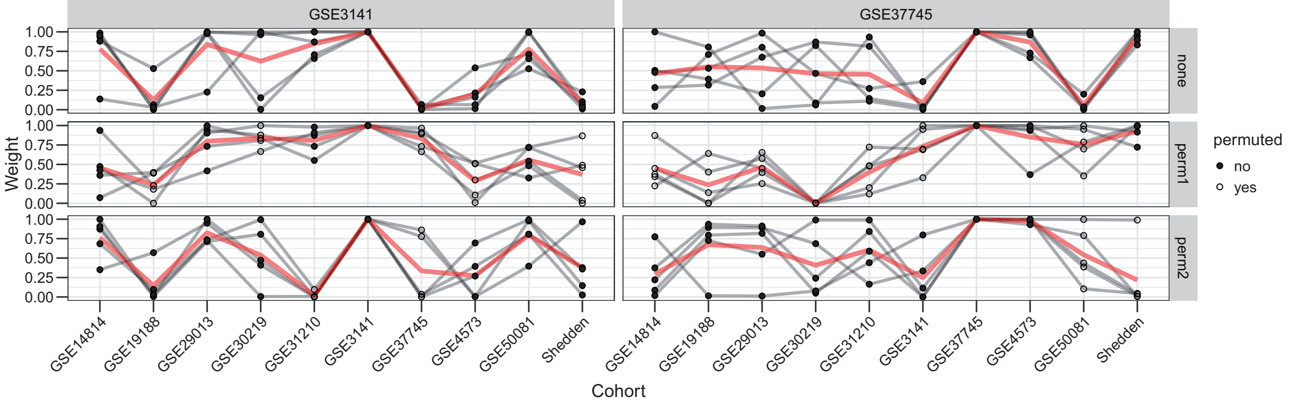

Fig. 6. Similar to Figure 4 the subgroup weights corresponding to the best predictive performance found by the model-based optimization are shown for two exemplary target subgroups. For perm1 and perm2 the survival data of the subgroups that are marked with an empty circle are permuted. The weights with no permutation in the upper panel are the same as in Figure 4 and drawn for comparison

similar feature–outcome relationship. For optimizing the weights of the other cohorts, we used MBO. In a lung cancer survival study, it turned out that this strategy often leads to an improved C-index as performance criterion, in a cross-validation setting.

Some important aspects for future research remain. It will be interesting to analyze in which way the size of the weight for a subgroup can be related to other properties of the corresponding patient subgroup, especially regarding sample size and the distributions of clinical covariates. Furthermore, our results indicate that the inconsistencies in the obtained weights are likely due to the small sample sizes. It remains to be examined to what extent more features, i.e. no filtering, and more observations lead to more stable and better predictions. For proposing the final weight MBO returns, the weight configuration that is estimated to perform best. This estimation can be nearly the same for various weight configurations. Therefore it might be interesting to introduce some slight regularization to give preference to specific values, e.g. $0, 0.5, 1$. Setting weights strictly to zero would lead to clearer decisions which subgroups to include for the model building process.

Funding

This work was partly supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876, A3.

Conflict of Interest: none declared.

References

Bergersen,L.C. *et al.* (2011) Weighted lasso with data integration. *Statist. Appl. Genet. Mol. Biol.*, **10**, 666.

Bernau,C. *et al.* (2014) Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*, **30**, i105–i112.

Bickel,S. *et al.* (2008) Multi-task learning for HIV therapy screening. In: *Proceedings of the 25th International Conference on Machine Learning, ICML '08, Helsinki, Finland, 2008*, pp. 56–63. ACM, New York.

Binder,H. *et al.* (2012) Cluster-localized sparse logistic regression for SNP data. *Statist. Appl. Genet. Mol. Biol.*, **11**, 1–28.

Bischl,B. *et al.* (2016) Mlr: machine learning in R. *J. Mach. Learn. Res.*, **17**, 1–5.

Bischl,B. *et al.* (2017) mlrMBO: a modular framework for model-based optimization of expensive black-box functions. *arXiv*: 1703.03373 [stat], pp. 1–26.

Bogojeska,J. and Lengauer,T. (2012) Hierarchical Bayes model for predicting effectiveness of HIV combination therapies. *Statist. Appl. Genet. Mol. Biol.*, **11**, 1–19.

Boulesteix,A.-L. *et al.* (2017) IPF-LASSO: integrative L 1-penalized regression with penalty factors for prediction based on multi-omics data. *Comput. Math. Methods Med.*, **2017**, 1.

Cox,D.R. (1972) Regression models and life-tables. *J. Royal Statist. Soc.*, **34**, 187–220.

Demšar,J. (2006) Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, **7**, 1–30.

Edgar,R. *et al.* (2002) Gene expression omnibus: nCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Hastie,T. *et al.* (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. Springer Science & Business Media, New York.

Hellwig,B. *et al.* (2016) Epsin family member 3 and ribosome-related genes are associated with late metastasis in estrogen receptor-positive breast cancer and long-term survival in non-small cell lung cancer using a genome-wide identification and validation strategy. *PLoS One*, **11**, e0167585.

Huang,D. *et al.* (2006) Global optimization of stochastic black-box systems via sequential kriging meta-models. *J. Global Optim.*, **34**, 441–466.

Huang,Y. *et al.* (2011) Borrowing information across populations in estimating positive and negative predictive values. *J. Royal Statist. Soc.*, **60**, 633–653.

Jones,D.R. *et al.* (1998) Efficient global optimization of expensive black-box functions. *J. Global Optim.*, **13**, 455–492.

Klein,J.P. and Moeschberger,M.L. (2003) *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health, 2nd edn. Springer, New York.

Kratz,J.R. *et al.* (2012) A practical molecular assay to predict survival in resected non-squamous, non-small-cell lung cancer: development and international validation studies. *Lancet*, **379**, 823–832.

Liu,J. *et al.* (2014a) Integrative analysis of cancer diagnosis studies with composite penalization. *Scand. J. Statist. Theory Appl.*, **41**, 87–103.

Liu,J. *et al.* (2014b) Integrative analysis of prognosis data on multiple cancer subtypes. *Biometrics*, **70**, 480–488.

McCall,M.N. *et al.* (2010) Frozen robust multiarray analysis (fRMA). *Biostatistics*, **11**, 242–253.

Roustant,O. *et al.* (2012) DiceKriging, DiceOptim: two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *J. Statist. Softw. Art.*, **51**, 1–55.

Shahriari,B. *et al.* (2016) Taking the Human out of the loop: a review of Bayesian optimization. *Proc. IEEE*, **104**, 148–175.

Simon,R. (2002) Bayesian subset analysis: application to studying treatment-by-gender interactions. *Statist. Med.*, **21**, 2909–2916.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc.*, **58**, 267–288.

Tibshirani,R. (1997) The lasso method for variable selection in the Cox model. *Statist. Med.*, **16**, 385–395.

Tutz,G. and Binder,H. (2005) Localized classification. *Statist. Comput.*, **15**, 155–166.

Weyer,V. and Binder,H. (2015) A weighting approach for judging the effect of patient strata on high-dimensional risk prediction signatures. *BMC Bioinformatics*, **16**, 294.

Zhao,S.D. *et al.* (2014) Más-o-menos: a simple sign averaging method for discrimination in genomic data analysis. *Bioinformatics*, **30**, 3062–3069.