Technical Report

Model-Based Optimization of Subgroup Weights for Survival Analysis

Jakob Richter, Katrin Madjar, Jörg Rahnenführer

05/2018

technische universität dortmund

SFB 876 Verfügbarkeit von Information durch Analyse unter Ressourcenbeschränkung

Part of the work on this technical report has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", project A3.

Speaker: Prof. Dr. Katharina Morik
Address: TU Dortmund University
Joseph-von-Fraunhofer-Str. 23
D-44227 Dortmund
Web: http://sfb876.tu-dortmund.de

Abstract

To obtain a reliable prediction model for a specific cancer subgroup or cohort is often difficult due to the limited number of samples and, in survival analysis, even more due to potentially high censoring rates. Sometimes similar datasets are available for other patient subgroups with the same or a similar disease and treatment, 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 potential high heterogeneity between the cohorts.

A promising compromise is to identify which subgroups are similar enough to the specific subgroup of interest and then include only these for model building. Similarity here refers to the relationship between input and output in the prediction model, and not necessarily to the distributions of the input and output variables themselves.

Here, we propose a subgroup-based weighted likelihood approach and evaluate it on a set of lung cancer cohorts. When interested in a prediction model for a specific subgroup, then for every other subgroup, an individual weight determines the strength with which its observations enter into the likelihood-based optimization of the model parameters. A weight close to 0 indicates that a subgroup should be discarded, and a weight close to 1 indicates that the subgroup fully enters into the model building process.

MBO (model based optimization) can be used to quickly find a good prediction model in the presence of a large number of hyperparameters to be tuned. Here, we use MBO to identify the best model for survival prediction in lung cancer subgroups, where besides the parameters of a Cox model additionally the individual values of the subgroup weights are optimized. Interestingly, often the resulting models with highest prediction quality are obtained for a mixed weight structure, i.e. both weights close to 0, weights close to 1, and medium weights are optimal, reflecting the similarity of the corresponding cancer subgroups.

1 Introduction

Survival analysis is a central aspect in cancer research with the aim of predicting the survival time of a patient on the basis of his covariates. Often it can be assumed that the relation between covariates and survival time is not the same across different subgroups of patients (e.g. cohorts from different clinical centers). Then, the aim is to improve the prediction of the survival function for a specific subgroup by appropriately adding data from the other subgroups, in order to benefit from the larger sample size.

In standard subgroup analysis only the patients of the subgroup of interest G are included in the model. This can lead to unstable results, especially for smaller subgroups. As an alternative we propose a model that potentially uses all subgroups but assigns them subgroup-dependent weights. When the relationship between covariates and survival time in a subgroup is more similar to the model for subgroup G, this subgroup enters with a higher weight into the model building process.

This idea extends the work of Weyer and Binder [14] who aim at improving stability and prediction quality of a model for a specific subgroup by including one additional weighted subgroup. They 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 through Bayesian optimization [7]. In an adapted version of classical cross-validation, only the patients of the subgroup G 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 G. Those subgroups that deteriorate the predictive performance (mainly due to a different relationship between covariates and survival time) are assigned lower weights.

In this report 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 ten non-small-cell lung cancer studies as subgroups and optimize the prediction quality for each subgroup, respectively, using all other subgroups with optimized weights.

2 Model-based Optimization

Sequential **model-based optimization (MBO)** [7] (also known as Bayesian Optimization) is a state-of-the-art [11] technique for expensive black-box optimization problems. In comparison to other black-box optimization methods, like Genetic Algorithms or Simulated Annealing, MBO is especially applicable when evaluating a configuration (model with its parameters and hyperparameters, her denoted by θ) takes a lot of time. MBO solves the optimization problem within a bounded search space Θ:

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

where $f(\theta)$ denotes the evaluation of the black-box with the input 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 θ of already evaluated configurations is needed. Then, 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 on regions that promise best outcomes. As infill criterion we use the augmented expected improvement [6] that is well suited for noisy functions. The steps are repeated until a budget is exhausted. The setting θ^* that leads to the best outcome is returned as the result.

2.1 Surrogate Model

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 DiceKriging package and configure it to apply the Mattern$\frac{3}{2}$ kernel with an estimated *nugget effect* [10] to account for the noisy response of f. Due to numerical instabilities in some situations the maximum likelihood estimation of the covariance matrix 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 is constant, then the noisy response values $f(\theta)$ for a specific θ are aggregated by their means. These simplified data usually lead to models without constant predictions.

3 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 [4] 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) [9], 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 (non-tumorous) samples, as well as samples with missing survival endpoint were removed. More details on the data curation process can be found in [5].

The resulting ten non-small cell lung cancer (`NSCLC`) cohorts comprise $n = 1779$ patients with available overall survival endpoint and gene expression data. These data, as well as the `ADENOS` subset containing only adenocarcinoma samples ($n = 1142$), are used for analysis. Estimated survival functions of each cohort are plotted in Figure 1. A summary of the clinicopathologic variables is provided in Table 1.

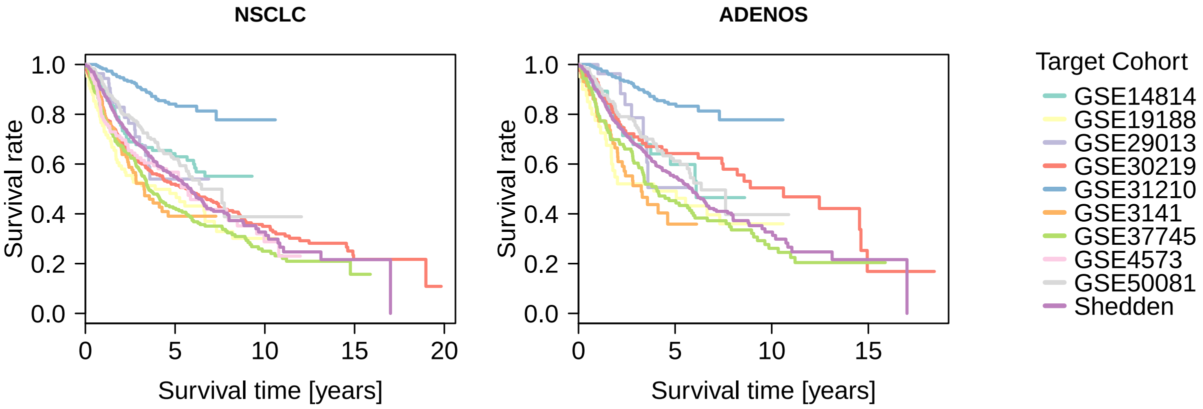

Figure 1: Kaplan-Meier plots of the estimated survival functions for all ten lung cancer cohorts.

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
Age (years)	min.	38		32	15	30		39	42	40	33
	mean	62		64	61	60		64	67	68	64
	max.	81		76	84	76		84	91	87	87
Sex	male	67	59	38	228	105	0	105	82	98	223
	female	23	23	17	40	121	0	89	47	83	219
	NA	0	0	0	1	0	110	0	1	0	0
pTNM stage	I	45	0	24	183	168	0	128	73	127	0
	II	45	0	14	35	58	0	35	34	54	0
	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
Histology	SQC	52	24	25	61	0	52	64	130	43	0
	ADC	28	40	30	85	226	58	106	0	127	442
	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
Smoking status	never-smoker	0	0	2	0	115	0	15	0	24	49
	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

4 Weighted Cox model

Let S be the number of subgroups in the dataset. Assume that the observed data 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 $s_i \in \{1, \ldots, S\}$, $i = 1, \ldots, n$. $t_i = \min(T_i, C_i)$ 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 [3]. 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 ([8], chapter 8.3).

In order to take subgroups into account, a weighted version of the partial log-likelihood as in [14] is used:

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

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

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

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 [12, 13] 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.

5 Evaluation

We apply the methods described above to estimate a model separately for each of the ten cancer cohorts. We use the 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 partial log-likelihood in (1). Subgroup specific weights (2) are optimized using MBO, with budget 300 evaluations (Cox model parameters plus weight vector). The initial design for MBO consists of $2 \cdot (S - 1)$ randomly sampled subgroup weights and additional specific extreme cases: Exactly one other subgroup has weight 1 and all others weight 0, all other subgroups have weight 0, all other subgroups have weight 1 (full model). The target subgroup 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 one weight configuration, the C-index is averaged using a modified 10-fold cross-validation: 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 data set. The C-index is only calculated on the one chunk of the target subgroup that was not used for model building. To judge the stability of the optimization results the whole optimization process is repeated 5 times, with different random samples for the initial design and different cross-validation splits of the target subgroup.

Optimization of subgroup weights is carried out twice: Once all patients are included (dataset `NSCLC`) and once only patients with tumor type adenocarcinoma (dataset `ADENOS`, with patients from only eight out of the ten cohorts). Each cohort is treated once as target subgroup, respectively. Only gene expression data are used as covariates and the number of genes included in analyses is initially reduced to the 1000 features with highest variance across all ten subgroups.

The algorithms for this work are implemented in R, for the model-based optimization the R-package mlrMBO [1] is used, and survival analysis is performed using the R-package mlr [2].

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

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

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

Init uses the subgroup weights that led to the best performance in the initial design (includes the above cases).

MBO uses the weights that led to the best performance during the model-based optimization process (does not include weights tried in the initial design).

Figure 2 shows for `ADENOS` (left) and `NSCLC` (right) the predictive performance of the so far best model during the MBO optimization process. For some cohorts the predictive performance increases strongly over time, while for others no improvement is observed. A strong increase can be seen especially for GSE14814 (`ADENOS`) and GSE29013 (`NSCLC`). The latter is also the smallest subgroup, with only 55 patients.

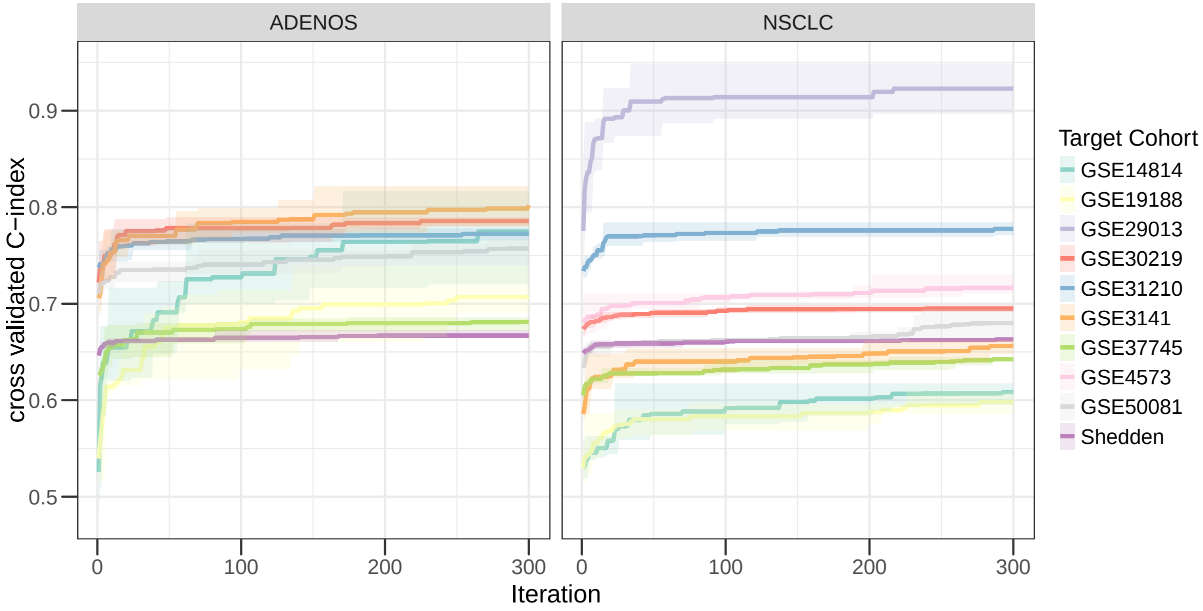

Figure 2: Progress of the MBO optimization over time, averaged over 5 replications. The shaded area indicates the range of all replications. Iteration 0 corresponds to the best performance from the initial design.

Figure 3 shows the optimal weight vectors for Init and MBO, for `ADENOS` (left) and `NSCLC` (right). 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 over 5 repetitions. Overall, we see different patterns with weights close to 0 and close to 1, but sometimes also medium weights.

Consider the two examples highlighted above with substantially improved prediction performance due to MBO. For GSE29013 (`NSCLC`) as target subgroup, most other subgroups obtain weights close to 0.5. Interestingly, the optimal weights from the initial design look similar. For GSE14814 (`ADENOS`), with only 90 patients, the other subgroups obtain larger weights after running the MBO optimization.

An immediate question is if weighting values are bidirectional, meaning that an additional subgroup that is weighted highly for predicting the target subgroup also uses the latter with a high weight if it is the target subgroup itself. In Figure 3 we can especially notice that for some additional subgroups a weight near 0 or 1 is clearly chosen by MBO. For example GSE14814 (`NSCLC`) clearly benefits most from GSE30219 and GSE50081 as highly weighted additional subgroups. The other way round, for GSE30219 and GSE50081 as target subgroup, also GSE14814 is a highly weighted additional subgroup. One can suspect that these datasets are similar w.r.t. to good models and thus having a larger training dataset helps to increase the predictive performance.

A different scenario can be observed for the target subgroup GSE31210 (`NSCLC`). There is no clear preference for additional subgroups, but GSE3141, GSE4573 and GSE50081 benefit from including GSE31210 with a high weight as additional subgroup. One reason could be that GSE31210 has a fairly large sample size with 226 observations, while the other cohorts are a bit smaller.

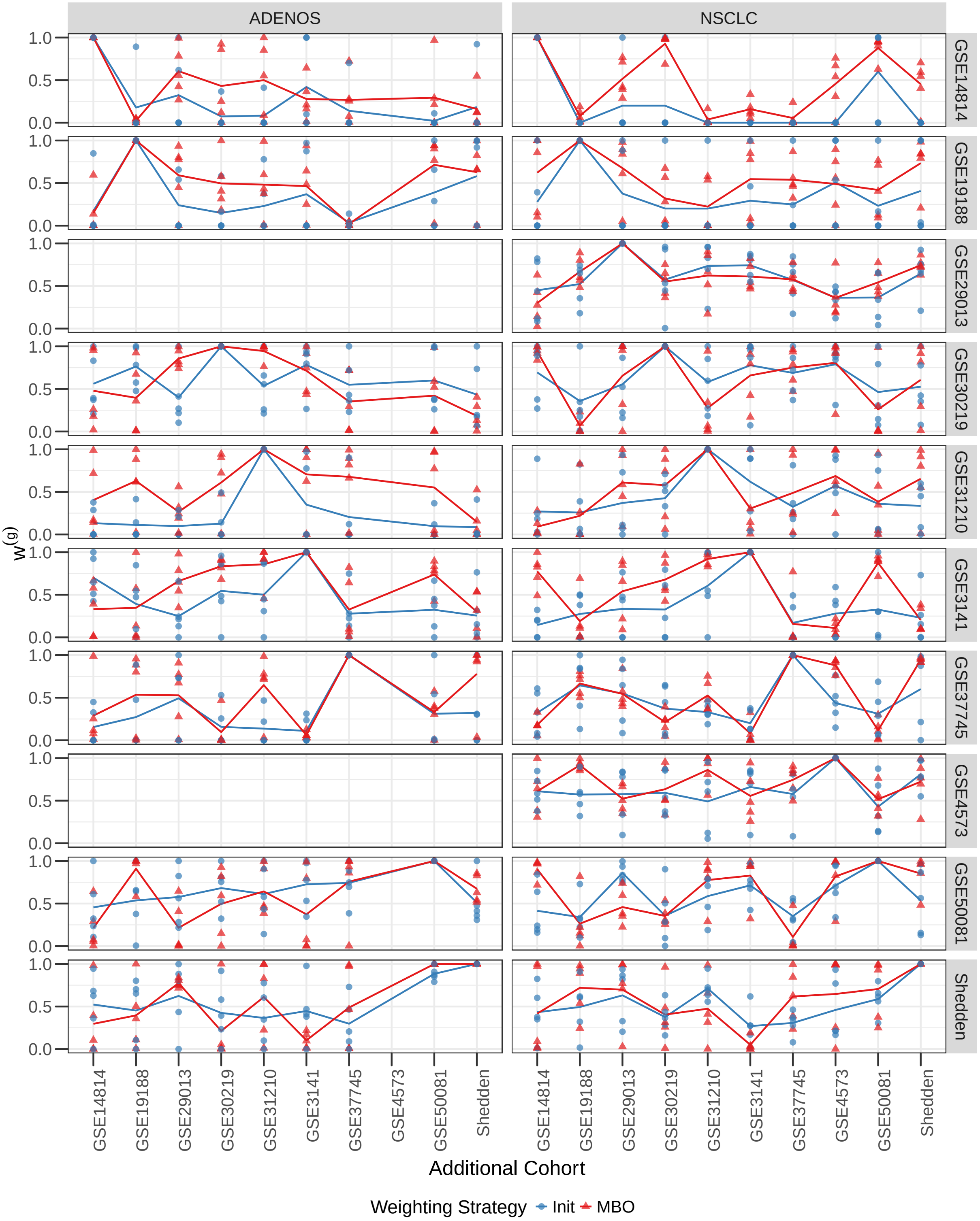

Figure 3: Subgroup weights corresponding to the best predictive performance within the initial design and after optimization with MBO. The row indicates the target subgroup, the columns per figure indicate the subgroups to be used for model building. Each dot represents the optimal weight for the respective subgroup obtained in one repetition of the optimization run. The line denotes the mean over the 5 repetitions. If the dots per subgroup scatter heavily this indicates an unstable result.

The C-index for predictions of the GSE29013 target cohort could not be validated for the `ADENOS` subset because sample size is too low (30 observations with 8 events). Cohort GSE4583 has no observations in the `ADENOS` subset. Therefore it is also not included in the weights.

Figure 4 shows the predictive performance of the best weight configuration identified with the four different weighting strategies. Overall, we observe that for almost all cases the C-index obtained with MBO is in average the highest, compared to the other weighting strategies. The only exceptions are the results for the target subgroups GSE3141 and GSE31210 (`NSCLC`). Using only the target subgroup to train the Cox model yields the worst C-index for the majority of cases, except for GSE31210, GSE19188 in the `ADENOS` dataset and GSE19188 and GSE14814 in the `NSCLC` dataset.

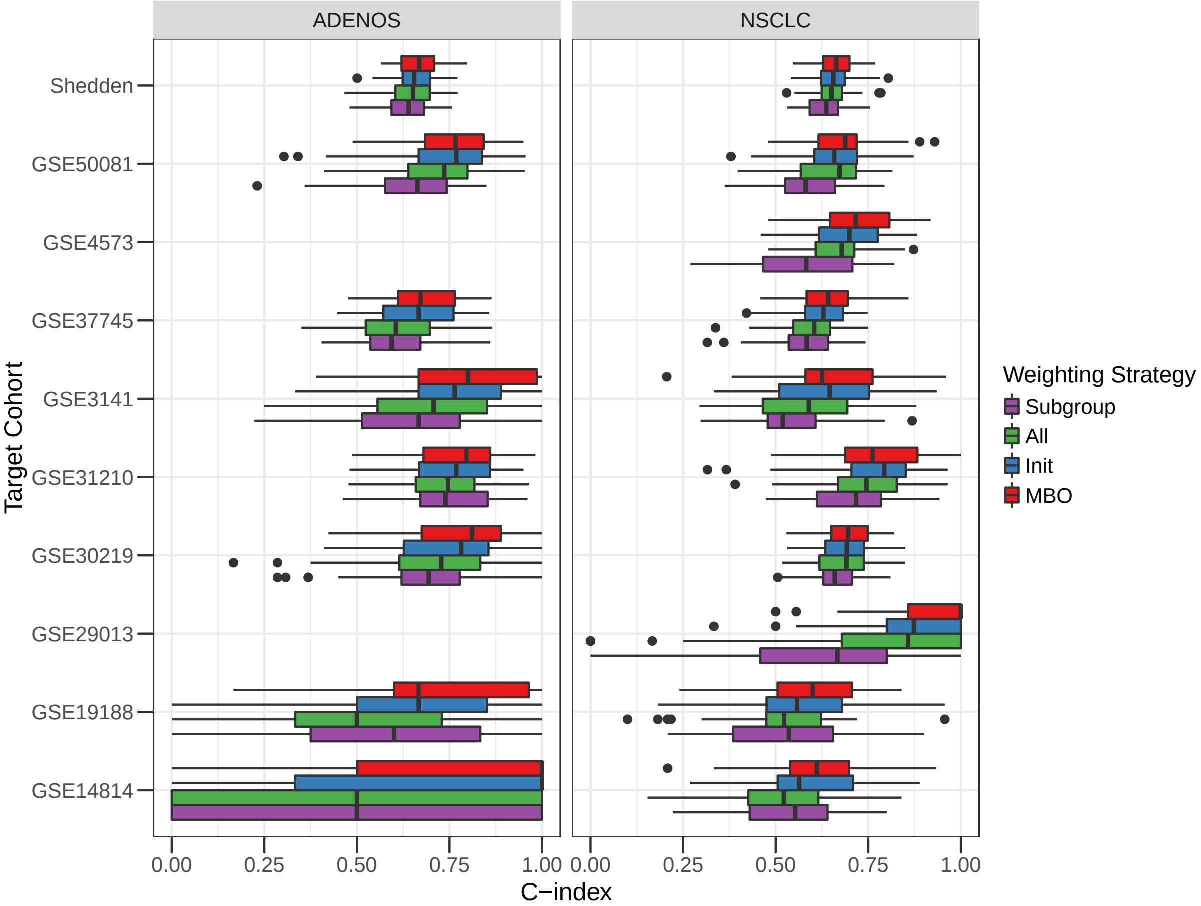

Figure 4: The predictive performance of the best weight configuration found using the different strategies. Each box plot includes the C-indices measured for the best weight setting of the 10-fold cross-validation and the 5 repetitions.

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

Two important aspects for future research remain. The implementation of a nested cross-validation setting will avoid overfitting of the optimization process. Further 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.

References

[1] B. Bischl, J. Richter, J. Bossek, D. Horn, J. Thomas, and M. Lang. mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions. URL `http://arxiv.org/abs/1703.03373`.

[2] B. Bischl, M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, and Z. M. Jones. mlr: Machine learning in r. *Journal of Machine Learning Research*, 17(170):1–5, 2016. URL `http://jmlr.org/papers/v17/15-066.html`.

[3] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 0035-9246. URL `http://www.jstor.org/stable/2985181`.

[4] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1): 207–210, Jan. 2002. ISSN 0305-1048. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC99122/`.

[5] B. Hellwig, K. Madjar, K. Edlund, R. Marchan, C. Cadenas, A.-S. Heimes, K. Almstedt, A. Lebrecht, I. Sicking, M. J. Battista, P. Micke, M. Schmidt, J. G. Hengstler, and J. Rahnenführer. 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(12), Dec. 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0167585. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5142791/`.

[6] D. Huang, T. T. Allen, W. I. Notz, and N. Zeng. Global Optimization of Stochastic Black-Box Systems via Sequential Kriging Meta-Models. 34(3):441–466. ISSN 0925-5001, 1573-2916. doi: 10.1007/s10898-005-2454-3. URL `http://link.springer.com/article/10.1007/s10898-005-2454-3`.

[7] D. R. Jones. A Taxonomy of Global Optimization Methods Based on Response Surfaces. 21(4):345–383. doi: 10.1023/A:1012771025575.

[8] J. P. Klein and M. L. Moeschberger. *Survival analysis.* Statistics for biology and health. Springer, New York [u.a.], 2. ed. edition, 2003. ISBN 978-0-387-95399-1.

[9] M. N. McCall, B. M. Bolstad, and R. A. Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics (Oxford, England)*, 11(2):242–253, Apr. 2010. ISSN 1465-4644. doi: 10.1093/biostatistics/kxp059. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2830579/.

[10] O. Roustant, D. Ginsbourger, and Y. Deville. DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization. 51(1):1–55. ISSN 1548-7660. doi: 10.18637/jss.v051.i01. URL https://www.jstatsoft.org/v051/i01.

[11] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. Taking the Human Out of the Loop: A Review of Bayesian Optimization. 104(1):148–175. ISSN 0018-9219. doi: 10.1109/JPROC.2015.2494218.

[12] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 0035-9246. URL http://www.jstor.org/stable/2346178.

[13] R. Tibshirani. The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine*, 16(4):385–395, Feb. 1997. ISSN 1097-0258. doi: 10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3. URL http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3/abstract.

[14] V. Weyer and H. Binder. A weighting approach for judging the effect of patient strata on high-dimensional risk prediction signatures. 16:294. doi: 10.1186/s12859-015-0716-8.