Contemporary Clinical Trials Communications 5 (2017) 107–115

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

Innovative methods for the identification of predictive biomarker signatures in oncology: Application to bevacizumab

Paul Delmar [a], [*], Cornelia Irl [b], Lu Tian [c]

[a] *Department of Biostatistics, F. Hoffmann-La Roche Ltd., Basel, Switzerland*
[b] *Department of Biostatistics, Genentech Inc., South San Francisco, CA, USA*
[c] *Department of Biomedical Data Science, Stanford University School of Medicine, Palo Alto, CA, USA*

ARTICLE INFO

Article history:
Received 18 December 2015
Received in revised form
6 December 2016
Accepted 17 January 2017
Available online 19 January 2017

Keywords:
Bevacizumab
Breast cancer
Personalized medicine
Subgroup analysis
Multivariate statistical classification
Machine learning

ABSTRACT

Current methods for subgroup analyses of data collected from randomized clinical trials (RCTs) may lead to false-positives from multiple testing, lack power to detect moderate but clinically meaningful differences, or be too simplistic in characterizing patients who may benefit from treatment. Herein, we present a general procedure based on a set of newly developed statistical methods for the identification and evaluation of complex multivariate predictors of treatment effect. Furthermore, we implemented this procedure to identify a subgroup of patients who may receive the largest benefit from bevacizumab treatment using a panel of 10 biomarkers measured at baseline in patients enrolled on two RCTs investigating bevacizumab in metastatic breast cancer. Data were collected from patients with human epidermal growth factor receptor 2 (HER2)-negative (AVADO) and HER2-positive (AVEREL) metastatic breast cancer. We first developed a classification rule based on an estimated individual scoring system, using data from the AVADO study only. The classification rule takes into consideration a panel of biomarkers, including vascular endothelial growth factor (VEGF)-A. We then classified the patients in the independent AVEREL study into patient groups according to "promising" or "not-promising" treatment benefit based on this rule and conducted a statistical analysis within these subgroups to compute point estimates, confidence intervals, and p-values for treatment effect and its interaction. In the group with promising treatment benefit in the AVEREL study, the estimated hazard ratio of bevacizumab versus placebo for progression-free survival was 0.687 (95% confidence interval [CI]: 0.462–1.024, p = 0.065), while in the not-promising group the hazard ratio (HR) was 1.152 (95% CI: 0.526–2.524, p = 0.723). Using the median level of VEGF-A from the AVEREL study to divide the study population, then the HR becomes 0.711 (95% CI: 0.435–1.163, p = 0.174) in the promising group and 0.828 (95% CI: 0.496–1.380, p = 0.468) in the not-promising group. Similar results were obtained with the median VEGF-A levels from the AVADO study ("promising" group: HR = 0.709, 95%CI: 0.444–1.133, p = 0.151; "not-promising" group: HR = 0.851, 95% CI: 0.497–1.458, p = 0.556). Our analysis shows it is feasible to employ statistical methods for empirically constructing and validating a scoring system based on a panel of biomarkers. This scoring system can be used to estimate the treatment effect for individual patients and identify a subgroup of patients who may benefit from treatment. The proposed procedure can provide a general framework to organize many statistical methods (existing or to be developed) into a coherent set of analyses for the development of personalized medicines and has the potential of broad applications.

1. Introduction

Randomized clinical trials are designed to assess the efficacy of a new treatment compared with placebo or standard of care. Oftentimes, in addition to the main comparison of the overall population enrolled in the study, subgroup analyses are performed to examine whether the benefit of the new treatment is consistent across patient populations [1]. Specifically, subgroup analyses aim

Abbreviations: FGF, fibroblast growth factor; FLT, fms-like tyrosine kinase; HER2, human epidermal growth factor receptor 2; MRST, in mean restricted survival time; RCT, randomized clinical trial; VEGF, vascular endothelial growth factor.

[*] Corresponding author. Department of Biostatistics, F.Hoffmann-La Roche Ltd., 4070 Basel, Switzerland.

E-mail addresses: paul.delmar@roche.com (P. Delmar), cornelia.irl@roche.com (C. Irl), lutian@stanford.edu (L. Tian).

http://dx.doi.org/10.1016/j.conctc.2017.01.007

to estimate and test the treatment effect on pre-determined subgroups. The subgroups are usually characterized by simple criterions measured at baseline, such as sex, race, comorbidities, and pre-existing treatment status. The final results are often presented graphically in a forest plot (e.g., Fig. 1), where each tree represents the point, as well as the interval estimates of the treatment effect within a subgroup. If one or several trees stand(s) out of the forest, this may indicate non-homogeneity of the treatment effect.

This simplicity, however, may be misleading [2,3]. The first difficulty associated with subgroup analyses is multiple testing [4]. If one tries to estimate the treatment effect in a sufficiently large number of subgroups, there will always be significant findings. This opens the door for subjective interpretation of the subgroups identified based on the significance level or the point estimator itself: it could be either a simple false-positive result due to multiple testing or a promising subgroup worthy of further investigation. Various statistical adjustments have been proposed but are rarely used in practice for good reasons [5]. For example, the Bonferroni correction is one of the most robust approaches to ensure that the treatment effect in at least one of the identified subgroups truly exists with the claimed significance level [6]. However, the adjustment is highly conservative and may fail to detect a moderate subgroup-specific treatment effect. This raises the second difficulty in subgroup analyses, i.e., lack of power to detect moderate yet clinically meaningful treatment effects [3]. Finally, the definition of the pre-defined subgroup may be too simplistic to characterize patients who may (or may not) benefit from the treatment. If we are willing to consider subgroups defined by a combination of characteristics, the number of candidate subgroups increases very rapidly, exacerbating the difficulties associated with multiple testing and lack of statistical power. For example, 10 binary characteristics can define up to 2048 different subgroups of patients. Even after acknowledging that some subgroups may be too small to be of interest, it is likely that we still need to deal with hundreds of subgroups. When some of the characteristics are continuous, such as systolic blood pressure or gene expression level, there are an infinite number of subgroups and it becomes infeasible to conduct subgroup analyses. More sophisticated methods that allow automatic identification of the subgroups of interest are needed [7–9].

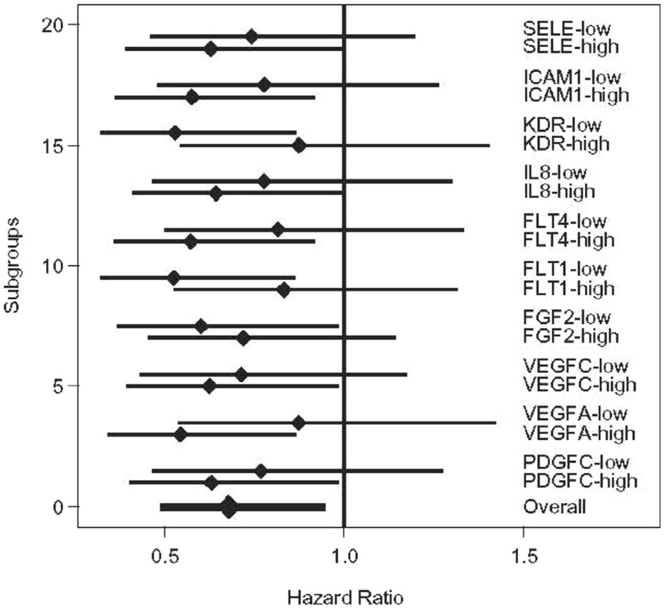

Fig. 1. Forest plot for subgroup analysis in AVADO study. The high and low groups are defined using the median of the corresponding biomarkers.

In light of these drawbacks of the simple subgroup analyses, there are many recent developments in statistical methodology for personalized medicine [7–19]. Among them, many adopt various modern machine learning techniques to relax conventional statistical model assumptions [11,14–19].

However, most of these recent developments are fragmentary and there is no practical guideline for conducting the complete statistical analysis for personalized medicine. For example, in the presence of multiple approaches for estimating personalized treatment effect and even different metrics for quantifying the personalized treatment effect, there is a lack of methods for selecting the optimal approach.

We have identified three goals for statistical methods in personalized medicine: (1) estimating the treatment effect for the individual patient, i.e., the individualized treatment effect [9,10,14–19], (2) building a classification rule for identifying patients who may (or may not) benefit from the treatment, or (3) making valid statistical inferences about treatment effect in the identified subgroup.

In this paper, we propose a coherent stage-wise procedure for addressing all three objectives. It has a clearly defined target at each step. The procedure is also flexible and can easily be extended to leverage new or future developments in the field. This procedure will be illustrated by analyzing the data from two randomized clinical oncology trials conducted by Hoffmann-La Roche Inc. In both trials, the overall comparisons showed moderate treatment effect in the entire study population and it is desirable to identify a subgroup of patients having more substantial treatment benefit [26,27]. However, the simple subgroup analysis failed to detect and confirm the existence of the heterogeneous treatment effect [28].

2. Methods

2.1. Procedure for subgroup selection

The procedure consists of two major steps: training and testing. The outcome of the training step is a classification rule for selecting a subgroup of patients based on baseline features including biomarker levels, demographic information, comorbidities, etc. The classification rule can be complex and depends on multiple features. The outcome of the testing step is the verification and evaluation of the treatment effect in the subgroup identified by the classification rule, as well as in the complementary subgroup. In the ideal case, there are data from two randomized clinical trials and we use the first for training (Part I) and the second for testing (Part II). If all the data are from a single trial, we need to split the data into two non-overlapping parts (Parts I and II).

2.2. Training step

In this stage, we estimate the treatment effect for individual patients and construct a classification rule for selecting patients with promising treatment effect. However, several estimation methods can be used and we need to select the optimal one based on the data. To this end, the estimation and validation steps need to be built within the training step. Specifically, the training data will be randomly split into two parts: the first part (Part I-E) will be used to estimate the treatment effect for individual patients with different methods; the second part (Part I-V) will be used to evaluate the performance of each of the estimated treatment effects in stratifying patient population into strata of different treatment effects.

2.3. Estimation

In general, there are two steps in estimating the personalized treatment effect using the Part I-E data.

(1) Select a measure for the treatment effect. Ideally the measure is estimable and has meaningful clinical interpretation. For time-to-event data, such as overall survival or progression-free survival, a popular choice is the hazard ratio (HR). However, the interpretation of the HR relies on the proportional hazards (PH) assumption, which is not always met in practice [20]. Alternatively, one may choose the difference in mean restricted survival time (MRST), which has the appealing intuitive interpretation of the area under the survival curve (AUC) over a given interval [21,22]. Once the measure is determined, one can define the personalized treatment effect accordingly. For example, if we decide that the difference in MRST is of interest, then the individualized treatment effect can be defined as

$$D(z) = E(\min(T,\ \tau)|R = 1,\ Z = z) - E(\min(T,\ \tau)|R = 0,\ Z = z),$$

where the binary indicator Rdenotes the treatment assignment, Z is a set of baseline covariates characterizing the patients, T is the survival time of interest, and τ is a given constant such that $D(z)$is identifiable from the observed data subject to right censoring. Note that a limitation of MRST is its dependence on cut-off time point τ, whose choice can be subjective in practice.

(2) Choose a regression model for estimating the individualized treatment effect specified in step 1. The output of the regression model is a scoring system which is a multivariate function of the baseline covariate Z = z approximating the treatment effect for individual patients. As an illustrative example for the aforementioned steps, considering the difference in MRST as the treatment effect measure, one may further assume PH models for two arms separately:

$$P(T > t|R = 1, Z = z) = S_1(t)^{\exp\left(\beta_1' Z\right)} \text{ and } P(T > t|R = 0, Z = z) = S_0(t)^{\exp\left(\beta_0' Z\right)},$$

where $S_j(t),\ j = 0,\ 1$ are survival functions corresponding to the baseline hazard at group j. Then $D(z)$can be estimated by the “treatment effect score”

$$\widehat{D}(z) = \int_0^{\tau} \widehat{S_1}(t)^{\exp\left(\widehat{\beta_1}' z\right)} dt - \int_0^{\tau} \widehat{S_0}(t)^{\exp\left(\widehat{\beta_0}' z\right)} dt$$

where $\widehat{\beta}_j$and $\widehat{S}_j(t)$are estimators for $\beta_j$and $S_j(t)$under the PH model, respectively [9,23].

2.4. Validation

Validation will be conducted on Part I-V data. A scoring system for the individualized treatment effect may be developed from different combinations of regression model and measure for the treatment effect. To compare multiple scoring systems, we need to determine a treatment effect measure of the primary interest, such as HR or MRST difference, which may not necessarily be the same as those used to construct the scoring systems. Note that a scoring system always yields a ranking for all the patients, which can be used to select patients having the largest treatment benefit with respect to the selected treatment effect measure. Therefore, the quality of the ranking can be used to choose the optimal scoring system. A scoring system is considered useful if the corresponding ranking is consistent with that of the true underlying individualized treatment effect. The obstacle is that the true individualized treatment effect and its ranking are unknown, since each patient only receives one treatment. Our solution is to consider the average treatment effect for all patients with the treatment effect score above a threshold. For each fixed threshold, we have a pair of measures: the fraction of unselected patients and the average treatment effect for selected patients. By varying the threshold, the corresponding pairs can be graphically presented as the “treatment effect curve”. The treatment effect curve only depends on the ranking of the score, and a high-quality ranking yields a monotone increasing treatment effect curve: the smaller the selected patient subgroup (more selective), the bigger the average treatment effect. Fig. 2 presents an example of an estimated treatment-effect curve, suggesting that the difference of MRST between treatment and control arms is 72 days among the top 50% of patients, 57 days among the top 70% of patients, and 26 days among all patients. The Y-axis of the figure represents the estimated treatment effect in the selected subgroup based on the statistical method for estimating the treatment effect in the entire study, e.g., the nonparametrically estimated difference in MRST. Of note, the estimation method is independent of the working models employed to construct the scoring system. Since all treatment-effect curves start from the same point, $(0\%,\ \widehat{D}(1))$ with $\widehat{D}(1)$ being the estimated treatment effect for the entire population, the slope of curve can be measured by its weighted AUC defined as

$$\int_0^1 (1 - \mathrm{q})\widehat{\mathrm{D}}(\mathrm{q})\mathrm{dq}$$

where $1 - q$ is the fraction of the selected subgroup and $\widehat{D}(q)$ is the estimated treatment effect of the subgroup. This weighted AUC actually is proportional to the correlation between the score ranking and the true individualized treatment effect [23]. In addition to the weighted AUC value, the overall shape of the average treatment effect curve is also important: scoring systems with

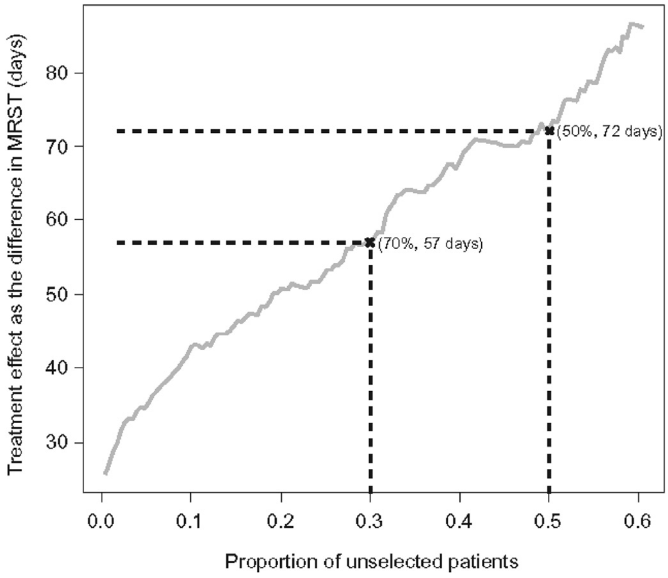

Fig. 2. Example of the treatment-effect curve.

persistent and steep monotone treatment effect curve (and thus a large AUC) are considered of high quality.

2.5. Cross-validation

Oftentimes, several different scoring systems have very similar AUCs and we may want to avoid the dependence of subtle differences on the specific split of Part I-E and Part I-V. To this end, we may employ the cross-validation method, i.e., we repeatedly split Part I into Parts I-E and I-V in a random fashion, and the optimal score will be selected based on the evaluation result in terms of the treatment effect curve averaged across all replications. Characteristics such as AUC and the size of the potential subgroup of patients with clinically meaningful treatment effect are considered in the selection process. Furthermore, the average of the estimated scores in those replications can then be used as the final score for selecting a patient subgroup for further testing. Similar to the bootstrap aggregating technique in machine learning, this scoring system ensemble often performs better than applying the selected estimation method to the entire Part I data, especially when the estimation procedure is irregular in nature and involves variable selection, etc [24,29].

Once the optimal scoring system is selected, one may determine a threshold level such that all patients with a score above the threshold are considered to have promising treatment benefit. The selection of this threshold needs to balance factors such as the size of the selected subgroup, the magnitude of the expected average treatment effect, and the likelihood of successful verification in the testing stage. In the simplest term, one may want to select the largest subgroup with a reasonable treatment effect, which can be detected by the testing data with sufficient power.

Once an effective scoring system is obtained, it is important to have a transparent interpretation to the score. Since the resulting scoring system could be a complex multivariate function of baseline features, there is no simple method to decipher the contribution role of each individual feature. If we denote the treatment effect score by a multivariate function $D(Z) = D(z_1, z_2, \cdots, z_p)$, one way to measure the importance of individual features is to use

$$I_i = E\left\{ \left[\frac{\partial D(z_1, \cdots, z_p)}{\partial z_i} \right]^2 \times var(Z_i) \right\},$$

One can see, for example, if linear regression is involved, then

$$D(z_1, \cdots, z_p) = \beta_1 z_1 + \cdots + \beta_p z_p,$$

$$I_i = \beta_i^2 var(Z_i) = \tilde{\beta}_i^2,$$

where $\tilde{\beta}_i$ is the standardized regression coefficient for covariate Z_i normalized to have a unit variance. In general, I_i approximately measures the average change in the score caused by a "typical size" perturbation of the *i*th input feature, i.e., one standard deviation change in the feature value. Furthermore, the marginal expectation

$$\mu_i(z) = E\{D(Z)|Z_i = z\}$$

can be used to summarize the role of individual features in the score [25]. However, even with measures such as I_i and $\mu_i(z)$, it still may be difficult to decipher the contribution of each individual feature and understand the underlying mechanism of the score, since input features affect output score jointly

2.6. Testing step

Using the optimal scoring system and threshold determined via the training data (Part I), we identify the "promising" subgroup of patients in the independent test set (Part II). We then conducted a statistical analysis to evaluate the treatment effect in this subgroup of patients, and compare it with the complementary subgroup using an interaction test. The point estimates, CIs, and p-values for the treatment effect and its interactions, can be computed as in the standard statistical analysis for a clinical trial. The testing step is based on a fresh test set and independent of the complex analytical procedure used for identifying the promising subgroup. The statistical analyses used in testing stage are analogous to those employed in the primary analysis for treatment effect for the entire study, which is based on minimum model assumption and generates results with causal interpretation. The testing is only conducted for the finally selected subgroup to avoid false positives and potential biases from multiple testing and model selections.

2.7. Operational considerations

The entire procedure is graphically summarized in Fig. 3. In general, the analysis is conducted in the order of estimation, cross-validation, subgroup identification, and final testing. There are several important practical issues to consider in implementing the procedure. The first step is to select the training and testing sets (Part I and Part II, respectively). There is a delicate tradeoff in allocating samples between two sets: while assigning too few samples to Part I data may harm the chance of identifying a good subgroup, a small Part II set lacks power to validate underlying treatment benefit even in a "good" subgroup. Since we will only recommend a successfully validated subgroup, in practice we propose to determine the Part II test set first considering the treatment benefit of the targeted subgroup. For example, if we target a subgroup consisting of ~50% of the patients with a HR of 0.5, then we can choose the Part II size according to the planned power if such a desired subgroup was actually identified based on Part I data. For the same reason, we prefer to preserve high-quality data, such as those from well-conducted randomized clinical trials, to the Part II data to ensure the reproducibility of the final testing results.

Second, in the estimation stage, we need to choose working models for developing scoring systems gauging the personalized treatment effect. We have the flexibility in employing competitive models and incorporating prior knowledge about the treatment effect since all the resulting scoring systems will be evaluated via cross-validation.

Third, in implementing the cross-validation for identifying the subgroup of interest based on Part I data, one needs to repeatedly separate Part I data into Part I-E and Part I-V for estimation and validation, respectively. Since the validation will be made by aggregating results from multiple Part I-Vs, we recommend assigning most (85–95%) of the data to Part I-E to maximize the chance of successfully constructing a high quality scoring system and identifying corresponding subgroup of patients. Next, despite the potentially large number of working models used at the estimation step, it is advisable to select a consistent and interpretable metric for validating the competing scoring systems based on Part I-Es and testing the final subgroup based on the Part II data. We used the difference in RMST as an illustrative example, but one may use difference in survival probability or HR depending on the application context. We don't encourage to select subgroups by directly examining all potential combinations of the estimated score and corresponding cut-off value since a scoring system of poor quality may still generate a promising subgroup by chance, which would lack reproducibility in independent data. Therefore, in general, we suggest to select the best scoring system first and then identify the corresponding subgroup. In this step, we do not recommend using a single rigid criterion such as the targeted

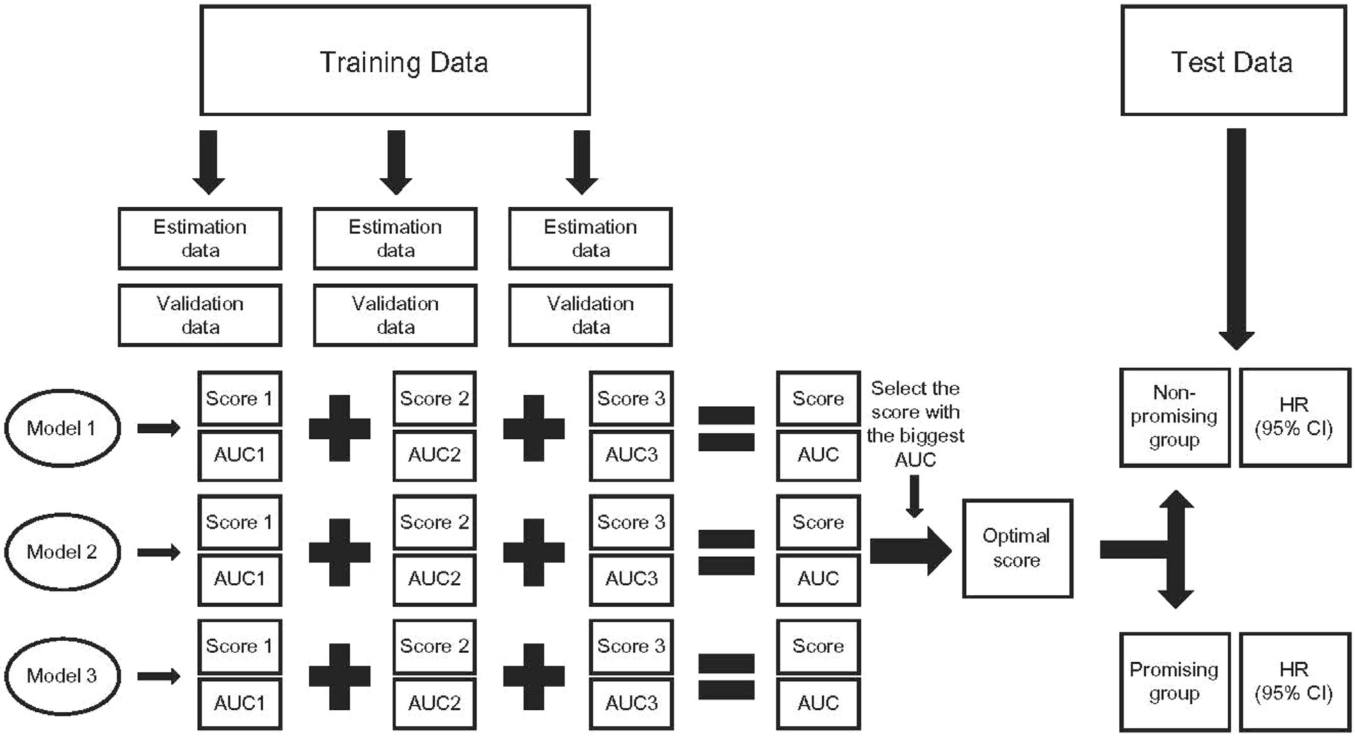

Fig. 3. The flow chart of the procedure for selecting a subgroup of patients with promising treatment effect.

treatment effect size in determining the final subgroup of patients. Instead, multiple criteria such as sensitivity to the cut-off values and subgroup sizes, need to be considered. For example, one may be willing to trade larger observed treatment effects with a larger subgroup size and more robust cut-off values. However, although the decision process for selecting the subgroup can be complex and adaptive to the application, once it is selected, one cannot change it based on the testing results, since this practice would render the statistical inference from Part II data invalid.

3. Example

3.1. Data

The Part I and II data are from the randomized clinical trials AVADO and AVEREL, respectively. The randomized, double-blind, clinical trial AVADO was conducted to test the efficacy and safety of the combination of bevacizumab and docetaxel among patients with HER2-negative metastatic breast cancer [26]. The study recruited 736 patients with 241 randomized into the control arm receiving placebo and docetaxel (100 mg/m^2), 248 into the low-dose arm receiving bevacizumab (7.5 mg/kg) and docetaxel (100 mg/m^2), and 247 into the high-dose arm receiving bevacizumab (15 mg/kg) and docetaxel (100 mg/m^2). The primary endpoint of the study is the progression-free survival, i.e., the time from randomization to the first documented disease progression or death. Patients who did not experience any disease related to progression or die during the study were right censored at the last tumor assessment at which they were known to be progression free. We consider the following 10 biomarkers baseline features for characterizing the subgroup of patients who may benefit from the treatment: platelet-derived growth factor C, vascular endothelial growth factor (VEGF)-A, VEGF-C, fibroblast growth factor (FGF)2, fms-like tyrosine kinase (FLT)1, FLT4, interleukin-8, kinase insert domain receptor, intracellular adhesion molecule-1, and E-SELECTIN. The 10 biomarkers were measured for the 345 patients enrolled in the study.

The randomized, open-label, clinical trial AVEREL was conducted to test the efficacy and safety of bevacizumab in combination with trastuzumab/docetaxel in patients with human epidermal growth factor receptor (HER)2-positive metastatic breast cancer [27]. In AVEREL, 208 patients had been randomized into the treatment arm receiving bevacizumab (15 mg/kg) plus docetaxel (100 mg/m^2) plus trastuzumab (8 mg/kg followed by 6 mg/kg), and 216 patients to the control arm receiving docetaxel (100 mg/m^2) plus trastuzumab (8 mg/kg followed by 6 mg/kg). The outcome of interest is progression-free survival as in AVADO. The 10 biomarkers used to characterize the patient subgroup were measured for 158 patients in the AVEREL study.

3.2. Statistical analysis

Our objective was to identify patients who may benefit from bevacizumab treatment using baseline biomarker measurements. It has been reported that the baseline level of VEGF-A may be predictive for the treatment benefit from bevacizumab [28]. However, we plan to empirically construct a rule for selecting patients with promising treatment benefit based on a panel of biomarkers without any special treatment of VEGF-A, since our purpose here is to use the AVADO and AVEREL studies to illustrate our method in general and such knowledge on predictive properties of individual biomarkers may not be available in other situations. The AVADO study was used as a training set (Part I). We focus on the comparison between the high-dose and control arms only. The rational of this decision is (1) the high-dose bevacizumab showed superior PFS compared with placebo while the benefit of low-dose bevacizumab was less pronounced in the AVADO study; (2) it is desirable to identify patients who benefitted from a clearly defined treatment regimen including the dosage and (3) the dosage of bevacizumab in the AVEREL study was 15 mg/kg, the same as that in the high-dose arm of the AVADO study. Each arm consists of 115 patients with complete biomarker measurements. In order to construct a scoring algorithm approximating the treatment effect for individual patients, we employ 10 different regression models in the training stage. The 10 regression models are listed in the Appendix. Several of the proposed regression models, such as boosting method coupled with the modified covariate approach for Cox models as well as the accelerated failure time model, are new [14]. However, we will not discuss their implementation details as well as theoretical justifications in the main text since it is not the focus of this paper. Indeed, with more and better methods to be developed in the future, this list of 10 regressions can grow as needed. The optimal scoring algorithm is selected based on the estimated treatment effected curves with 2000 cross-validation replications. We classify the patient population into two groups based on the selected optimal treatment effect scores: the "promising" subgroup consists of patients with the top 70% individualized treatment effect scores, and the "not-promising" subgroup consists of patients with the bottom 30% "individualized treatment effect" scores. The proportions of promising and not-promising patients are chosen because of our preference to a larger promising subgroup and the empirical observation that the cross-validated treatment effect estimator only slightly reduced when the proportion of promising patients ranges from 50% to 70%. In the testing stage, we classify the patients in the AVEREL study into two subgroups using the same classification rule and separately estimate the treatment effect. The HRs and 95% CIs are obtained. We also compare the performance of this procedure with simply using VEGF-A levels with the median as the cut-point for identifying the subgroup of patients with promising treatment effect [28].

4. Results

In the training step, we estimate the survival functions for the high-dose and control arms in the AVADO study, the result is plotted in Fig. 4. The estimated HR is 0.680 (95% confidence interval [CI]: 0.488–0.947, p = 0.023) favoring the treatment arm. In the cross-validation step, the optimal score is selected by examining both the shape of the resulting average treatment effect curve and its weighted? AUC. The chosen final score is constructed from the MRST-based regression model with non-parametric covariate treatment interaction,

$$E(\log(\min(T,\ \tau))|R = r,\ Z = z) \\ = \alpha_0 + \alpha_1 r + \alpha_2' z + D(z) \times (2r - 1)$$

with $D(z)$ $D(z)$ being approximated by a boosting algorithm of aggregating a set of adaptively constructed depth-2 classification trees [29]. The analytic form of the resulting score function is too complicated to present, but its computation for any given patient is easy and fast with a computer. Based on the cross-validated treatment effect score in the training data, we split the 230 patients into "promising" and "not-promising" groups of 161 and 69 patients, respectively. The estimated HR is 0.620 (95% CI: 0.420–0.914, p = 0.016) and 0.913 (95% CI: 0.472–1.765, p = 0.786) in the "promising" and "not-promising" groups, respectively (Fig. 5). The results of simple subgroup analysis of the AVADO study using selected baseline characteristics were in general consistent with results for the overall study population and failed to pinpoint a subgroup with comparable size and treatment benefit as ours [26]. A subsequent paper reports that VEGF-A and VEGFR-2 are potential predictive markers for bevacizumab efficacy [28]. However, the findings are based on simple post-hoc analysis and need independent validations. Note that our results may also be overly optimistic because they are based on the same training dataset (Part I) used to select the optimal scoring system. The independent test data (Part II data) consist of 158 patients from the AVEREL study with complete biomarker information. The distributions of biomarker levels in AVEREL and AVADO studies are compared in Supplemental Table 1. We calculate the treatment effect score for all 158 patients in the test set. A total of 123 patients are assigned to the "promising" group and 35 patients are assigned to the "not-

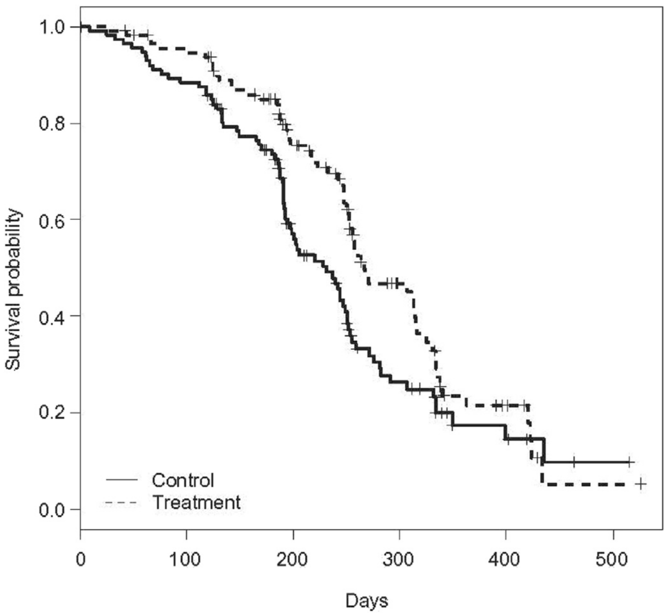

Fig. 4. The estimated survival curves for progression-free survival in the high-dose treatment arm and the control arm in the AVADO study.

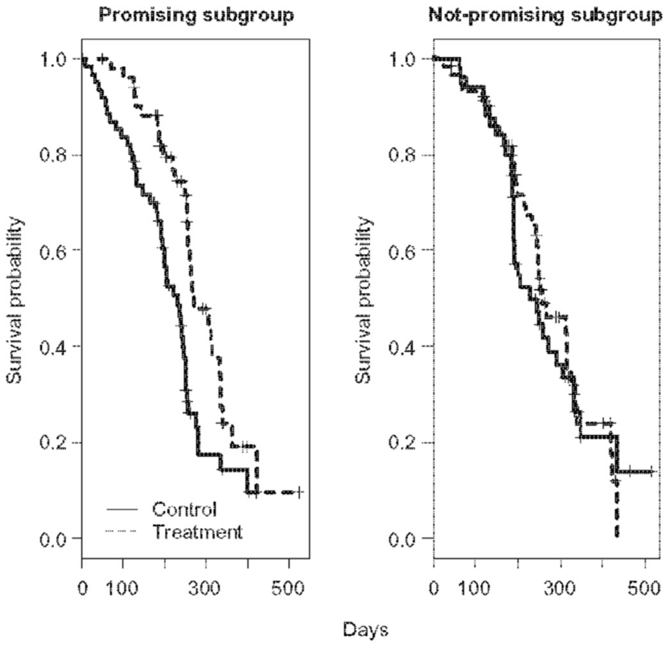

Fig. 5. The estimated survival curves of progression-free survival in the "promising" and "not-promising" subgroups determined by the estimated personalized treatment effect in the AVADO study.

promising" group. We estimate the survival functions of the treatment and control arms in both subgroups (Fig. 6). The HR is 0.687 (95% CI: 0.462–1.024, p = 0.065) and 1.152 (95% CI: 0.526–2.524, p = 0.723) for the "promising" and "not-promising" subgroups, respectively. We also plotted the estimated survival curves if bevacuzumab is given to (1) all patients; (2) only the selected patients based on the proposed method and (3) no patient (Supplemental Fig. 7). This. suggests that the benefit of treating all patients with bevacizumab can be maintained by targeting only the top 70% of patients according to the estimated treatment scoring system. The benefit of treating additional 30% of the patients is minimal and need to be balanced against the cost and the risk of potential adverse events.

If we use the median of VEGF-A in AVEREL study to divide the study population, then the estimated HR is 0.711 (0.435–1.163, p = 0.174) in the "promising" group and 0.828 (0.496–1.380, p = 0.468) in the "not-promising" group. The results are similar if we use the median of VEGF-A in AVADO study to divide the population: the estimated HR is 0.709 (0.444–1.133, p = 0.151) in the "promising" group of 87 patients and 0.851 (0.497–1.458, p = 0.556) in the "not-promising" group of 71 patients. This result is clearly inferior to that based on the constructed scores. The median VEGF-A is chosen as the cut-off value out of convenience as well as based on preliminary analysis results of the AVADO study. There may be other cut-off values of VEGF-A generating better results. How to search and validate the "optimal" cut-off values of VEGF-A is analogous to the problem of identifying a good threshold value for the estimated personalized treatment effect score, in which many factors need to be considered in the context of the application.

Here the constructed score approximating the individualized treatment effect is a complex function of the 10 baseline biomarkers but can be conveniently computed using a personal computer. In our scoring system, VEGF-A, FLT1, and FGF2 are the top three most important biomarkers. Furthermore, the marginal expectation with respect to VEGF-A is monotone increasing, suggesting that patients with higher VEGF-A level tend to have a bigger treatment benefit, which is in consistent with our prior knowledge on VEGF-A. It takes about 6 h to complete the entire analysis on a PC with intel 3.40 GHz CPU and 16 GB RAM using R.

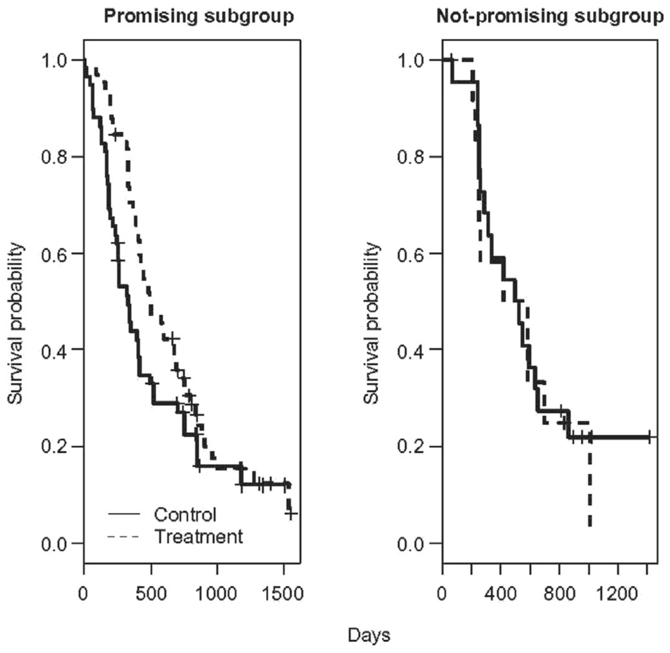

Fig. 6. The estimated survival curves of progression-free survival in the "promising" and "not-promising" subgroups determined by the estimated treatment effect of individual patient in AVEREL study.

5. Discussion

There is substantial recent development in statistical methodology for personalized medicine. However, how to appropriately choose and apply these methods in practice remains a great challenge. Different methods have different merits and limitations. For example, while the tree-based learning method tends to select over-simplified subgroups and thus, have unsatisfactory performance, the performance of the more sophisticated outcome-weighted learning methods depends on the choice of target function and the machine learning algorithms optimizing it. Furthermore, the statistical validity of many methods such as asymptotical consistency often relies on theoretical assumptions, whose verifications are difficult in practice. In this paper, we have argued that the selection of the methods can be achieved empirically without the need of examining all the theoretical assumptions for each of the models. We have also highlighted the importance of robust, independent validation with minimum model assumptions. We have demonstrated the merit this systematic approach to identify subgroups of patients who may benefit from a treatment based on data from two oncology trials that tested the efficacy of bevacizumab. Specifically, we constructed a scoring algorithm assessing the treatment effect of individual patients and the independent testing data supported that the treatment effect is higher among patients with higher estimated treatment effect scores.

In the training stage, we treat the different regression models for deriving the treatment effect scores as working models employed for convenience. Different regression models can be used at the same time without assuming any one of them as the true model. In practice, one can have substantial flexibility in expanding the tool box by adding novel regression or machine learning methods to approximate the individualized treatment effect. For example, when the number of features is large, such as gene expression levels, one may apply various regularization methods for feature selection in the regression analysis [30,31].

The independence between the training and test sets is important. It ensures that all statistical inferences, including estimation and hypothesis testing, conducted in the test set are valid and can be interpreted in the conventional manner, regardless of the novel techniques employed in the training stage [32–34]. Ideally, the statisticians performing the analysis for the training set should remain blinded to the test data until the testing stage. The testing step with independent data cannot be replaced by the commonly used cross-validation. First, the cross-validation result was already used for selecting the optimal scoring system and, therefore, the performance of the selected scoring system from cross-validation is prone to bias in the optimistic direction, especially when the number of candidate scoring systems is not small. Furthermore, there is no valid inference procedure for the treatment effect in subgroups based on models selected with the cross-validation procedure.

As a major limitation, the method requires data from two independent clinical trials (or a large trial allowing splitting the data). It may not be feasible in many practical applications. The essence of the problem is lack of power in detecting the treatment-covariate interaction. In the example presented here, the interaction between the treatment and group assignment ("promising" vs. "not-promising") in the test set (AVEREL study) is not statistically significant (p = 0.225), which is likely due to the limited power/small sample size. It highlights the need for cautious interpretation and further verification of the final result. Another limitation is the selection of the appropriate cut-off value of the estimated scoring system. Since many factors need to be considered, there is no rigorous data-driven procedure to automatically give optimal cut-off values and, therefore, the choice is still by and large subjective. In addition, the complexity and the "black box" nature of the scoring algorithm, which prevent a straightforward interpretation of the results, could be perceived as a hurdle to its application, especially in a medical context. Understanding the relative importance of the different features in the final score is important. In some cases, it may be possible to construct "simplified" scoring algorithms based only on a small subset of the most important features.

In summary, when we want to construct and recommend a rule for selecting subgroups of patients with promising treatment effect, many important, sensitive issues need to be carefully considered, such as the strength of the evidence for the presence of heterogeneous treatment effect from the testing data, the understanding of potential mechanism of the rule, and the consistency with our prior knowledge. Therefore, while it may be advantageous that the proposal is very flexible and adaptive, the implementation is not automatic. The successful discovery of the targeted subgroup depends on good choice on issues such as training and testing splitting, candidate estimation procedures, and the metric for the treatment effect. It is unfortunate that there is no universal answer to these important questions. Carefully designed simulation studies can be conducted and are helpful to examine the empirical performance of the proposal and provide practical guidelines on implementing the proposal under different settings. The analysis presented in this paper is limited and intended as proof of concept for the proposed statistical and machine learning approach. The bevacizumab data were used strictly as an example for illustrative purposes. The results presented herein are purely exploratory and should not be interpreted in terms of a recommendation for clinical practice.

Lastly, we want to emphasize the fact that, although the automatic statistical modeling and machine learning methods are powerful tools in identifying multi-biomarker signatures for approximating the treatment effect of individual patients and in identifying subgroups of patients with promising treatment effect, they are not replacements for subject-matter knowledge. Since there is always a price to pay in eliminating "noise", i.e., irrelevant features, relevant prior knowledge of the underlying biological mechanism may greatly help to boost the performance of the procedure by focusing on truly important biomarkers and patient characteristics.

Funding sources

This study was mainly funded by Genentech, Inc., which was involved in the study design, collection, analysis, and interpretation of data; writing of the report; and the decision to submit for publication. The research of Dr. Tian was also partially supported by NIH grant R01 HL089778-05.

Conflict of interest

PD is a full-time employee of F. Hoffmann-La Roche Ltd. CI is a full-time employee of Genentech, Inc. LT is a full-time employee of Stanford University.

Acknowledgments

The authors thank Stefan Scherer for supporting the work and for inspirational discussion. Third-party editorial assistance for this manuscript was provided by CodonMedical, an Ashfield Company, and was funded by Genentech, Inc.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.conctc.2017.01.007.

Appendix. The regression models used to estimate the personalized treatment effect

(1) PH model for two arms separately, i.e., we assume

$$P(T>t|R=1,Z=z)=S_1(t)^{\exp\left(\beta_1' Z\right)}$$

$$P(T>t|R=0,Z=z)=S_0(t)^{\exp\left(\beta_0' Z\right)}$$

where $S_j(t),\ j=1,\ 2$ are survival functions corresponding to the baseline hazard. The "treatment effect score" measuring the difference in MRST is

$$D(z)=\int\limits_0^\tau S_1(t)^{\exp\left(\beta_1' Z\right)}dt-\int\limits_0^\tau S_0(t)^{\exp\left(\beta_0' Z\right)}dt.$$

(2) PH model with treatment-covariate interaction

$$P(T>t|R=r,Z=z)=S_0(t)^{\exp\left(\alpha_0' z+\alpha_1 r+\beta_0' z\times(2r-1)\right)}$$

where $S_0(t)$ is the baseline survival functions. The "treatment effect score" is $D(z)=\beta_0' z$.

(3) PH model with modified covariates

$$P(T>t|R=r,Z=z)=S_0(t)^{\exp\left(\left(\alpha_1+\beta_0' z\right)\times(2r-1)\right)}.$$

The "treatment effect score" $D(z)=\beta_0' z$.

(4) PH model with non-parametric covariate and treatment interactions

$$P(T>t|R=r,Z=z)=S_0(t)^{\exp(g(z)\times(2r-1))},$$

where $g(z)=\sum w_{jk}I(x_j\geq c_j,\ x_k\geq c_k),\ (x_1,x_2,\cdots,x_p,\ x_{p+1},\cdots,\ x_{2p})=(z_1,z_2,\cdots,\ z_p,\ -z_1,\cdots,\ -z_p)$. The "treatment effect score" $D(z)=g(z)$.

(5) MRST regression model for two arms

$$E(\min(T,\ \tau)|R=1,\ Z=z)=\exp\left(\alpha_1+\beta_1' z\right)$$

$$E(\min(T,\ \tau)|R=0,\ Z=z)=\exp\left(\alpha_0+\beta_0' z\right),$$

The "treatment effect score" $D(z)=\exp(\alpha_1+\beta_1' z)-\exp(\alpha_0+\beta_0' z)$.

(6) MRST regression model with covariate treatment interactions:

$$E(\min(T,\ \tau)|R=r,Z=z)=\exp\left(\alpha_0+\alpha_1^{'}z+\alpha_2r+\beta_0^{'}z\times(2r-1)\right).$$

The “treatment effect score” $D(z)=\beta_0^{'}z$.

(7) MRST regression model with modified covariate:

$$E(\min(T,\ \tau)|R=r,\ Z=z)=\exp\left(\left(\alpha_0+\beta_0^{'}z\right)\times(2r-1)\right).$$

The “treatment effect score” $D(z)=\beta_0^{'}z$.

(8) The log-transformed MRST regression model with nonparametric covariate treatment interactions

$$E(\log\{\min(T,\ \tau)\}|R=r,\ Z=z) = \alpha_0+\alpha_1r+\alpha_2^{'}z+g(z)\times(2r-1),$$

where $g(z)=\sum w_{jk}I(x_j\geq c_j,\ x_k\geq c_k)$, $(x_1,x_2,\cdots,x_p,\ x_{p+1},\cdots,\ x_{2p})=(z_1,z_2,\cdots,\ z_p,\ -z_1,\cdots,\ -z_p)$. The “treatment effect score” $D(z)=g(z)$.

(9) The log-transformed MRST regression model with covariate treatment interactions:

$$E(\log\{\min(T,\ \tau)\}|R=r,Z=z)=\alpha_0+\alpha_1^{'}z+\alpha_2r+\beta_0^{'}z\times(2r-1).$$

The “treatment effect score” $D(z)=\beta_0^{'}z$.

(10) The log-transformed MRST regression model with modified covariate:

$$E(\log\{\min(T,\ \tau)\}|R=r,\ Z=z)=\left(\alpha_0+\beta_0^{'}z\right)\times(2r-1).$$

The “treatment effect score” $D(z)=\beta_0^{'}z$.

For fitting PH models in (1), (2), and (3), we maximize the lasso-regularized log-partial likelihood functions. For fitting model (4), we use the gradient boosting algorithm to maximize the partial likelihood function with tree as the base learner. For fitting the MRST regression models in (5), (6), and (7), we maximize the inverse-probability weighted loss function with lasso-regularization, where the loss function is in the form of $\min(T,\ \tau)(\alpha+\beta^{'}z)-\exp(\alpha+\beta^{'}z)$.

For fitting the log-transformed MRST in (8), we minimize the inverse-probability weighted squared loss function with the gradient boosting algorithm using tree as the base learner. For fitting the log-transformed MRST regression models in (9) and (10), we minimize the inverse-probability weighted squared loss function with lasso-regularization.

References

[1] S.T. Brookes, E. Whitley, T.J. Peters, P.A. Mulheran, M. Egger, G. Davey Smith, Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false- negatives, Health Technol. Assess. 5 (2001) 1–56.

[2] S.W. Lagakos, The challenge of subgroup analyses—reporting without distorting, N. Engl. J. Med. 354 (2006) 1667–1669.

[3] S.T. Brookes, E. Whitely, M. Egger, G.D. Smith, P.A. Mulheran, T.J. Peters, Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test, J. Clin. Epidemiol. 57 (2004) 229–236.

[4] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. B 57 (1995) 125–133.

[5] L. Gunter, J. Zhu, S. Murphy, Variable selection for qualitative interactions in personalized medicine while controlling the family-wise error rate, J. Biopharm. Stat. 21 (2011) 1063–1078.

[6] O.J. Dunn, Multiple comparisons among means, J. Am. Stat. Assoc. 56 (1961) 52–64.

[7] J.C. Foster, J.M. Taylor, S.J. Ruberg, Subgroup identification from randomized clinical trial data, Stat. Med. 30 (2011) 2867–2880.

[8] L. Tian, R. Tibshirani, Adaptive index models for marker-based risk stratification, Biostatistics 12 (2011) 68–86.

[9] T. Cai, L. Tian, P.H. Wong, L.J. Wei, Analysis of randomized comparative clinical trial data for personalized treatment selections, Biostatistics 12 (2011) 270–282.

[10] M. Qian, S. Murphy, Performance guarantees for individualized treatment rules, Ann. Stat. 39 (2011) 1180–1210.

[11] Y. Zhao, D. Zeng, A.J. Rush, M.R. Kosorok, Estimating individualized treatment rules using outcome weighted learning, J. Am. Stat. Assoc. 107 (2012) 1106–1118.

[12] W. Lu, H.H. Zhang, D. Zeng, Variable selection for optimal treatment decision, Stat. Methods Med. Res. 22 (2013) 493–504.

[13] C. Kang, H. Janes, Y. Huang, Combining biomarkers to optimize patient treatment recommendations, Biometrics 70 (2014) 695–707.

[14] L. Tian, A. Alizadeh, A. Gentles, R. Tibshirani, A simple method for estimating interactions between a treatment and a large number of covariates, J. Am. Stat. Assoc. 109 (2014) 1517–1532.

[15] E.B. Laber, Y. Zhao, Tree-based methods for individualized treatment regimes, Biometrika 102 (2015) 503–514.

[16] Y. Zhao, D. Zheng, E.B. Laber, M.R. Kosorok, New statistical learning methods for estimating optimal dynamic treatment regimes, J. Am. Stat. Assoc. 110 (2015) 583–598.

[17] Y. Zhang, E.B. Laber, A. Tsiatis, M. Davidian, Using decision lists to construct interpretable and parsimonious treatment regimes, Biometrics 71 (2015) 895–904.

[18] J. Weiss, F. Kuusisto, K. Boyd, J. Liu, D. Page, Machine learning for treatment assignment: improving individualized risk attribution, AMIA. Annu. Symp. Proc. (2015) 1306–1315.

[19] R. Song, M.R. Kosorok, D. Zeng, Y. Zhao, E.B. Laber, M. Yuan, On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning, Stat 4 (2015) 59–68.

[20] D.R. Cox, Regression models and life-tables, J. R. Stat. Soc. B 34 (1972) 187–220.

[21] L. Zhao, L. Tian, H. Uno, S.D. Solomone, M.A. Pfeffere, J.S. Schindler, et al., Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study, Clin. Trials 9 (2012) 570–577.

[22] H. Uno, B. Claggett, L. Tian, E. Inoue, P. Gallo, T. Miyata, et al., Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis, J. Clin. Oncol. 32 (2014) 2380–2385.

[23] L. Zhao, L. Tian, T. Cai, B. Claggett, L.J. Wei, Effectively selecting a target population for a future comparative study, J. Am. Stat. Assoc. 108 (2013) 527–539.

[24] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.

[25] T. Hastie, R. Tibshirani, J. Friedman, Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer, New York, 2009.

[26] X. Pivot, A. Schneeweiss, S. Verma, C. Thomssen, J.L. Passos-Coelho, G. Benedetti, et al., Efficacy and safety of bevacizumab in combination with docetaxel for the first-line treatment of elderly patients with locally recurrent or metastatic breast cancer: results from AVADO, Eur. J. Cancer 47 (2011) 2387–2395.

[27] L. Gianni, G.H. Romieu, M. Lichinitser, S.V. Serrano, M. Mansutti, X. Pivot, et al., AVEREL: a randomized phase III Trial evaluating bevacizumab in combination with docetaxel and trastuzumab as first-line therapy for HER2-positive locally recurrent/metastatic breast cancer, J. Clin. Oncol. 31 (2013) 1719–1725.

[28] D.W. Miles, S.L. de Haas, L.Y. Dirix, G. Romieu, A. Chan, X. Pivot, et al., Biomarker results from the AVADO phase 3 trial of first-line bevacizumab plus docetaxel for HER2-negative metastatic breast cancer, Br. J. Cancer 108 (2013) 1052–1060.

[29] R.E. Schapire, The strength of weak learnability, Mach. Learn. 5 (1990) 197–227.

[30] R. Tibshirani, The lasso method for variable selection in the Cox model, Stat. Med. 16 (1997) 385–395.

[31] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (2010) 1–22.

[32] S.G. Baker, D.J. Sargent, Designing a randomized clinical trial to evaluate personalized medicine: a new approach based on risk prediction, J. Natl. Cancer Inst. 102 (2010) 1–4.

[33] B. Freidlin, W. Jiang, R. Simon, The cross-validated adaptive signature design, Clin. Cancer Res. 16 (2010) 691–698.

[34] L.M. McShane, M.-Y.C. Polley, Development of omics-based clinical tests for prognosis and therapy selection: the challenge of achieving statistical robustness and clinical utility, Clin. Trials 10 (2013) 653–665.