Chapter 9
A New Paradigm for Subset Analysis in Randomized Clinical Trials

Richard Simon and Noah Simon

Abstract There are numerous methods for identifying subsets of patients in a randomized clinical trial who appear to benefit from the test treatment to a greater or lesser extent than average. Generally such claims are based multiple hypothesis testing and re-substitution estimates of treatment effect that are known to be highly optimistically biased. In this chapter we describe a new paradigm for subset analysis. Rather than being based on multiple hypothesis testing, it is based on training a single predictive classifier and provides an almost unbiased estimate of treatment effect for the selected subset.

Keywords Predictive classifier · Re-sampling · Pre-validation · Personalized treatment · Bootstrap sampling · Cross-validation · Subset analysis

9.1 Introduction

The main objective of most randomized clinical trials is to determine whether the test treatment is beneficial on average for the population of all eligible patients with regard to the primary endpoint. For biologically heterogeneous diseases like most forms of cancer, it has become increasingly apparent that for most treatments the treatment effect is not uniform across the eligible population. Consequently the average treatment effect is an imperfect guide for basing treatment strategies and there is often interest in identifying subsets of patients who have treatment effects greater than or less than the average. Statisticians often dismiss this objective as "exploratory" because they are not familiar with reliable methods which can perform discovery and inference on the same dataset. This problem of reliably

R. Simon (✉)
R Simon Consulting, Bethesda, MD, USA

N. Simon
Department of Biostatistics, University of Washington, Seattle, WA, USA

N. Ting et al. (eds.), *Design and Analysis of Subgroups with Biopharmaceutical Applications*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-030-40105-4_9

characterizing treatment effect heterogeneity is not a hypothesis testing problem although it is often treated as if it were.

There is no lack of subset identification methods. Because of the high false positive rate for tests of treatment effect in subsets selected from the data, such analyses usually elicit skepticism and are viewed as hypothesis generation to be tested on independent data. Often however such independent data is not available. In this chapter we shall describe a new paradigm for subset analysis based on developing a "predictive classifier" (Freidlin and Simon 2005). We shall also describe how this predictive classifier can be internally validated using measures of performance appropriate for classifiers.

9.2 Methods

9.2.1 Predictive Classifiers

Let D denote the data from a randomized clinical trial comparing a test treatment to a control regimen. The data consists of covariate vectors (X) for the patients, treatment indicators (z) and outcomes (y). If our clinical trial is "negative" with regard to average treatment effect for all eligible patients, then our objective may be to identify and validate a subset of patients who benefit from the test treatment. If our clinical trial is "positive" overall, the objective may be to identify an "intended excluding patients" who do not seem to benefit from the test treatment. More generally, we may want to stratify the population with regard to the likelihood that they benefit from the test treatment.

A predictive classifier is not like the usual prognostic classifier relating baseline covariates to prognosis. When there are two treatments, a predictive classifier is a function which indicates whether the patient is likely to benefit from the test treatment or not. Here we will discuss tri-level classifiers with $\mathrm{C(X)} = 2$ indicating that a patient with covariate vector X is very likely to benefit from the test treatment, $\mathrm{C(X)} = 1$ meaning that the patient is moderately likely to benefit and $\mathrm{C(X)} = 0$ meaning that the patient is unlikely to benefit or may have better outcome on the control treatment.

We may denote the classifier as $\mathrm{C}\,(\mathrm{X}; \mathcal{A}, \mathrm{D})$ meaning it is a function of the covariate vector X and that it was developed by applying a predictive classifier development algorithm $\mathcal{A}$ to the dataset D. Specifying $\mathcal{A}$ means that the user is required to specify in advance the types of analyses that will be performed to develop a fully specified classifier. This is essential for using re-sampling methods for evaluating classifiers because the same classifier development algorithm must be applied to several re-sampled training sets.

A predictive classifier is not a "risk classifier". Instead it classifies patients with regard to their likelihood of benefit from the test treatment relative to the control regimen. Predictive classifiers have been called "regimes" by some investigators (Bai et al. 2017).

There are many types of predictive classifiers. For example one could develop separate prognostic models for the test treatment T and for the standard treatment S. Denoting these models as f(X;T) and f(X;S), they provide expected outcome or a function of expected outcome for a patient with covariate vector X. These models might be based on penalized logistic regression, random forest, support vector machines, etc. Our predictive classifier C might be defined based on these models as

$$\begin{aligned} &\mathrm{C}\,(\mathrm{X};\mathcal{A},\mathrm{D}) = 2 \ \text{ if } \ \mathrm{f}\,(\mathrm{X};\mathrm{T})\text{–}\mathrm{f}\,(\mathrm{X};\mathrm{S}) > \mathrm{k}_2 \\ &\mathrm{C}\,(\mathrm{X};\mathcal{A},\mathrm{D}) = 1 \ \text{ if } \ \mathrm{k}_2 > \mathrm{f}\,(\mathrm{X};\mathrm{T})\text{–}\mathrm{f}\,(\mathrm{X};\mathrm{S}) > \mathrm{k}_1 \\ &\mathrm{C}\,(\mathrm{X};\mathcal{A},\mathrm{D}) = 0 \ \text{ otherwise.} \end{aligned} \tag{9.1}$$

The set of covariate vectors

$$\mathcal{S}_2 = \{X : C\,(X;\mathcal{A},D) = 2\}$$

might be taken as the intended use population for the new treatment. $\mathcal{S}_1$ and $\mathcal{S}_0$. can be analogously defined. The characterization of the covariate vectors in these subsets can be used for product labeling if T is a new treatment. Otherwise the subsets can be used for patient management; i.e. patients with covariate vectors in $\mathcal{S}_2$ would generally receive the test treatment and those with covariate vectors in $\mathcal{S}_0$ generally would receive the control. For patients with covariate vectors in $\mathcal{S}_{1,}$ treatment selection would be influenced by secondary endpoints and patient preference. The constants k_1 and k_2 can be specified based on clinical significance, cost or adverse effects of the test treatment. For example, with survival outcome k_2 might be defined as the natural logarithm of 0.90 taking a 10% decrease in hazard as minimally clinically significance. Defining k_1 as zero would identify $\mathcal{S}_2$ as the class in which expected outcome on the control is better than on the test treatment.

With survival modeling, one might fit a proportional hazards model

$$\log \frac{h\,(t;X,z)}{h_0(t)} = \alpha z + z\beta' X + (1-z)\,\gamma' X$$

where z is a (0,1) treatment indicator. The treatment effect on the log hazard ratio scale is the value of the log hazard ratio for z = 1 minus the value for z = 0; that is $\alpha + (\beta - \gamma)^{'}X$.The three class classifier described above is C = 2 if $\alpha + (\beta - \gamma)^{'}X \le k_2$ and the other classes defined similarly (sign reversed because lower hazard is better). C = 1 if $k_1 \le \alpha + (\beta - \gamma)^{'}X < k_2$ and C = 0 otherwise. In classifying cases we use the maximum likelihood estimates of the parameters. The sets S_0 S_1 and S_2 thus are a partition of the cases. If there are a large number of candidate covariates, then penalized regression methods can be utilized in training the classifier.

Our objective here is not to provide advice about what types of predictive classifiers are best nor to develop a new type of predictive classifier, but to show how to internally validate a predictive classifier once it has been defined.

9.2.2 De-biasing the Re-substitution Estimates

The usual approach to subset analysis involves some type of analysis of the full dataset D to identify a subset $\mathcal{S}_2$ for which the treatment effect seems large. The empirical estimate of treatment effect for $\mathcal{S}_2$ in these circumstances is called a "re-substitution estimate". $\mathcal{S}_2$ was used as part of D for subset identification and then as the basis for computing treatment effect and this often results in a large bias in the estimate of treatment effect.

Although the re-substitution estimates of treatment effect based on the sets $\mathcal{S}_2$, $\mathcal{S}_1$, and $\mathcal{S}_0$ are biased estimates, they can be de-biased in the following manner as suggested by Zhang et al. (2017).

Let $\mathrm{D_b}$ denote a non-parametric bootstrap sample of cases and let $C_b = C\left(X; \mathcal{A}, D_b\right)$ denote the predictive classifier developed on $\mathrm{D_b}$ using the classifier development algorithm $\mathcal{A}$. Define

$$\varDelta\left(C_b, D_b\right)$$

to be the empirical average treatment effect for patients in $\mathrm{D_b}$ for whom $\mathrm{C_b} = 2$. Since $\mathrm{D_b}$ was the data on which classifier $\mathrm{C_b}$ was trained, this is a re-substitution estimate of treatment effect.

We can also use the classifier $\mathrm{C_b}$ to classify the withheld cases $\overline{D}_b = D - D_b$ i.e. those not used to develop the classifier. That classification determines $\varDelta\left(C_b, \overline{D}_b\right)$ the empirical estimate of treatment effect for the subset of the hold-out subset for which $\mathrm{C_b} = 2$. Since the hold-out set was not included with the bootstrap data used to train $\mathrm{C_b}$, $\varDelta\left(C_b, \overline{D}_b\right)$ is an unbiased estimate of the treatment effect to be expected in the future for cases with $\mathrm{C_b} = 2$. Also, the differences

$$\eta_b = \varDelta\left(C_b, D_b\right) - \varDelta\left(C_b, \overline{D}_b\right)$$

are estimates of the re-substitution bias in estimating treatment effect in $\mathcal{S}_2$ using our algorithm $\mathcal{A}$ for classifier development. These estimates can be averaged over bootstrap samples and then used to debias the re-substitution estimates. We have described it here for $\mathcal{S}_2$ but it can be done similarly for $\mathcal{S}_1$ and $\mathcal{S}_{0.}$

9.2.3 Pre-validated Estimates of Treatment Effect

An alternative approach for estimating the treatment effects is to classify each patient i using a classifier trained on a dataset not including case i. This approach was first developed for use in the Cross-Validated Adaptive Signature Design (Freidlin et al. 2010). Suppose we perform a leave-one-out cross validation. When case i is omitted we train a classifier and use it to classify the omitted case i. Let $\mathrm{C}\left(\mathrm{X_i};\mathcal{A},\mathrm{D}^{(-\mathrm{i})}\right)$ denote the classification of this omitted observation. This is called "pre-validated" classifications because each observation i is classified using a classifier trained on a dataset not containing case i.

After all the folds of the cross-validation are completed, we have pre-validated classifications for all the cases. We can thus collect together the cases classified $\mathrm{C}\left(\mathrm{X_i};\mathcal{A},\mathrm{D}^{(-\mathrm{i})}\right) = 2$. These cases define $\mathcal{S}_2$ and we can compute the empirical treatment effect within this subset. Pre-validated subsets $\mathcal{S}_1$ and $\mathcal{S}_0$ can be analyzed analogously.

We simulated clinical trials to illustrate the bias of the re-substitution estimate of treatment effect on the $\mathcal{S}_2$ subset and the effectiveness of defining $\mathcal{S}_2$ based on pre-validated classifications. The simulations involved 300 patients with exponentially distributed survival and 40 binary covariates each with equal prevalence. The intended use subset $\mathcal{S}_2$ was determined by fitting a full proportional hazards model (2). For each patient the predictive index was computed for the patient receiving the test treatment and for receiving the control. If the difference was less than -0.2 then the patient was classified in $\mathcal{S}_2$. Table 9.1 shows the results of 10 simulated clinical trials with no treatment effect. For the first three columns the classifier was trained on the full dataset and then applied to the same full data to obtain $\mathcal{S}_2$. Consequently it provides biased re-substitution estimates. Column 2 shows the hazard ratio estimates of treatment effect in these $\mathcal{S}_2$ subsets and column 3 shows the computed log-rank test statistics of treatment effect which should have a chi-square distribution on one degree of freedom for the usual setting of no treatment effect and an independent test set. It is seen that the hazard ratios are not close to 1.0 as they should be and the log-rank distribution looks shifted to larger values.

Columns 4 and 5 of Table 9.1 show results of cross-validation for the same ten simulated clinical trials. The classifiers were fit to the training sets of each fold of a tenfold cross validation. Those ten classifiers were used to classify the patients in the ten respective hold-out sets. That is, for purposes of cross-validated evaluation, the classifier used to classify a case was trained on a subset of the full dataset with that target case omitted. These cross-validation based classifiers are not used for classifying future patients, but they provide a way of evaluating the classifier developed on the full dataset that avoids the bias of the re-substitution estimator. The patients classified in $\mathcal{S}_2$ in this way were taken as constituting the pre-validated $\mathcal{S}_2$ set. The empirical treatment effect was computed on these pre-validated sets and the hazard ratios and log-rank statistics are shown. The hazard ratios are all expressed as less than 1. The estimated hazard ratios are closer to 1.0 and the log-rank statistics are smaller.

Table 9.1 Simulation of 10 null clinical trials

	Re-substitution		Cross-validated	
Trial	HR	LR-chisq	HR	LR-chisq
1	0.59	4.5	0.82	1.0
2	0.60	4.0	0.73	2.7
3	0.64	1.7	0.83	0.67
4	0.62	2.4	0.94	0.07
5	0.68	2.0	0.84	0.66
6	0.72	1.3	0.83	0.67
7	0.49	9.8	0.77	2.0
8	0.48	12.2	0.69	4.9
9	0.54	6.7	0.69	3.7
10	0.56	6.7	0.77	2.1

Estimated HR and log-rank chi-squared in adaptively determined intended use subset

Table 9.2 Simulation of 10 clinical trials with treatment effect for subset with marker 1 equal to 1

	Re-substitution		Cross-validated	
Trial	HR	LR-chisq	HR	LR-chisq
1	0.49	11.8	0.62	7.2
2	0.28	46.9	0.38	34.0
3	0.54	4.8	0.72	2.0
4	0.55	7.3	0.60	8.2
5	0.41	20.3	0.62	7.8
6	0.43	21.2	0.61	9.4
7	0.54	7.6	0.79	1.5
8	0.38	24.0	0.56	11.0
9	0.38	20.9	0.53	12.8
10	0.63	4.7	0.72	3.1

True HR $= 0.6$ in subset
Estimated HR and log-rank chi-squared in adaptively determined intended use subset

Table 9.2 shows analogous results for 10 clinical trials simulated with a treatment effect of hazard ratio 0.6 for the half of patients with covariate 1 equal to 1. The same type of proportional hazards predictive classifier was fit as before. The cross-validated chi-square values for treatment effect within the adaptively determined intended use subset is not as inflated as the re-substitution values and the hazard ratio estimates within the intended use subset are closer to the true 0.6 values used for simulating the data. The R software used to compute Tables 9.1 and 9.2 are available from the first author.

9.2.4 Testing Treatment Effects in Subsets $\mathcal{S}_2$, $\mathcal{S}_1$ and $\mathcal{S}_0$

We can estimate the expected treatment effects in these subsets as described in the previous section but we would also like to test the null hypothesis that these treatment effects are zero. We can test the null hypothesis that the expected treatment effect is zero in $\mathcal{S}_2$ by permuting the treatment assignments, re-computing the adaptively determined $\mathcal{S}_2$ and using the empirical treatment effect in the new $\mathcal{S}_2$ as a test statistic for the permutation test.

9.2.5 PPV and NPV of the Predictive Classifier

If we take classification into subset $\mathcal{S}_2$ as indicating that the patient is more likely to benefit from the new treatment, then what is the PPV and NPV of the classifier? If outcomes are survival times and the treatments have proportional hazards within each subset, then the probability that a patient classified in S_2 benefits from the test treatment is approximately

$$PPV = \frac{1}{1 + e^{\delta_2}}$$

where δ_2 is the hazard ratio of the test treatment to control in S_2. This is shown by Simon (2015) under the assumption of independence of treatment effects for a patient. Similarly, the NPV for a case classified in $\mathcal{S}_2$ is approximately

$$NPV_0 = \frac{e^{\delta_0}}{1 + e^{\delta_0}}$$

where δ_0 denotes the hazard ratio for cases in S_0. For a case classified in $\mathcal{S}_1$ the NPV is approximately

$$NPV_1 = \frac{e^{\delta_1}}{1 + e^{\delta_1}}.$$

9.2.6 Calibration of Pre-Validated Treatment Effects

The development above enables the classification of future patients into the three subsets, $\mathcal{S}_2$ representing very likely to benefit from the test treatment, $\mathcal{S}_0$, very unlikely to benefit from the test treatment and an intermediate group $\mathcal{S}_1$. The cases in $\mathcal{S}_0$ may have better outcomes on the control. This is individualized prediction because it is based on the covariate vector X. These estimates are discretized into three sets, however, and are based on the parametric prognostic models f(X,T) and f(X, S). An alternative approach is to focus on the pre-validated treatment effect difference f(X,T) − f(X,S) for each case. Then, if our outcome is survival, these difference scores can be smoothed by fitting a proportional hazards model containing a main effect of treatment. By using a spline we can estimate the relationship of difference score to treatment effect. This is similar to the approach as suggested by (Matsui et al. 2012).

Instead of fitting the proportional hazards model with the splines of the pre-validated $d_i^{(p)}$ values, a simple window smoother can possibly be used. For every small window on the d axis we compute an estimate of the hazard ratio of the two treatments. The empirical hazard ratio is $(e_1/m_1)/(e_0/m_0)$ where e_1 and e_0 denote the number of events in the window for the treatment and control groups respectively and m_1 and m_0 are the numbers of patients at risk at the start of the window for those groups. This is only used for windows for which m_1 and m_0 are both positive. This is related to the approach suggested by Cai (2011).

9.3 Discussion

In the new paradigm of subset analysis that we have described multiple hypothesis testing is replaced with the development of a single predictive classifier. We have shown how to obtain approximately unbiased estimates of the treatment effect for the set of future patients selected based on this predictive classifier and testing the significance of this treatment effect. Simulation studies have shown that the residual bias is very small (Simon and Simon 2019). We have also shown how to estimate the PPV, NPV for the predictive classifier.

The bootstrap de-biasing approach described provides a method of estimating and correcting the bias of the re-substitution estimate of treatment effect in an adaptively defined subset like $\mathcal{S}_2$. The re-substitution estimate is the empirical treatment effect in $\mathcal{S}_2$. It is biased because $\mathcal{S}_2$ was included in the application of the algorithm $\mathcal{A}$. The estimate of bias is based on comparing the re-substitution estimate for each bootstrap sample to the treatment effect in the subset of the “out of box” cases which have covariate vectors characteristic of $\mathcal{S}_2$. These bias estimates are averaged over the bootstrap samples. The method will fail, however, if the sample size is too small because there will be insufficient “out of box” cases to estimate the treatment effect in the $\mathcal{S}_2$ subset.

The method based on pre-validated classification of the cases remains effective with smaller sample sizes. This method evaluates treatment effect in the set of cases which were classified in $\mathcal{S}_2$ during the fold of the cross-validation in which they were left out. Under the null, those expected treatment effects should all be zero.

In a prospective randomized clinical trial, we recommend that this approach be part of the primary analysis. The other part is the usual test of average treatment effect for the entire eligible population. The threshold significance levels for the overall test and the test of treatment effect in the adaptively defined intended use subset can be chosen to ensure that the overall type I error of the trial is limited to the desired 0.05. If the null hypothesis of no average treatment effect for the overall eligible population is rejected, one can still use the approach described above for identifying the subset of patients most likely to benefit from the test treatment. This can be clinically useful if the proportion with benefit is quite limited as it is in many clinical trials. The re-sampling procedure can also provide a de-biased estimate of the treatment effect the complement of the intended use subset.

Rather than use a binary classifier, one can use a three level classifier to identify patients most likely to benefit from the test treatment, those least likely and those intermediate. The pre-validated scores can be divided into three sets either based on the 33rd and 66th percentiles of the difference scores or on pre-specified constants representing clinical significance as shown here.

We have emphasized here valid evaluation of the predictive classifier, not advocating using one type of classifier or another as is more usual. Although predictive classifiers have not been nearly as extensively studied as prognostic classifiers, many approaches to predictive classification are possible. The prognostic methods literature can be utilized by training prognostic classifiers for the treatment and control groups and then combining them into a predictive classifier or predictive score. The prognostic models can be based on logistic regression, random forest, support vector machines, proportional hazards regression etc.

Although there are many subset identification methods in the literature, there are very few subset validation methods. Dixon and Simon (1991) described an empirical Bayesian method that can be used with proportional hazards or logistic modeling with a large number of binary covariates. Hierarchical priors are placed on the interaction effects. The posterior distributions of treatment effect for any subset defined by one or more covariates are easily computed. These distributions are shrunken towards zero thereby providing a type of internally validated subset analysis. The methods presented here, however, avoid the assumption of hierarchical prior distributions.

Two final points deserve emphasis. First, all aspects of the development should be described prospectively in the statistical analysis plan. Secondly, fully external validation of a "subset effect" is always valuable. Generally there is no valid internal evaluation of the treatment effect in adaptively defined subsets and the claims are based solely on the biased re-substitution estimates. With the paradigm proposed here, there will be much stronger evidence of the value of a predictive classifier based on the internal evaluation. This can guide investigators about whether a confirmatory study is warranted.

References

Bai X, Tsiatis AA, Lu W, Song R (2017) Optimal treatment regimes for survival endpoints using a locally-efficient doubly-robust estimator from a classification perspective. Lifetime Data Anal 23:585–604

Cai T, Tian L, Wong PH, Wei LJ (2011) Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics (Oxford, England) 12:270–282

Dixon DO, Simon R (1991) Bayesian subset analysis. Biometrics 47(3):871–881

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

Freidlin B, Simon R (2005) Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Cancer Res 11:7872–7878

Matsui S, Simon R, Qu P, Shaughnessy JD, Barlogie B, Crowley J (2012) Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine. Clin Cancer Res 18:6065–6073

Simon R (2015) Sensitivity, specificity, PPV, and NPV for predictive biomarkers. J Natl Cancer Inst 107:153–156

Simon R, Simon N (2019) Finding the intended use population for a new treatment. J Biopharm Stat 29(4):675–684

Zhang Z, Li M, Lin M, Soon G, Greene T, Shen C (2017) Subgroup selection in adaptive signature designs of confirmatory clinical trials. J R Stat Soc C 66:345–361