Chapter 25
A Framework of Statistical Methods for Identification of Subgroups with Differential Treatment Effects in Randomized Trials

Lei Shen, Ying Ding and Chakib Battioui

Abstract The problem of identifying subgroups of patients with differential treatment effects in randomized trials plays an important role in the effort to tailor therapies to patients who are most likely to get benefit from them. It has attracted active research effort in recent years, and a growing number of statistical methods have been developed. In this chapter, after first examining the major challenges with subgroup identification that these methods are designed to address, we create a structured framework into which many of the methods can be placed. Such a framework provides insight into the subgroup identification problem and methods, and can be utilized to generate additional methods from existing ones. Using a small simulation study, we also demonstrate a recently proposed approach to systematically evaluate the performance of subgroup identification methods. Together, the methodological framework and systematic assessment of performance can help to determine the optimal analyses for various applications.

25.1 Introduction

In the drug development process, there is now an increasing amount of attention on tailoring a new therapy to those patients who are most likely to benefit from it. An important part of the effort to develop tailored therapeutics is the identification, using data from randomized clinical trials, of patient subgroups that enjoy an enhanced treatment response.

A number of statistical methods for the identification of such subgroups have been proposed (Loh 2002; Su et al. 2008; Lipkovich et al. 2011; Dusseldrop and Van

L. Shen (✉) · C. Battioui
Eli Lilly and Company, Indianapolis, USA
e-mail: Shen_Lei@Lilly.com

C. Battioui
e-mail: Battioui_Chakib@Lilly.com

Y. Ding
Department of Biostatistics, University of Pittsburgh, Pittsburgh, USA
e-mail: YingDing@Pitt.edu

Z. Chen et al. (eds.), *Applied Statistics in Biomedicine and Clinical Trials Design*,
ICSA Book Series in Statistics, DOI 10.1007/978-3-319-12694-4_25

Mechelen 2014; Foster et al. 2011; Battioui et al. 2014; Bell et al. 2012), and new ones regularly appear in the literature. With methods becoming more numerous, there comes an acute need to understand these methods and their performance in various settings. Although publications that present novel methods often contain simulation studies, the many differences in the setup of these simulation studies make it difficult to understand the relative performance of various methods. There is a strong interest in consistent and rigorous evaluation of subgroup identification methods, a topic addressed in Zink et al. (2015). In this chapter, we focus on a different question: Can we create a framework into which most of these methods would fit? Such a framework could help us gain much insight into the subgroup identification problem itself and its desirable solutions. On the surface, many of the statistical methods for subgroup identification look quite different from each other. However, in this chapter, we attempt to show that a useful framework can indeed be used to capture the key components of these methods. We will then demonstrate some important benefits of this framework and new insights gleaned from it.

In Sect. 25.2, we briefly review some of the subgroup identification methods in preparation for the discussion that follows. A methodological framework is proposed in Sect. 25.3, and we show a few important applications of this framework in Sect. 25.4 before concluding with Sect. 25.5.

25.2 Subgroup Identification Problem and Methods

25.2.1 Major Challenges

We first discuss the major challenges inherent in the problem of subgroup identification, which the various methods attempt to address in different ways. Perhaps the most often mentioned of these challenges is multiplicity, appropriately so, given the potentially severe impact it has on inflated type I error rate as well as on overly optimistic estimates of treatment effect. An analysis to identify interesting subgroups almost always includes multiple predictors—numbering in dozens for baseline pathophysiological variables and sometimes thousands for genomic or genetic variables. The number of predictors in an analysis is, however, not the only source of multiplicity, as there are at least two others. If a predictor is measured on a continuous scale—such as expression level of a gene or the amount of a protein—the same predictor can define many different patient subgroups when various cutoff values are used. In addition, when an analysis attempts to explore beyond subgroups defined by a single biomarker, the number of potential subgroups defined by the same set of predictors increases exponentially with higher complexity of the subgroups under consideration. For example, 100 binary biomarkers define 200 single-marker subgroups, but about 20,000 subgroups when two biomarkers are used jointly.

Another major challenge, also derived from the potentially large number of candidate subgroups, is computational. Not only do we need to efficiently search through a large number of subgroups in order to identify the most promising ones but we also

often need to apply resampling approaches to address the aforementioned multiplicity issue. Any of the various resampling techniques require an additional computational loop around the search for subgroups. When these two factors—searching among many subgroups and repeating the search for a larger number of resampled datasets—are put together, the computational burden can be so severe as to render an otherwise reasonable method infeasible.

High degree of multiplicity is not unique to the problem of subgroup identification; rather it is prevalent in fields such as "high dimensional data analyses" or "statistical learning," of which subgroup identification can be considered a special case. It is therefore natural to assume that many of the modern statistical techniques developed for these fields can be applied to subgroup identification, and indeed many of them prove to be useful. But now we come to a third major challenge of subgroup identification. If we analyze data from patients receiving the same treatment and try to identify subgroups with higher average response, we can directly utilize methods such as classification-and-regression trees (CART) (Breiman et al. 1984). However, since we are interested in subgroups with differential *treatment effects,* with "treatment effect" defined as the difference in average responses between two treatment groups (typically a new treatment and a control, the latter often in the form of placebo or standard-of-care), the problem is one of identifying treatment-by-subgroup interactions. Many statistical learning algorithms such as CART cannot be directly applied to solve this more complex problem.

It should be noted that, while these challenges are the most important ones, there are certainly others. For example, the naïve estimate of treatment effect in the identified subgroup is known to be overly optimistic due to ascertainment bias associated with the process of searching for the best subgroups. It is therefore desirable if a subgroup identification method can provide bias-corrected estimates of treatment effect so that the clinical importance of an identified subgroup can be properly judged.

25.2.2 Subgroup Identification Methods

Having discussed three major challenges in subgroup identification, we now provide a brief survey of three methods that have been proposed for this problem.

In what is traditionally termed as "subgroup analysis," most phase 2 and phase 3 clinical trials have in their statistical analysis plans lists of prespecified subgroups to be investigated using interaction testing. In this chapter, we will refer to this method as the "traditional" method. The testing for treatment by subgroup interaction is performed one-at-a-time. Often, no formal multiplicity adjustment is made, although the Bonferroni correction is sometimes used (if informally) in the interpretation of results.

Recursive partitioning techniques are utilized by many modern statistical methods for subgroup identification, including the next two methods to be reviewed in this chapter (as well as Loh 2002; Su et al. 2008; Lipkovich et al. 2011; Dusseldrop and Van Mechelen 2014; Bell et al. 2012). A detailed review of recursive partitioning

can be found in the references (Breiman et al. 1984; Loh 2014). Briefly, a recursive partitioning method creates a decision tree that classifies patients into subgroups with differential treatment effects using sequential splits based on dichotomous (or dichotomized) predictors.

The second method we consider in detail is the "virtual twin" method by Foster et al. (2011). It borrows concepts from counterfactual models for causal inference. As a first step, this method applies random forest model (Breiman 2001) to impute the unobserved outcome for each patient; that is, the outcome of the patient if he or she had been randomized to the other treatment group. This allows an individualized treatment effect to be calculated for each patient since his or her responses to both treatments are now available, for example, by subtracting one treatment response from the other if the response variable is continuous. Recursive partitioning is then applied to these individualized treatment effects in order to identify subgroups with enhanced treatment effect. The authors considered a number of techniques to account for multiplicity.

The final method to be discussed here is the "treatment-specified subgroup detection tool" (TSDT) method by Battioui et al. (2014). It also utilizes recursive partitioning to identify promising subgroups, albeit in two steps. First, one of the treatment groups is selected based on practical considerations; this is often—although not always—the group receiving the new treatment, since a hypothesized subgroup effect is such that response to the new treatment is impacted much more by the group status than is response to placebo or standard-of-care. Recursive partitioning is applied to this selected treatment group to yield a list of candidate subgroups that manifest differential response (note, not differential *treatment effect,* at this point). As the second step, data from the other treatment group are utilized to ensure that a given candidate subgroup does not reflect a similar differential response in the other treatment group, which would render the subgroup uninteresting since there would be little or no differential treatment effect. This two-step analysis is performed on a number of datasets resampled from the original dataset using bootstrap or subsampling. And finally, response values are permuted within each treatment arm to allow the calculation of an adjusted p value for the best subgroup identified.

While other methods for subgroup identification have been proposed (e.g., Loh 2002; Su et al. 2008; Lipkovich et al. 2011; Dusseldrop and Van Mechelen 2014; Bell et al. 2012), the above review of three representative methods is sufficient for the introduction of a general methodological framework in the next section.

25.3 A Framework for Subgroup Identification Methods

Although the three methods reviewed above have many differences among them, a number of important components emerge when we examine how they handle the major challenges of subgroup identification presented in the previous section. We will discuss each of these components below.

25.3.1 Component "T": How to Handle Treatment-By-Subgroup Interaction

The "traditional" method deals with this directly by testing for interactions. An interesting idea used by the "virtual twin" method is to first impute unobserved outcomes, hence changing a problem of differential treatment effect (interaction) to a simpler problem of differential response (main effect). Yet another strategy is used by the "TSDT" method, where one treatment group is analyzed first, before the other group is incorporated into the analysis to ensure an interaction effect.

By reviewing these and other methods, we can see at least the following approaches:

1. "Model": Testing for treatment-by-subgroup interaction in a regression model.
2. "Transformation": Transforming the observed response, such as imputing for unobserved outcome and then calculating individualized treatment effect (Foster et al. 2011).
3. "Sequential": Analyzing one treatment group first, before incorporating the other group (Battioui et al. 2014).
4. "Direct": Directly contrasting the observed average responses to two treatments for any given subgroup (Lipkovich et al. 2011).

25.3.2 Component "S": How to Search for Candidate Subgroups, Ideally in a Computationally Efficient Manner

In this regard, the "traditional" method simply considers all possible subgroups, but in doing so, essentially limits itself to considering only single-marker subgroups, since testing treatment-by-subgroup interactions for more complex subgroups is often computationally prohibitive in practice. The other two reviewed methods both utilize recursive partitioning, which counts computational efficiency as one of its main strength. Although not all possible subgroups are considered, the recursive nature of the algorithm allows much more complex subgroups to be considered, such as those defined by two or even more predictors.

We therefore have the following options for this component:

1. "Exhaustive": Studying all possible subgroups.
2. "Recursive partitioning": Creating a decision tree that classifies patients into subgroups with differential treatment effects using sequential splits based on dichotomous (or dichotomized) predictors (Loh 2014).
3. "Stepwise modeling": We use this option to represent the various penalized regression techniques (Zou and Zhang 2012), which can also identify candidate subgroups efficiently without considering all possible subgroups, but (unlike option #2) does so in a regression setting.

25.3.3 Component "M": How to Address Multiplicity

The Bonferroni correction sometimes used in traditional subgroup analysis can be impractical and overly conservative, and most subgroup identification methods utilize one or a combination of resampling techniques. We have the following options regarding this component:

1. "Simple": Such as the Bonferroni correction.
2. "Permutation": Using permutations of the original data to generate a reference distribution of the test statistic under an appropriate null.
3. "Bootstrap": Bootstrapping the original data to estimate the sampling distribution of the test statistic and/or a bias-corrected estimates of effect sizes using out-of-bag samples.
4. "Cross-validation": Using m-fold cross-validation to estimate prediction accuracy or other key quantities associated with a particular application.
5. "Subsampling": Randomly dividing the original data into two smaller datasets with prespecified proportions, with one used as training data and the other testing data; this is often repeated a number of times with results then averaged over subsamples.
6. "Combinations": Using a combination of above approaches, such as "subsampling & permutation."

It should be noted that there are other options for each of the components above, as the lists are not intended to be comprehensive. For example, some methods utilize variable importance to further control false-positive findings. One could also say that some of the options are fairly broad. For example, "recursive partitioning" covers a wide range of actual methods, with one of the key differences being the criteria used to determine whether and how to split at each node. In this regard, the "TSDT" method uses a specific approach, while the method by Bell et al. (2012) allows any user-defined criteria to be used. In theory, the user-defined criteria can optimize the desirability of the identified subgroup according to practical considerations for the specific application, such as the proper balance between subgroup size and the magnitude of treatment effect in the subgroup. Nevertheless, we will see in the next section that such a framework, even with simplifications on the options for each component, can be quite useful.

25.4 Utilizing the Framework

An immediate application of this framework is that we can now catalogue seemingly different methods for subgroup identification. For example, the "TSDT" method can be represented by *T(sequential)* × *S(recursive partitioning)* × *M(subsampling & permutation)*. As another example, the method by Lipkovich et al. (2011) can be represented by the following entry in the framework: *T(direct)* × *S(recursive partitioning)* × *M(permutation)*. Of course, it should be stated that such representation

captures the key elements of each method, but not all its details. The "TSDT" method utilizes out-of-bag samples from bootstrapping or subsampling to correct ascertainment bias in estimating the treatment effect size in the identified subgroup, and such details are not easily captured in a framework.

By considering the key components of subgroup identification methods, we are able to enumerate multiple options for each component, hence gaining valuable insight. By dissecting even a small number of methods, we now have a "toolbox" where options for each component can be combined. This leads to an even more interesting application, namely many "new" methods for subgroup identification generated by this toolbox. For example, one can naturally combine *T(transformation)* with *S(exhaustive)*. In other words, we can perform the first step of the "virtual twin" method and calculate individualized treatment effects, then perform a test for each subgroup that is simpler than interaction tests. Intuitively, in situations where the imputed outcome is of high quality, this method should outperform the "traditional" method. With the options given above for each component, we have $4 \times 3 \times 6 = 72$ combinations, each of which corresponds to a unique "method." Some of these methods, once described, are clearly impractical or inferior; but at the same time, many of these methods appear reasonable, yet are "novel" in the sense that they have not been proposed in the literature.

25.4.1 Systematic Method Evaluation

In addition to the value described above, we posit that such a framework of numerous methods for subgroup identification should work very well with a system to consistently and rigorously evaluate these methods, as proposed by Zink et al. (2015). There are three components in this evaluation system: data generation, application of analysis methods, and performance measurement. Consistency in data generation and performance measurement allows a wide array of analysis methods to be compared directly, thus leading to insight on strengths and weaknesses of each method.

Of both technical and practical importance is the proposal to evaluate the performance of a method on three levels: marker-level, subgroup-level, and subject-level. Briefly, the marker-level performance measures capture the accuracy in which the markers are correctly identified as predictive markers (or not); the subgroup-level performance measures include the average size and treatment effect of the identified subgroups, while the quality of associated treatment decisions for individual patients is measured at the subject level. Section 25.4.2 will elaborate on these measures in the context of a simulation study; additional details can be found in Zink et al. (2015).

Table 25.1 Three subgroup identification methods compared in the simulation study

Method	Component "T"	Component "S"	Component "M"
"Traditional"	Model	Exhaustive	Simple (Sidak correction)
"VT"	Transformation	Recursive partitioning	Permutation
"TSDT"	Sequential	Recursive partitioning	Subsampling + permutation

25.4.2 Simulation Study

Here, as an example to demonstrate how this system works, we present a small simulation study to compare three subgroup identification methods.

25.4.2.1 Subgroup Identification Methods

The methods have been briefly described in Sect. 25.2 and presented in Table 25.1 according to the framework we established. Here, we provide additional details of each method:

- Traditional Method: Test for treatment by subgroup interaction ("T: model") one-at-a-time for all variables ("S: exhaustive"), with multiplicity adjustment made using Sidak correction ("M: simple").
- Virtual Twin Method: First apply random forest model to impute for each patient the unobserved outcome as if he or she had been randomized to the other treatment group ("T: transformation"). Then apply recursive partitioning ("S: recursive partitioning") to the individualized treatment effects calculated by subtracting the "control outcome" from the "new treatment outcome" of the same patient. Finally, use permutations of the original data to estimate a reference null distribution of the test statistics for differential treatment effect in an identified subgroup, which in turn provides a multiplicity adjusted p value ("M: permutation").
- TSDT Method: In a subsample of the original data, construct candidate subgroups with differential response based solely on the new treatment arm, and then incorporate data from the control arm to exclude any candidate subgroup that does not show sufficient treatment-by-subgroup interaction ("T: sequential"). Candidate subgroups are constructed using recursive partitioning ("S: recursive partitioning"). Confirm the directional consistency of any remaining candidate subgroup in the corresponding out-of-bag sample. Averaging the results over all the random subsamples, for each candidate subgroup, and calculate the proportion of subsamples for which the subgroup is identified and shown to be consistent in the out-of-bag sample. Finally, apply permutation of the original data to obtain a reference null distribution of the consistency measure, which in turn provides a multiplicity adjusted p value ("M: subsampling + permutation").

For each of the subgroup identification methods, three different α levels ($\alpha = 0.1$, 0.2, 0.3) are used for controlling type I error rate.

Table 25.2 Five scenarios used in the simulation study

Scenario	Number of subjects	Number of markers	Number of predictive markers
A	240	20	1
B	240	50	2
C	240	50	1
D	240	20	0
E	240	50	0

25.4.2.2 Simulation Scenarios

We generated 200 datasets in each of five scenarios, with Table 25.2 providing a summary of these scenarios. In scenario A, each dataset contains 20 predictors, with one of them being a predictive marker and hence the target of identification. The number of predictors is increased to 50 in scenario B, with two of them being predictive markers. Scenario C is chosen to provide comparisons with the first two scenarios. It calls for generation of datasets with 50 predictors, one of which is a predictive marker. Contrasting scenarios A and C will allow us to observe the impact of the total number of predictors, while the comparison between scenarios B and C can demonstrate the impact of the number of predictive markers. Scenarios D and E are null scenarios with no predictive marker, included here for the purpose of evaluating control of type I errors.

In all scenarios, there are 240 subjects, with a 3:1 randomization ratio between the new treatment and control. For each dataset, an appropriate number (20 or 50) of genetic markers with identical distribution were generated. Specifically, each marker is a three-level ordinal variable with proportions of the three levels being 49, 42, and 9 %. According to the scenario, responses on a continuous scale were then generated with either zero, one, or two of the genetic markers being predictive. The predictive markers each confer the same magnitude of effect. When there is one predictive marker (scenario A and C), the population consists of two subpopulations that are both about 50 % in size and have average treatment effects 0.1 and 0.55, respectively. When there are two predictive markers (scenario B), the population is divided into four subpopulations that are each about 25 % in size and have average treatment effects 0.1, 0.55, 0.55, and 1.00, respectively.

25.4.2.3 Performance Measures

The aforementioned performance measures were calculated for each method across datasets. Specifically:

- Marker-level performance measures: Natural choices for presenting the accuracy with which predictive markers are correctly identified by an analysis method are sensitivity, specificity, positive predictive value, and negative predictive value.

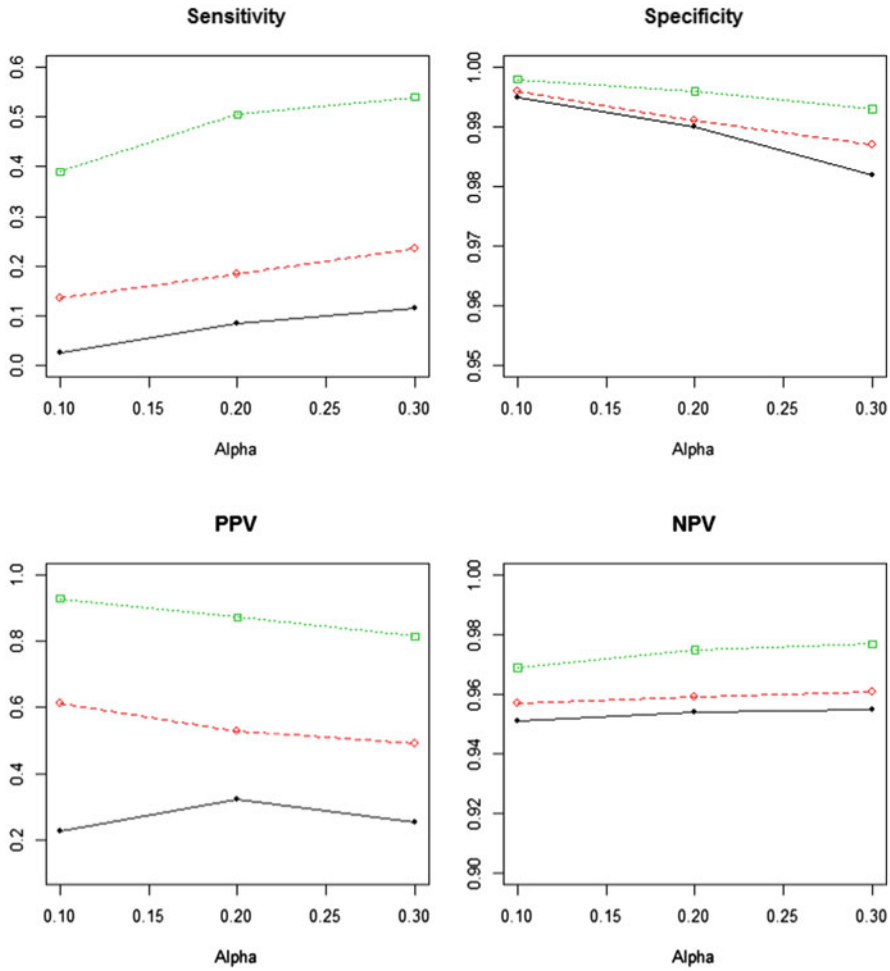

Fig. 25.1 *Marker level performance for scenario A* (solid line/solid dots = "Traditional," dashed line/hollow dots = "Virtual Twin," dotted line/square dots = "TSDT"). **a** *Sensitivity* = proportion of times that true predictive marker(s) are identified as predictive. **b** *Specificity* = proportion of times that nonpredictive markers are identified as nonpredictive. **c** *PPV* = proportion of true predictive markers among the markers identified as predictive. **d** *NPV* = proportion of nonpredictive markers among the markers identified as nonpredictive

These values for a single analysis are easily calculated from the 2×2 table with rows being the true status of a marker (predictive or not) and columns being the results of identification (identified as predictive or not). The proportions are then averaged across datasets.

- Subgroup-level performance measures: Toward the eventual objective of subgroup identification—to tailor a potential medicine to those patients who are more likely to respond—it is often desired that subsequent clinical trials would focus on the

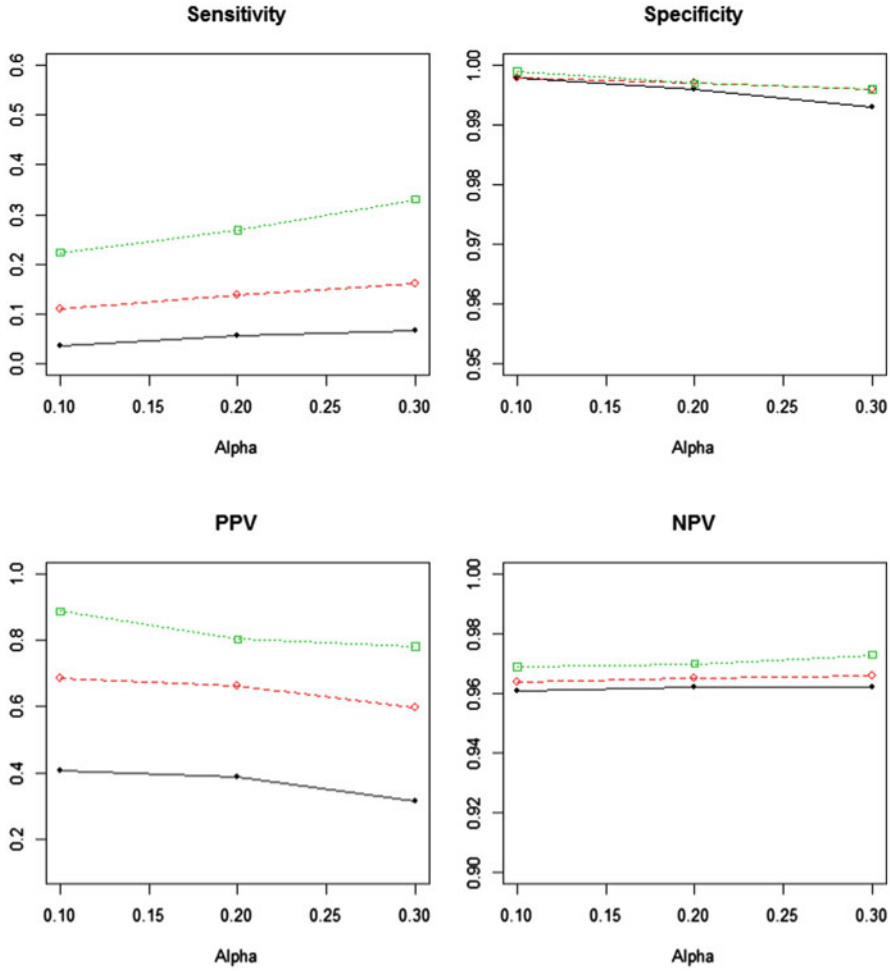

Fig. 25.2 *Marker level performance for scenario B* (solid line/solid dots = "Traditional," dashed line/hollow dots = "Virtual Twin," dotted line/square dots = "TSDT"). **a** *Sensitivity* = proportion of times that true predictive marker(s) are identified as predictive. **b** *Specificity* = proportion of times that nonpredictive markers are identified as nonpredictive. **c** *PPV* = proportion of true predictive markers among the markers identified as predictive. **d** *NPV* = proportion of nonpredictive markers among the markers identified as nonpredictive

subgroup that has been identified. Whether such a tailored drug development program is clinically and commercially prudent depends critically on the size and treatment effect associated with the subgroup. Therefore, it is important to capture these quantities (averaged over datasets) in simulation studies. While other summaries across simulated datasets can be constructed, we start with the most obvious ones by simply averaging the size and treatment effect of the identified subgroup for each dataset.

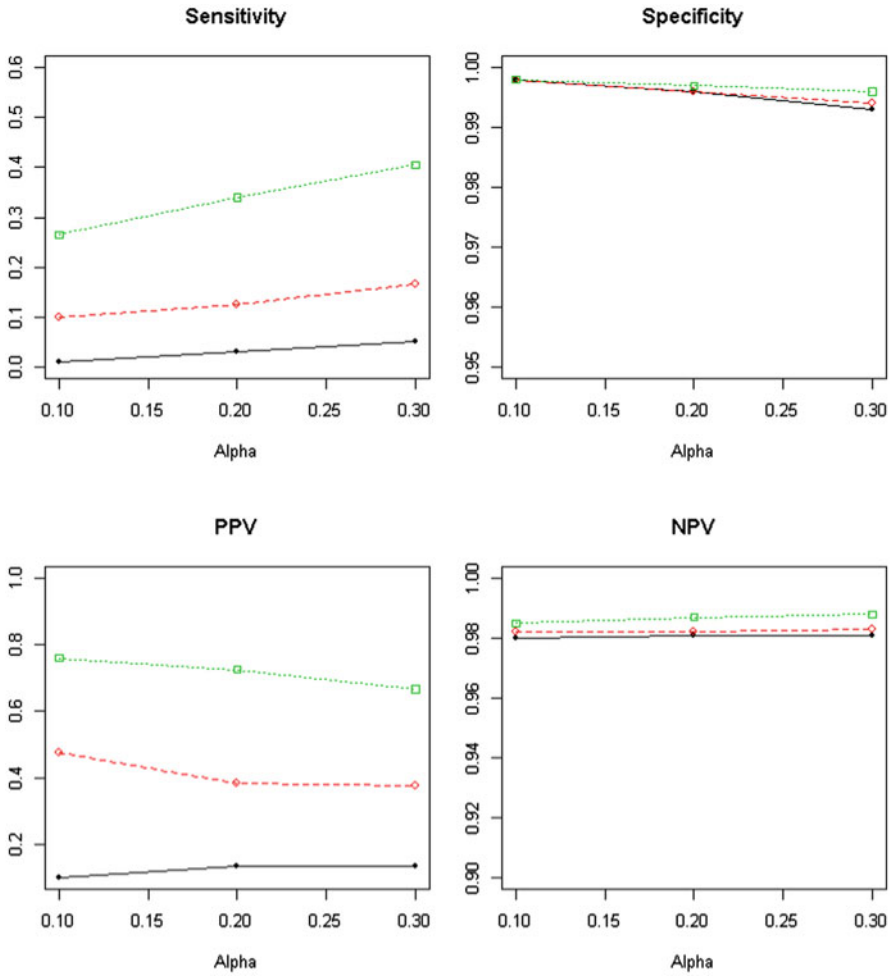

Fig. 25.3 *Marker-level performance for scenario C* (solid line/solid dots = Traditional," dashed line/hollow dots = "Virtual Twin," dotted line/square dots = "TSDT"). **a** *Sensitivity* = proportion of times that true predictive marker(s) are identified as predictive. **b** *Specificity* = proportion of times that nonpredictive markers are identified as nonpredictive. **c** *PPV* = proportion of true predictive markers among the markers identified as predictive. **d** *NPV* = proportion of nonpredictive markers among the markers identified as nonpredictive

- Subject-level performance measures: Upon approval of a potential treatment by regulatory agencies, the subgroup identified and confirmed in the drug development program will impact clinical decision making. The status of each patient—in terms of whether he or she belongs to the subgroup—can be considered as a decision rule of whether the patient should be given the new treatment. Naturally, the quality of this decision rule can be measured using sensitivity, specificity, positive

Table 25.3 Subgroup-level performance ($\alpha = 0.1$)

Scenario	Method	Subgroup identified (%)	Subgroup size (%)	Subgroup treatment effect
A	T	11	93.1	0.335
	VT	22	88.8	0.359
	TSDT	42	79.2	0.415
B	T	16	92.5	0.574
	VT	32	83.6	0.609
	TSDT	49	75.8	0.658
C	T	10	95.2	0.332
	VT	21	89.7	0.352
	TSDT	34	83.0	0.389
D	T	9	93.7	–
	VT	10	94.8	–
	TSDT	10	94.9	–
E	T	9	95.7	–
	VT	12	94.3	–
	TSDT	12	94.0	–

predictive value, and negative predictive value—this time with each subject as a unit. However, since clinical decision making does not become important until the new medicine is successfully developed, our simulation study here will not focus on these measures.

25.4.2.4 Results

Figures 25.1, 25.2, 25.3 present the marker-level performance for each non-null scenario, method, and α level. Across all scenarios and all measures, we can see that the "TSDT" method performed the best, while the "traditional" method performed the worst. The choice of α level had a moderate impact on the results. When comparing between scenarios, we can see that when the number of predictors increased (scenario C vs. A), sensitivity decreased for all three methods, as expected. On the other hand, an interesting observation is that, when the number of predictive markers increased (scenario B vs. C), sensitivity did not seem to improve.

The first column ("Subgroup identified") of Table 25.3 provides further information on how often each method identified a subgroup in these scenarios. We start with the two null scenarios D and E, where all three methods appear to do a good job of controlling the type I error rate at the stated nominal α level of 0.1. When we look at the three non-null scenarios A, B, and C, we see that in every scenario, the

"TSDT" method identified subgroups most often, whereas the "traditional" method did so the least often.

The final two columns of Table 25.3 present the subgroup-level performance measures (for $\alpha = 0.1$). It is evident that when the "TSDT" method identified subgroups in non-null scenarios, the subgroups also tended to be of the best quality in terms of having the largest treatment effect ("Subgroup Treatment Effect" column). Comparing across scenarios, it is clear that identification of subgroup, especially high-quality subgroups, is the most difficult for scenario C and easiest for scenario B, as one would expect. The average size of subgroups identified by each method is closely related to the frequency of identifying subgroups (since the size is 100 % of the population when no subgroup is identified), and in this case it is not otherwise informative given the identical distribution of all the predictors.

In summary, since all three methods control type I error rate at the same level in the null scenarios, the performance in non-null scenarios indicates that the "TSDT" is the most powerful method among the three in this simulation study.

25.5 Conclusions

In this chapter, we established a framework for statistical methods to identify patient subgroups with differential treatment effects in randomized clinical trials. By focusing on three major challenges with subgroup identification, we submit that the methods can be viewed as combinations of three key components: how treatment by subgroup interaction is handled, how candidate subgroups are searched, and how multiplicity is accounted for. This framework allows us to dissect existing methods, identify the options they utilize for each component, and then combine these options in other ways to easily generate additional methods. Such a system to catalogue and index various methods also works well with the framework proposed by Zink et al. (2015) to consistently evaluate performance of subgroup identification methods.

Acknowledgment The authors would like to sincerely thank their colleagues Hollins Showalter and Brian Denton for assistance with computation.

References

Battioui C, Shen L, Ruberg SJ (2014) A resampling-based ensemble tree method to identify patient subgroups with enhanced treatment effect. Proceedings of 2014 Joint Statistical Meetings, pp 4013–4023

Bell M, Higgs R, Lipkovich I, Lu Y, Ruberg S (2012) Flexible subgroup search tool. Presented at 2012 FDA/DIA Statistics Forum

Breiman L (2001) Random forests. Mach Learn 45:5–32

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont

Dusseldorp E, Van Mechelen I (2014) Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions. Stat Med 33:219–217

Foster JC, Taylor JMG, Ruberg SJ (2011) Subgroup identification from randomized clinical trial data. Stat Med 30:2867–2880

Lipkovich I, Dmitrienko A, Denne J, Enas G (2011) Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. Stat Med 30:2601–2621

Loh WY (2002) Regression trees with unbiased variable selection and interaction detection. Stat Sin 12:361–386

Loh WY (2014) Fifty years of classification and regression trees. International Statistics Review 82(3):329–348. doi:10.1111/insr.12016

Su X, Zhou T, Yan X, Fan J, Yang S (2008) Interaction trees with censored survival data. Int J Biostatist 4(1):1557–4679. doi:10.2202/1557–4679.1071

Zink RC, Shen L, Wiolfinge RD, Showalter HD (2015) Assessment of methods to identify patient subgroups with enhanced treatment response in randomized clinical trials. Applied Statistics in Biomedicine and Clinical Trials Design: Selected Papers from 2013 ICSA/ISBS Joint Statistical Meetings. Springer (in press)

Zou H, Zhang HH (2012) On the adaptive elastic-net with a diverging number of parameters. Ann Stat 37:1733–1751