Featured Article

Statistics in Medicine

Received 1 November 2010, Accepted 26 May 2011 Published online 4 August 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4322

Subgroup identification from randomized clinical trial data

Jared C. Foster,[a] Jeremy M.G. Taylor[a*†] and Stephen J. Ruberg[b]

We consider the problem of identifying a subgroup of patients who may have an enhanced treatment effect in a randomized clinical trial, and it is desirable that the subgroup be defined by a limited number of covariates. For this problem, the development of a standard, pre-determined strategy may help to avoid the well-known dangers of subgroup analysis. We present a method developed to find subgroups of enhanced treatment effect. This method, referred to as 'Virtual Twins', involves predicting response probabilities for treatment and control 'twins' for each subject. The difference in these probabilities is then used as the outcome in a classification or regression tree, which can potentially include any set of the covariates. We define a measure $Q(\hat{A})$ to be the difference between the treatment effect in estimated subgroup $\hat{A}$ and the marginal treatment effect. We present several methods developed to obtain an estimate of $Q(\hat{A})$, including estimation of $Q(\hat{A})$ using estimated probabilities in the original data, using estimated probabilities in newly simulated data, two cross-validation-based approaches, and a bootstrap-based bias-corrected approach. Results of a simulation study indicate that the Virtual Twins method noticeably outperforms logistic regression with forward selection when a true subgroup of enhanced treatment effect exists. Generally, large sample sizes or strong enhanced treatment effects are needed for subgroup estimation. As an illustration, we apply the proposed methods to data from a randomized clinical trial. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: randomized clinical trials; subgroups; random forests; regression trees; tailored therapeutics

1. Introduction

Confirmatory randomized clinical trials are designed to provide definitive information about treatments and frequently compare a standard treatment with a new treatment. The conclusions from such a study are applicable to the whole population that has been considered. However, with increasing use of targeted therapies and with increasing understanding of the mechanisms of action of new agents and of the human response to those agents, it is quite plausible that there are subgroups of patients for whom the new treatment is especially effective. Likewise, there could be subgroups of patients for whom the new treatment is not effective or less effective than the standard therapy. There is a strong desire to find such subgroups if they exist [1]. From a statistical perspective, searching for subgroups is known to be a dangerous exercise, with the high possibility of finding false positives. There is a large literature on this topic, with many statisticians and clinical trialists writing about the dangers of subgroup analyses [2–11]. The general opinion is that if subgroups are going to be examined they should be defined before looking at the data and that post hoc mining of the data in an uncontrolled or undefined fashion is likely to lead to unreliable results. On the other hand, there are those who believe that the biological rationale for subgroups is so strong that the statistical concerns about mining the data have been emphasized too much and are a barrier to progress [12].

An alternative strategy to predefining the subgroups is to predefine the statistical approach that is going to be used to find subgroups [1]. Such an approach is reproducible, and its statistical properties can be understood. The research we present in this paper is to describe several different strategies for finding subgroups. We believe the methodology described here would be particularly applicable for

[a] *Department of Biostatistics, University of Michigan , Ann Arbor, MI, 48109, USA*
[b] *Global Statistical Sciences, Advanced Analytics, Eli Lilly , Indianapolis, IN, USA*
**Correspondence to: Jeremy M.G. Taylor, Department of Biostatistics, University of Michigan , Ann Arbor, MI, 48109, USA.*
[†] *E-mail: jmgt@umich.edu*

two situations: (i) those where a new treatment is shown overall to be slightly better than the standard therapy but not sufficiently better to be widely adopted and (ii) those situations in which the new treatment appears better, although not significantly, but where there may be a subgroup of patients for which there is a substantial benefit of the new treatment. The challenge in such situations is finding such a subgroup and then demonstrating that the benefit is likely to be real for future patients.

The setting we consider is a two-group randomized clinical trial with a binary outcome variable Y, treatment group, indicator T, and covariates $\boldsymbol{X}$. The dimension of $\boldsymbol{X}$ is moderate, for example 8 to 100; these covariates are measured pretreatment and could be demographic, laboratory, or questionnaire variables. The goal is to find a subgroup of patients defined by a region (denoted A) of the covariate space of $\boldsymbol{X}$, in which the treatment effect is substantially better than the average treatment effect or better than some prespecified threshold. Because it is generally desirable to have a relatively simple way of defining the subgroup, we want A to depend on a small number of variables. For example, the region A could be $X_4 \geqslant 2$ and $X_7 \leqslant 5$.

The classical approach to identifying subgroups is nicely described by Kehl and Ulm [13] and involves the fitting of a model which includes interactions between treatment and the covariates. For example, for logistic regression, we could consider models of the general form

$$\text{logit}(P(Y = 1|T\boldsymbol{X})) = \alpha + \beta T + \gamma h(\boldsymbol{X}) + \theta T w(\boldsymbol{X}),$$

for which the main interest would be in the term $w(\boldsymbol{X})$. One problem with such an approach is that only those factors or combinations of factors that are included as interactions in the model may potentially be identified as important in defining the subgroup of enhanced treatment effect. Additionally, even if the form of interactions between factors is assumed to be linear, the order of such terms is unknown [13]. Thus, this method is not very feasible when the dimension of $\boldsymbol{X}$ is even moderately large because the number of potential interactions is massive even when one only considers one-way and two-way interactions of T with the Xs. It is well known that large sample sizes are needed to find interactions in models, so for similar reasons, we may expect large samples to be necessary to find and confirm the existence of subgroups and accurately define them. Although subgroups and interactions in statistical models are very related topics, they are not identical. What one means by a subgroup with enhanced treatment effect depends on how one defines enhanced. What one means by an interaction in a model depends on the scale of the observations because interactions on one scale can disappear when the data have been transformed to another scale.

Much of the statistical literature on model building and validation has some relevance to the problems of subgroup estimation and testing for subgroup effects. We will not review it here, except to mention that tests for interactions in clinical trials [14–16] appear promising and that tree-based methods of directly finding treatment–covariate interactions have been suggested [17–19].

In this paper, we develop and compare some different methods for defining a subgroup which shows an enhanced treatment effect. A challenge in this setting is to give an accurate estimate of the enhanced treatment effect in the subgroup. The procedures we develop essentially mine the data in a defined way, so there is a considerable danger of overfitting. We will investigate a number of different schemes for obtaining an honest estimate of the magnitude of the treatment effect in the subgroup, including resampling schemes such as cross-validation [20].

2. Methods for estimating region A

2.1. Notation

The data consist of $(Y_i, T_i, X_{1i}, \ldots, X_{pi})$, $i = 1, \ldots, n$. Let n_j = number of observations with $T_i = j$, $j = 0, 1$. We expect n_0 and n_1 to be close to $n/2$ in 1:1 randomized trial. We will focus on the situation where Y is a binary outcome. The Xs could be continuous or categorical, and they may be correlated with each other. There could be a marginal effect of the treatment on Y, and some of the Xs could be prognostic in the sense that they are marginally associated with Y. The goal is to partition the covariate space into two regions A and A^c, with A being defined by a relatively small subset of the Xs. Define $|A_j|$ to be the number of observations with $\boldsymbol{X}_i \in A$ and $T_i = j$, $j = 0, 1$.

2.2. Forward logistic regression

As a standard and simple method, we consider forward selection in a logistic regression model. The terms that are considered for inclusion in the model are main effects for all Xs and T, and all $X \times T$,

$X \times X$, and $X \times X \times T$ interactions. The forward selection procedure starts with an intercept-only model and, at each step, the term which gives the smallest Akaike information criterion is added to the model. The final model is used to calculate $\hat{P}_{1i} = P(Y_i = 1|T_i = 1, X_i)$ and $\hat{P}_{0i} = P(Y_i = 1|T_i = 0, X_i)$ for each person i. A new variable $Z_i = \hat{P}_{1i} - \hat{P}_{0i}$ is then created, and subjects are defined to be in group $\hat{A}$ if Z_i is greater than some cutoff c, which we generally take to be either $\delta + 0.1$ or $\delta + 0.05$, where δ is an estimate of the treatment effect $P(Y_i = 1|T_i = 1) - P(Y_i = 1|T_i = 0)$. In addition, although all Xs in the final model are needed to define $\hat{A}$, the Xs which are involved in first-order and second-order interactions with T are noted, as these are the Xs which are most important in defining $\hat{A}$. If no Z_i is greater than c, then $\hat{A}$ is the null set. An alternative approach to defining the treatment effect for each subject is to define the difference on the logit scale, i.e., define $Z_i = \text{logit}(\hat{P}_{1i}) - \text{logit}(\hat{P}_{0i})$, and subjects are defined to be in group $\hat{A}$ if Z_i is greater than some cutoff. The form of Z_i that is preferred will depend on the context. In this paper, we will only consider differences on the probability scale.

2.3. *Virtual Twins method*

This approach borrows concepts from counterfactual models, in which there are two possible outcomes for each person (one under each treatment assignment), only one of which can be observed, and it is the difference between the two outcomes that is important. We investigate two versions of Virtual Twins, VT(R) and VT(C), which have the same first step but in the second step have either a regression procedure or a classification procedure.

2.3.1. Step 1. Apply random forests to the data. A random forest [21] is an ensemble predictor based on multiple regression trees. For our purposes, the random forest is simply a black box predictor which takes as input covariate values ($\boldsymbol{X}_i, T_i, \boldsymbol{X}_i I(T_i = 0)$, and $\boldsymbol{X}_i I(T_i = 1)$ in our case) and gives as output an estimate of $P(Y_i = 1)$ for that set of covariate values. The inclusion of $\boldsymbol{X}_i I(T_i = 0)$ and $\boldsymbol{X}_i I(T_i = 1)$ as covariates is not essential, but in numerical work we found that their inclusion improved the properties of the method. Fitting of the random forest is done using the R function *randomForest* with all default settings except for the number of trees per forest, which we set at 1000. As with the logistic method, the random forest is used to predict $\hat{P}_{0i}$ and $\hat{P}_{1i}$. If the actual treatment group for subject i is j, then $\hat{P}_{ji}$ is obtained from the out-of-bag estimate from the random forest, whereas $\hat{P}_{(1-j)i}$ is obtained by applying the random forest to that person's data, with the treatment group switched. Once this is done, we define $Z_i = \hat{P}_{1i} - \hat{P}_{0i}$, which can be regarded as an estimate of the treatment effect for subject i. A variation of this step would be to use two separate random forests, one for each treatment group, and predict for each subject using the forest from the other group. All of these strategies for obtaining Z_i involve some form of extrapolation, that is, estimating the probability of an outcome in a treatment group that is different from the treatment they actually received.

2.3.2. Step 2. Estimate a regression or classification tree. The purpose of this tree is to find a small number of Xs that are strongly associated with Z and hence can define A. We consider two alternative methods.

Virtual Twins (regression). In this method, denoted VT(R), we estimate a regression tree with Z as the response variable and covariates $\boldsymbol{X}$. The regression tree is then used to predict values of Z_i for each person. Then, subjects with predicted Z_i greater than some threshold c are considered to be in $\hat{A}$. Thus, $\hat{A}$ is defined by the paths down the tree which lead to terminal nodes with predicted Zs greater than c. We take c to be either $\delta + 0.1$ or $\delta + 0.05$. If none of the predicted Z_is are greater than c, then $\hat{A}$ is empty. We use the R function *rpart* with default settings, except that the minimal terminal node size is 20, and the complexity parameter is taken to be 0.02.

Virtual Twins (classification). Define a new binary variable Z^*, as $Z_i^* = 1$ if $Z_i > c$ and $Z_i^* = 0$ if $Z_i \leqslant c$. The value of c for this method is generally the same as that used in VT(R) and will affect the size of $\hat{A}$. Because Z_i is a difference in probabilities, we are essentially creating a classification variable immediately after the random forest stage. This Z^* is used as the outcome in building the classification tree, which is then used to classify individuals as being in $\hat{A}$ or not; thus, all Xs in the tree define $\hat{A}$. This method is denoted VT(C). Note that $\hat{A}$ is empty if the classification tree has no splits.

3. Properties of the estimated region $\hat{A}$

For any region A in the covariate space, define

$$Q(A) = (P(Y=1|T=1, \boldsymbol{X} \in A) - P(Y=1|T=0, \boldsymbol{X} \in A)) - (P(Y=1|T=1) - P(Y=1|T=0)) \quad (1)$$

as the measure of the enhanced treatment effect in A compared with the average treatment effect. We define $Q(A)$ to be zero if A is a null set. Let $\hat{A}$ be the estimated region, as determined by one of the methods described above. Because $\hat{A}$ is the region that would be recommended for use, it is important to understand how effective $\hat{A}$ will be in defining a region of enhanced treatment effect for future populations. Clearly large values of $Q(\hat{A})$ are desirable if $\hat{A}$ is to be useful.

3.1. *Estimation of* $Q(\hat{A})$

Below we describe some approaches to obtaining an estimate $\hat{Q}(\hat{A})$ of $Q(\hat{A})$, which will be evaluated in a simulation study. We consider six methods of estimating $Q(\hat{A})$. It is desirable that $\hat{Q}(\hat{A})$ be as close as possible to $Q(\hat{A})$, rather than $Q(A)$, as $Q(\hat{A})$ is the true measure of enhanced treatment effect for the estimated region $\hat{A}$, whereas $Q(A)$ is the corresponding measure for the unknown true region A, which in general will not be the same as $\hat{A}$.

Method 1. Resubstitution method. Estimate the four quantities $P(Y=1|T=1, \boldsymbol{X} \in \hat{A})$, $P(Y=1|T=0, \boldsymbol{X} \in \hat{A})$, $P(Y=1|T=1)$, and $P(Y=1|T=0)$ from the observed proportions in the data, which are then substituted into equation (1) to give $\hat{Q}(\hat{A})$. For this estimator, the same data that were used to construct $\hat{A}$ will be used to evaluate it. The methods of estimating $\hat{A}$ may be overfitting the data (i.e., modeling the noise), so a resubstituted estimate of $Q(\hat{A})$ is likely to be biased, especially in small samples and with many covariates.

Method 2. Simulate new data (SND). This method is a type of parametric bootstrap approach. For both the logistic regression and Virtual Twins methods, the first step gives estimates of P_{1i} and P_{0i}, and these estimates can be used to simulate new outcome data from Bernoulli distributions. The new data $(Y_i^*, \boldsymbol{X}_i, T_i)$ will 'look like' the original data in terms of marginal and conditional distributions but will be statistically independent of the original data. Specifically if $T_i = 1$, then $Y_i^* \sim Be(\hat{P}_{1i})$, and if $T_i = 0$, then $Y_i^* \sim Be(\hat{P}_{0i})$. We then obtain the estimates of the four quantities $P(Y=1|T=1, \boldsymbol{X} \in \hat{A})$, $P(Y=1|T=0, \boldsymbol{X} \in \hat{A})$, $P(Y=1|T=1)$, and $P(Y=1|T=0)$ from the empirical proportions in the simulated data Y^*, from which we obtain $\hat{Q}(\hat{A})$. Alternatively, we can avoid actually simulating new data by simply taking an appropriate average of the estimates of P_{1i} and P_{0i} to get $\hat{Q}(\hat{A})$:

$$\hat{Q}(\hat{A}) = \left[\frac{1}{|\hat{A}_1|} \sum_{\boldsymbol{X}_i \in \hat{A}, T_i = 1} \hat{P}_{1i} - \frac{1}{|\hat{A}_0|} \sum_{\boldsymbol{X}_i \in \hat{A}, T_i = 0} \hat{P}_{0i} \right] - \left[\frac{1}{n_1} \sum_{T_i=1} \hat{P}_{1i} - \frac{1}{n_0} \sum_{T_i=0} \hat{P}_{0i} \right].$$

Compared with the resubstitution estimator, we expect this method to have less bias because it is not based explicitly on the original data. However, it may not completely eliminate the bias because any idiosyncracies in the observed data which are causing overfitting will still be present but to a lesser degree in the estimates $\hat{P}_{1i}$ and $\hat{P}_{0i}$.

Method 3. Cross-validation of $\hat{P}_{1i}$ and $\hat{P}_{0i}$. A modification of method 2 is to obtain $\hat{P}_{1i}$ and $\hat{P}_{0i}$ via cross-validation. In this method, the specific data for subject i are not used to obtain $\hat{P}_{1i}$ and $\hat{P}_{0i}$. Using 10-fold cross-validation, we apply the random forest or logistic regression approach to 9/10 of the data and use the resulting predictor to obtain estimates of P_{1i} and P_{0i} for the remaining 1/10 of the observations. This is repeated 10 times, then the simple averaging approach of method 2 is applied to give $\hat{Q}(\hat{A})$.

Method 4. Full cross-validation. For all the methods, we apply the whole process of defining $\hat{A}$ to 9/10 of the data, which gives a region $\hat{A}_k$, and 1/10 is left out, which is used as an independent testing data set. This is repeated 10 times. Each left-out observation is then either in $\hat{A}_k$ or not. We count the number

of observations with $Y = 1$ for $T = 1$ and $\boldsymbol{X} \in \hat{A}_k$ and similarly the number with $Y = 1$ for $T = 0$ and $\boldsymbol{X} \in \hat{A}_k$. We pool the counts across the 10 values of k to give final estimates of $P(Y = 1|T = 1, \boldsymbol{X} \in \hat{A})$ and $P(Y = 1|T = 0, \boldsymbol{X} \in \hat{A})$, which are then used in equation (1) to give $\hat{Q}(\hat{A})$. Note that in contrast to method 3 in which $\hat{A}$ is fixed, the estimate of $\hat{A}$ in method 4 does vary.

Methods 5 and 6. Bootstrap bias corrected. The bootstrap is a method that can be used to evaluate the bias in an estimator. The original estimator is then adjusted by this estimated bias. In this method, the original data will be bootstrapped (resampled with replacement) 20 times, and for each data set a new estimated region will be obtained. Then the estimate of $Q(\hat{A})$ is given by (Q from original data applied to original $\hat{A}$) + (Q from original data applied to new $\hat{A}$) − (Q from new data applied to new $\hat{A}$). The justification for this adjusted estimate requires some notation. Let F = true unknown distribution of the data, $\hat{F}$ = distribution of bootstrapped data, A = true region, $\hat{A}_F$ = region estimated from the observed data, and $\hat{A}_{\hat{F}}$ = region estimated from bootstrapped data (new $\hat{A}$). It is necessary to consider three probability laws depending on whether Q is applied to future data, the current data, or bootstrapped data. Specifically, $Q(.)$ is governed by the probability law on the next data set (i.e., the true F), $\hat{Q}_F(.)$ is governed by the probability law on the observed data (i.e., the empirical distribution), and $\hat{Q}_{\hat{F}}(.)$ is governed by the probability law from the bootstrapped data (i.e., the empirical bootstrap distribution).

The quantity of interest is $Q(\hat{A}_F)$, which can be written as $[Q(\hat{A}_F) - Q(A)] + Q(A) = R + S$. Using the bootstrap, we approximate R by $[\hat{Q}_F(\hat{A}_{\hat{F}}) - \hat{Q}_F(\hat{A}_F)]$. To obtain S, we approximate $[\hat{Q}_F(\hat{A}_F) - Q(A)]$ by $[\hat{Q}_{\hat{F}}(\hat{A}_{\hat{F}}) - \hat{Q}_F(\hat{A}_F)]$, rearranging this to give an approximation to S of $\hat{Q}_F(\hat{A}_F) - [\hat{Q}_{\hat{F}}(\hat{A}_{\hat{F}}) - \hat{Q}_F(\hat{A}_F)]$. Then adding approximations of R and S gives $Q(\hat{A}_F) = \hat{Q}_F(\hat{A}_F) + \hat{Q}_F(\hat{A}_{\hat{F}}) - \hat{Q}_{\hat{F}}(\hat{A}_{\hat{F}})$. Because the second term will tend to be smaller than the third term, the bias-corrected estimate will likely be less optimistic than the original estimate $\hat{Q}_F(\hat{A}_F)$. To implement this bootstrap-corrected estimate, we use 20 bootstrap samples, and for the second and third terms, the average of the 20 values of Q is used. This bias correction can be applied to any method we have for calculating Q. Method 5 consists of applying it to the resubstitution method, and method 6 consists of applying it to the SND method.

3.2. *Sampling variability of* $\hat{Q}(\hat{A})$

It may be desirable to attach standard errors to $\hat{Q}(\hat{A})$. We suggest the following scheme. Simulate new data sets using the estimates of P_{1i} and P_{0i} obtained from the random forest procedure. For new data set j, find the new $\hat{A}_j$, and calculate $\hat{Q}(\hat{A}_j)$ using one of the methods above. The standard deviation of these quantities is an estimate of the standard error.

4. Simulations

4.1. *Simulation study design*

We generate data from a logistic model of the general form $\text{logit}(P(Y = 1|T, \boldsymbol{X})) = \alpha + \beta T + \gamma h(\boldsymbol{X}) + \theta T I(\boldsymbol{X} \in A)$, where Xs are independent $N(0, 1)$, and A is a known region in the covariate space defined by two Xs. There are a number of aspects of the design that may be important to consider, including sample size, dimension of $\boldsymbol{X}$, number of Xs that determine A, size of A, values of the parameters, and correlation between the Xs, as these might influence the performance of the methods. We only consider some of these potential scenarios in this paper.

4.2. *Criteria for evaluation and comparison of region* A *estimation methods*

We use a number of criteria to evaluate each method's ability to identify the region A.

Finding correct X**s.** For VT(R) and VT(C), the total number of unique Xs in the tree is recorded. For the logit method, we record the total number of unique X-by-T interactions in the selected model. For the VT(R), VT(C), and logit methods, we record how often the methods find specific covariates, including some that are important and some that are not associated with the outcome. For the VT methods, we record whether each covariate is in the tree, and for the logit method, we record whether the one-way X-by-T interaction is in the final model. To determine whether or not the trees are finding the correct

Xs, we record whether or not the two Xs that define the true region A both occur in the first two levels of the tree.

Closeness of $\hat{A}$ to the true A. This is measured using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the ROC curve (AUC). Each estimation method gives an $\hat{A}$, and because we know the true A we can calculate sensitivity, specificity, PPV and NPV for each A estimation method and each value of the threshold c. Furthermore, the predicted differences in probabilities from the logit model and the VT(R) approaches can be used along with true A membership to calculate AUC by varying c. When $\theta \neq 0$ and $\hat{A}$ is empty, the PPV is defined as zero. When $\theta = 0$, A does not exist. In these cases, only specificity is meaningful and ideally should be close to 1.

Closeness of the size of $\hat{A}$ to the size of the true A. To evaluate the closeness of $|\hat{A}|$ to $|A|$, we record $|\hat{A}|$. If $\hat{A}$ does not exist, then $|\hat{A}|$ is taken as zero.

Power. Another quantity of interest is the percentage of times our methods find a null $\hat{A}$ when $\theta \neq 0$ and when $\theta = 0$. For VT(C), we consider the method to have found a null $\hat{A}$ if the final tree consists only of a single node, i.e., the method failed to find any Xs. For VT(R), $\hat{A}$ is null if either the final tree is a single node or if a tree exists but all predicted Z_is are less than c. For the logit-based method, we consider $\hat{A}$ to be null if all $Z_i < c$. Lastly, to quantify if there are enough data to determine whether individuals in true A have an enhanced treatment effect, we fit a logit model with the correct terms (i.e., only the terms used to generate the data) and calculate the power by determining how often the A-defining term ($\hat{\theta}$) is significantly different from zero.

4.3. Properties of $\hat{Q}(\hat{A})$ as an estimator of $Q(\hat{A})$

For each simulated data set, we calculate $\hat{Q}(\hat{A})$ for each A estimation procedure using the six methods as described above in addition to calculating $Q(A)$ and $Q(\hat{A})$. Because this is done for multiple data sets, we can estimate the variability of the various $Q(A)$ estimates and can quantify how far $\hat{Q}(\hat{A})$ is from $Q(A)$ and $Q(\hat{A})$. Thus, this allows us to estimate both the bias and the variability of $\hat{Q}(\hat{A})$ and of $\hat{Q}(\hat{A}) - Q(\hat{A})$ for each A estimation procedure. Because they do not lead to any different conclusions, we do not show the results for $\hat{Q}(\hat{A}) - Q(\hat{A})$.

5. Simulation results

We consider a base-case simulation design and several modifications of this base case. In the base case, we simulate randomized trials with 1000 patients, and the Xs are generated as independent $X_j \sim N(0,1)$, $j = 1, \ldots, 15$. We consider logit models for data generation $\text{logit}(P(Y=1)) = -1 + 0.5X_1 + 0.5X_2 - 0.5X_7 + 0.1T + 0.5X_2X_7 + \theta TI(\boldsymbol{X} \in A)$, where θ determines the extent to which individuals in region A have an enhanced treatment effect. For most models, the true region A is $X_1 > 0 \cap X_2 < 0$, which contains approximately 25% of the observations. Results are based on 100 simulated data sets. We consider a null case (true $\theta = 0$) and a base case (true $\theta = 0.9$) chosen so that in most scenarios the correct logit model has approximately 90% power in a test of $\theta = 0$. Simulation results for the base case are summarized in Section 5.1 and the upper portions of Tables I, II, and III, the null-case results are summarized in Section 5.2 and the lower portions of Tables I, II, and III, and results for other modifications are presented in Section 5.3.

5.1. Base case results

From Table I, we can see that the tree-based methods tend to find too many Xs, whereas the logit-based method tends to be too conservative, finding only about one X on average. Additionally, the tree-based methods appear to be better able to identify main effects and covariates determining region A than the logit-based method, although they also tend to identify non-important covariates (e.g., X_3 and X_4) as important slightly more often than the logit-based method. Although the tree-based methods can typically identify one of the covariates determining subregion A, they are less able to simultaneously identify both covariates that define region A as the two most important. Between the tree-based methods, VT(R) finds the correct Xs in the top two levels more frequently than VT(C). Whereas VT(R) never fails to find a tree and $\text{VT(C)}_{0.05}$ almost never fails, $\text{VT(C)}_{0.1}$ fails to find a tree 11% of the time. From Table II,

Table I. 'Finding A' performance, results for base case ($\theta = 0.90$, true $A = \{X_1 > 0, X_2 < 0\}$) and null case ($\theta = 0$, true A null).

	Logit	VT(R)	$\text{VT(C)}_{0.05}$	$\text{VT(C)}_{0.1}$
$\theta = 0.9$				
Mean (# unique Xs)	1.13	3.62	3.94	3.72
SD (# unique Xs)	0.97	1.04	1.40	1.75
Proportion found X_1 (int*)	0.30	0.96	0.92	0.82
Proportion found X_2 (int)	0.42	0.68	0.67	0.64
Proportion found X_7 (main)	0.11	0.72	0.64	0.62
Proportion found X_3 (absent)	0.02	0.14	0.12	0.17
Proportion found X_4 (absent)	0.02	0.11	0.15	0.15
Proportion $\{X_1, X_2\}$ in top 2	NA	0.51	0.39	0.32
Proportion found nothing	0.29	0.00	0.01	0.11
$\theta = 0$				
Mean (# unique Xs)	0.18	3.29	3.73	2.08
SD (# unique Xs)	0.48	1.00	1.83	1.89
Proportion found X_1 (main)	0.06	0.72	0.67	0.35
Proportion found X_2 (main)	0.03	0.68	0.51	0.32
Proportion found X_7 (main)	0.02	0.59	0.59	0.37
Proportion found X_3 (absent)	0.00	0.14	0.17	0.05
Proportion found X_4 (absent)	0.02	0.05	0.18	0.08
Proportion found nothing	0.85	0.00	0.07	0.38

Note: * indicates whether the variable is in the true data-generating model as only a main effect, an interaction with T, or absent. Subscripts on VT(C) indicate the constant added to average treatment effect for $\hat{A}$ definition.

Table II. Comparison of A and $\hat{A}$ for base case ($\theta = 0.90$, true $A = \{X_1 > 0, X_2 < 0\}$) and null case ($\theta = 0$, true A null).

	$\text{Logit}_{0.05}$	$\text{Logit}_{0.1}$	$\text{VT(R)}_{0.05}$	$\text{VT(R)}_{0.1}$	$\text{VT(C)}_{0.05}$	$\text{VT(C)}_{0.1}$
$\theta = 0.9$						
Size of $\hat{A}$						
Proportion $\hat{A}$ null	0.29	0.29	0.07	0.32	0.01	0.11
5th percentile	0	0	0	0	99	0
50th percentile	199	73	189	114	222	103
95th percentile	327	194	397	252	333	186
Sensitivity	0.34	0.16	0.47	0.28	0.49	0.28
Specificity	0.89	0.96	0.89	0.95	0.87	0.96
PPV	0.37	0.41	0.55	0.45	0.56	0.59
NPV	0.81	0.78	0.84	0.81	0.84	0.80
AUC	0.69	0.69	0.77	0.77	–	–
$\theta = 0$						
Size of $\hat{A}$						
Proportion $\hat{A}$ null	0.84	0.85	0.25	0.65	0.07	0.38
5th percentile	0	0	0	0	0	0
50th percentile	0	0	131	0	162	40
95th percentile	188	55	308	142	274	118
Specificity	0.97	0.99	0.87	0.97	0.85	0.96

we can see that all tree-based methods except for $\text{VT(R)}_{0.1}$ find null $\hat{A}$s less frequently than the corresponding logit-based methods. Of the tree-based methods, VT(C) performs the best in this regard, and defining $\hat{A}$ by $Z_i > \delta + 0.05$ noticeably outperforms $Z_i > \delta + 0.1$. With one exception, the median predicted subgroup size tends to be closer to 25% (i.e., 250) for the tree-based methods than for the corresponding logit-based method. Also, although possibly lower than desirable, the sensitivities and positive predictive values for the tree-based methods are noticeably better than those of the logit-based

Table III. Q estimates for base case ($\theta = 0.90$, true $A = \{X_1 > 0, X_2 < 0\}$) and null case ($\theta = 0$, true A null).

	$\text{Logit}_{0.05}$	$\text{Logit}_{0.1}$	$\text{VT(R)}_{0.05}$	$\text{VT(R)}_{0.1}$	$\text{VT(C)}_{0.05}$	$\text{VT(C)}_{0.1}$
$\theta = 0.9$						
$Q(A)$	0.139	0.139	0.139	0.139	0.139	0.139
$Q(\hat{A})$	0.043	0.054	0.070	0.062	0.071	0.083
$\text{SD}(Q(\hat{A}))$	0.042	0.056	0.047	0.057	0.043	0.055
$\hat{Q}(\hat{A})$						
RS(Mean)	0.080	0.125	0.164	0.155	0.164	0.192
(SD)	0.063	0.127	0.078	0.122	0.062	0.106
SND(Mean)	0.073	0.104	0.111	0.101	0.115	0.144
(SD)	0.053	0.074	0.046	0.074	0.033	0.058
CV(Mean)	0.057	0.081	0.106	0.095	0.104	0.106
(SD)	0.045	0.063	0.045	0.070	0.041	0.079
FCV(Mean)	0.023	0.005	0.037	0.028	0.052	0.027
(SD)	0.088	0.163	0.110	0.164	0.091	0.182
RS(BC)(Mean)	0.060	0.096	0.103	0.118	0.105	0.133
(SD)	0.060	0.132	0.076	0.105	0.068	0.104
SND(BC)(Mean)	0.045	0.068	0.072	0.076	0.067	0.092
(SD)	0.043	0.060	0.044	0.061	0.041	0.049
$\theta = 0$						
$Q(\hat{A})$	0.000	−0.001	0.010	0.009	0.008	0.012
$\text{SD}(Q(\hat{A}))$	0.006	0.016	0.023	0.023	0.025	0.024
$\hat{Q}(\hat{A})$						
RS(Mean)	0.012	0.010	0.124	0.084	0.109	0.114
(SD)	0.034	0.094	0.099	0.127	0.092	0.135
SND(Mean)	0.014	0.019	0.074	0.051	0.089	0.090
(SD)	0.033	0.049	0.054	0.072	0.040	0.074
CV(Mean)	0.009	0.013	0.069	0.047	0.069	0.046
(SD)	0.024	0.037	0.052	0.068	0.048	0.071
FCV(Mean)	−0.040	−0.024	−0.020	−0.076	−0.014	−0.079
(SD)	0.149	0.188	0.132	0.232	0.105	0.219
RS(BC)(Mean)	0.004	0.002	0.068	0.062	0.043	0.068
(SD)	0.033	0.096	0.085	0.103	0.097	0.119
SND(BC)(Mean)	0.007	0.012	0.036	0.036	0.033	0.048
(SD)	0.019	0.031	0.045	0.054	0.041	0.047

Note: RS, SND, CV, and FCV indicate the resubstitution, simulate new data, cross-validation-based simulate new data, and cross-validation-based resubstitution methods, respectively. BC indicates bootstrap bias-corrected method.

methods, whereas the specificities and negative predictive values for all methods are similar. The AUC is better for the tree-based methods than for the logit-based methods.

From Table III, it is clear, because $Q(A) > Q(\hat{A})$, that all methods tend to identify subgroups that have less of a true enhanced treatment effect than the true subgroup A. The tree-based methods outperform the logit-based method, finding subgroups which have larger values of $Q(\hat{A})$. Also, VT(C) slightly outperforms VT(R), particularly when $Z_i^* > \delta + 0.1$ is used to define $\hat{A}$. With the exception of the full cross-validation-based estimates, all estimates (and standard deviations) are essentially identical between VT(R) and VT(C) when the threshold for defining $\hat{A}$ is $\delta + 0.05$; however, when $\hat{A}$ is instead defined using threshold $\delta + 0.1$, the cross-validation-based and bias-corrected estimates show similar levels of bias between VT(R) and VT(C), but the resubstitution and SND estimates are slightly less biased for VT(R). For all methods, the resubstitution, SND, and cross-validation-based SND estimates tend to overestimate $Q(\hat{A})$, although the two SND methods are generally much closer to $Q(\hat{A})$ than the corresponding resubstitution estimates. The full cross-validation-based resubstitution estimates tend to greatly underestimate $Q(\hat{A})$ and are extremely variable. The two bootstrap bias-corrected estimates are the best, with the bias-corrected SND estimate being the closest to $Q(\hat{A})$ and having smaller variability than other estimates.

From the magnitude of the standard deviations in Table III for SND with bootstrap bias correction SND(BC), we can see that the value of $\hat{Q}(\hat{A})$ would be considered greater than zero much more often

for the $\theta = 0.9$ case than for the null case. The standard errors of $\hat{Q}(\hat{A})$ were also evaluated using the method outlined in Section 3.2. We found that for VT(R) and the SND method of estimating $Q(\hat{A})$, the estimated standard errors were close to the standard deviations in Table III (results not shown), suggesting that these standard errors are reasonable measures of uncertainty to present along with the estimate $\hat{Q}(\hat{A})$.

5.2. *Null-case results*

From Table I, we can see that the tree-based methods again tend to identify too many covariates as important, whereas the logit-based method finds almost zero. Moreover, none of the tree-based methods fails to find a tree more than 38% of the time, whereas an intercept-only model is chosen 85% of the time for the logit-based method. This could be due to the fact that the logit-based method uses the same scale from which the data were generated, whereas the tree-based methods change to the probability scale. Between the tree-based methods, VT(C) finds empty trees more frequently than VT(R), and defining $\hat{A}$ using $Z_i^* > \delta + 0.1$ noticeably outperforms $Z_i^* > \delta + 0.05$. It appears that, even in the absence of a region of enhanced treatment effect, the tree-based methods still identify main effects as important, whereas the logit-based method does not. Similar results can be seen in Table II. The most conservative of the tree-based methods in this case ($\text{VT(R)}_{0.1}$) still identifies a subgroup of enhanced treatment effect 35% of the time, whereas the logit-based method identifies such a subgroup only about 15% of the time. Between the tree-based methods, the classification methods tend to identify subgroups more frequently than the corresponding regression-based methods. The specificities for all methods were reasonably good, although again the logit-based method tended to outperform the tree-based methods. As expected, specificities for $\hat{A}$ defined by $Z_i > \delta + 0.1$ are better than those for $Z_i > \delta + 0.05$.

From Table III, we can see that, as expected, subgroups identified by all methods have essentially no real enhanced treatment effect. The logit-based method has estimates ($\hat{Q}(\hat{A})$) which are generally much closer to zero than the corresponding tree-based methods. Between the tree-based methods, VT(R) appears to perform better for SND estimates, whereas the other estimates tend to be similar or slightly better for VT(C). In addition, except for the SND estimates, defining $\hat{A}$ using $Z_i > \delta + 0.05$ tends to be similar or slightly better than $Z_i > \delta + 0.1$ for VT(R), whereas for VT(C) this relationship is reversed. Also, the resubstitution, SND, and cross-validation-based SND methods tend to overestimate $Q(\hat{A})$, with the resubstitution method overestimating the most, whereas the full cross-validation-based resubstitution underestimates $Q(\hat{A})$. The bias-corrected estimates are again the closest to the true values of $Q(\hat{A})$; however, the bias-corrected estimates tend to somewhat overestimate $Q(\hat{A})$, particularly in the case of the tree-based methods.

5.3. *Modifications to the base case*

We also consider a number of modifications to the base case:

(i) 30 covariates instead of 15.
(ii) Correlated covariates, in which variables in the three clusters $\{X_1, X_3, X_7\}$, $\{X_2, X_4, X_{15}\}$, and {all remaining covariates} have internal correlations of 0.7 but are uncorrelated with variables from the other clusters.
(iii) Subject-specific effects, using data generation models $\text{logit}(P(Y_i = 1)) = a_i - 1 + 0.5X_1 + 0.5X_2 - 0.5X_7 + 0.1T + 0.5X_2X_7 + \theta TI(\boldsymbol{X} \in A)$ and $\text{logit}(P(Y_i = 1)) = a_i - 1 + 0.5X_1 + 0.5X_2 - 0.5X_7 + 0.1T + b_iT + 0.5X_2X_7 + \theta TI(\boldsymbol{X} \in A)$, where a_i and b_i are both normally distributed with mean zero and variance 0.25.
(iv) Sample sizes of 400 and 2000.
(v) True A redefined to be $X_1 \geqslant -0.545 \cap X_2 \leqslant 0.545$, so that A is approximately 50% of the data set.
(vi) $\theta = 1.5$.

For all situations, we present only the results for VT(R) with threshold $\delta + 0.05$. This choice was made for illustrative purposes. The approximate power in a test of $\theta = 0$ for the correct logit model ranged between 90% and 100% for all situations except for the $n = 400$ case, in which the power was found to be approximately 60%.

Results for these cases are compared with the base case in the upper sections of Tables IV and V, with Table V being limited to the SND method for Q estimation. As expected, the average value of $Q(\hat{A})$ is

Table IV. 'Finding A' performance: variations of base case ($\theta = 0.90$, true $A = \{X_1 > 0, X_2 < 0\}$) and null case ($\theta = 0$ true A null) for $\text{VT(R)}_{0.05}$.

	Sensitivity	Specificity	AUC	PPV	NPV	Proportion $\hat{A}$ null	Median $\lvert\hat{A}\rvert$	$\{X_1, X_2\}$ $\in$ tree	# Unique Xs
$\theta = 0.9$									
Base case	0.47	0.89	0.77	0.55	0.84	0.07	189	0.51	3.62
$p = 30$	0.48	0.89	0.78	0.51	0.85	0.15	213	0.52	2.89
Correlated	0.52	0.85	0.76	0.52	0.85	0.03	247	0.32	4.08
Subject-specific (a_i)	0.54	0.89	0.79	0.60	0.86	0.04	218	0.62	3.59
Subject-specific (a_i, b_i)	0.50	0.89	0.76	0.57	0.85	0.05	216	0.48	3.61
$n = 400$	0.40	0.82	0.66	0.42	0.81	0.05	93	0.19	4.09
$n = 2000$	0.68	0.93	0.87	0.74	0.90	0.03	462	0.78	3.03
Larger $\lvert A\rvert$	0.41	0.91	0.78	0.77	0.61	0.06	233	0.54	3.35
$\theta = 1.5$	0.78	0.91	0.89	0.77	0.93	0.00	245	0.83	3.39
$\theta = 0$									
Base case	–	0.87	–	–	–	0.25	131	–	3.29
$p = 30$	–	0.88	–	–	–	0.27	103	–	2.90
Correlated	–	0.82	–	–	–	0.09	164	–	4.19
Subject-specific (a_i)	–	0.84	–	–	–	0.17	152	–	3.44
Subject-specific (a_i, b_i)	–	0.86	–	–	–	0.28	131	–	3.25
$n = 400$	–	0.78	–	–	–	0.05	86	–	4.10
$n = 2000$	–	0.88	–	–	–	0.25	230	–	2.84

Table V. SND Q estimation: variations of base case ($\theta = 0.90$, true $A = \{X_1 > 0, X_2 < 0\}$) and null case ($\theta = 0$ true A null) for $\text{VT(R)}_{0.05}$.

	$Q(A)$	$Q(\hat{A})$	$\hat{Q}(\hat{A})$	$\hat{Q}(\hat{A})_{BC}$
$\theta = 0.9$				
Base case	0.139	0.070 (0.047)	0.111 (0.046)	0.072 (0.044)
$p = 30$	0.137	0.062 (0.047)	0.092 (0.049)	0.069 (0.048)
Correlated	0.129	0.055 (0.040)	0.110 (0.035)	0.061 (0.036)
Subject-specific (a_i)	0.133	0.078 (0.043)	0.113 (0.043)	0.073 (0.046)
Subject-specific (a_i, b_i)	0.126	0.072 (0.041)	0.106 (0.043)	0.064 (0.048)
$n = 400$	0.137	0.046 (0.052)	0.107 (0.050)	0.051 (0.056)
$n = 2000$	0.135	0.096 (0.044)	0.109 (0.036)	0.081 (0.042)
Larger $\lvert A\rvert$	0.094	0.066 (0.033)	0.105 (0.042)	0.065 (0.045)
$\theta = 1.5$	0.228	0.167 (0.057)	0.156 (0.040)	0.132 (0.056)
$\theta = 0$				
Base case	0.000	0.010 (0.023)	0.074 (0.054)	0.036 (0.045)
$p = 30$	0.000	0.016 (0.026)	0.072 (0.053)	0.043 (0.043)
Correlated	0.000	0.006 (0.023)	0.095 (0.043)	0.041 (0.038)
Subject-specific (a_i)	0.000	0.010 (0.029)	0.079 (0.046)	0.037 (0.036)
Subject-specific (a_i, b_i)	0.000	0.011 (0.028)	0.063 (0.049)	0.025 (0.037)
$n = 400$	0.000	0.010 (0.035)	0.099 (0.046)	0.038 (0.046)
$n = 2000$	0.000	0.003 (0.017)	0.064 (0.045)	0.035 (0.034)

Note: SD in parentheses.

noticeably smaller than $Q(A)$ when $n = 400$ and more similar to $Q(A)$ when $n = 2000$, indicating that the ability to find subgroups with good properties decreases with decreasing sample size. From Table IV we can see that, with the exception of the $n = 400$ case, which is noticeably worse than the others, and the $n = 2000$ and $\theta = 1.5$ cases, which are noticeably better, the method's ability to find the correct subgroup seems to be somewhat unaffected by moderate variations on the base case. The NPVs are quite good for

all cases other than the Larger $|A|$ case, and with the exception of the $n=400$ case, all cases lead to very good AUC values. Furthermore, the method seems to find subgroups that are similar in size to the true subgroup for all cases and rarely fails to find a subgroup of any sort. From the upper section of Table V, we can see that, other than $n=400$ and $n=2000$ cases, the method appears to estimate $Q(\hat{A})$ with similar accuracy under all cases when the bias correction is made. Although reasonably good, the bias-corrected estimates for the $n=2000$ and $\theta=1.5$ cases are negatively biased, which is somewhat counterintuitive, as one would expect increased sample size and increased signal to lead to more accurate estimates. The SND estimates for all cases other than $\theta = 1.5$ show some positive bias, which is, for the most part, removed once we implement the bias correction. Increasing the number of covariates to 30 does not change the properties much, except for increasing the percentage of times that no subgroup is found.

Null cases ($\theta = 0$) for each of these modifications are also considered, and comparisons of these with the null case are given in the lower sections of Tables IV and V. We can see from Table IV that for specificity, median $|\hat{A}|$, and number of unique Xs, the method seems somewhat insensitive to variation of the null case; however, the correlated and $n = 400$ cases show substantially fewer instances of $|\hat{A}| = 0$ than the other cases. Specificities under all cases are reasonably good; however, the method continues to find subgroups even when no true subgroup exists, identifying a subgroup at least 72% of the time in all cases. From Table V, we see that our method shows similar performance under all null cases, particularly when the bias correction is used. As expected, the estimated subgroups for all cases have, on average, no enhanced treatment effect. As we saw in the base case, the SND estimates are positively biased, but in this case, although the bias correction somewhat reduces this positive bias, some bias still remains. Although there is some positive bias, these estimates are relatively close to zero, so although the method continues to identify subgroups when no true subgroup exists, it is unlikely that such subgroups would be falsely identified as important.

6. Application to clinical trial data

The example is taken from a clinical trial conducted by Eli Lilly, and as the specific information is still confidential, the problem and solution will be described in general terms. The data come from a randomized, double-blind clinical trial in patients with a potentially fatal condition. Data from this study include 1019 individuals, 517 of whom received the experimental treatment in addition to the standard of care; the remaining patients received placebo with the standard of care. The intervention is a drug that is intended to improve survival, and as such, the agreed upon endpoint with the Food and Drug Administration was survival at 28 days post-randomization to treatment/placebo. We consider 44 covariates ($X_1 - X_{44}$), including demographic, laboratory, medical history, and questionnaire data. Of these, nine are binary, 14 are regarded as continuous, and 21 are categorical. The 21 categorical variables were subdivided using dummy variables, giving an overall total of 60 Xs. The overall treatment effect is 0.069 (SE = 0.028), indicating a modest overall survival benefit for the experimental treatment. For all methods, we define $\hat{A}$ using the threshold $\delta + 0.05 = 0.119$.

Applying the forward logistic approach found 30 main effects, 77 $X \times X$ interactions, and 11 $X \times T$ or $X \times X \times T$ interactions. The Xs that have the most significant interactions with T are X_1, X_2, X_{11}, X_{18}, X_{39}, and X_{41}, three of which are prognostic factors for survival and another two of which have biological plausibility. The resubstitution estimate of $Q(\hat{A})$ is 0.285, and the bias-corrected SND estimate is -0.042. For this method, the resulting estimated subgroup contains 290 individuals (151 treated). Although many main effects and interactions were included in the final logistic model, the bias-corrected estimates of $Q(\hat{A})$ are essentially zero, suggesting that no meaningful subgroup was found.

Applying VT(R) led to a tree in which $\hat{A} = \{X : X_8 < 58.22 \text{ and } X_{44} = 1, \text{ or } X_8 \in [58.22, 180.9) \text{ and } X_{12} \geq 9,938\}$, and similarly, VT(C) led to a tree in which $\hat{A} = \{X : X_8 < 59.19 \text{ and } X_{44} = 1 \text{ and } X_1 \geq 70.01, \text{ or } X_8 < 59.19 \text{ and } X_{44} = 1 \text{ and } X_1 < 70.01 \text{ and } X_{12} \geq 3,304\}$. These subgroups for VT(R) and VT(C) included 233 (126 treated) and 143 (77 treated) individuals, respectively. Although our analyses focus on the threshold $\delta + 0.05$ for defining $\hat{A}$, it should be noted that the Virtual Twins approaches failed to identify subgroups for the threshold $\delta + 0.1$. The resubstitution estimates of $Q(\hat{A})$ for VT(R) and VT(C) are 0.179 (SE = 0.107) and 0.197 (SE = 0.101), respectively, and the bias-corrected SND estimates are 0.047 (SE = 0.048) and 0.041 (SE = 0.039). The variables that appear to be potentially useful are X_1, X_8, X_{12}, and X_{44}, which include three laboratory biomarkers and one demographic variable. Three of these variables are related to the severity of the

patient's condition and are also prognostic for the survival outcome. All four variables have some biological plausibility. However, although the estimates of $Q(\hat{A})$ for the Virtual Twins methods are larger than the corresponding estimates for the logistic method, they are still modest in size and only slightly larger than their standard errors. This suggests that there is insufficient evidence from this trial that the subgroups found by the Virtual Twins methods have a significantly better outcome.

This example shows the benefits of the Virtual Twins method over the logistic method in three respects: the results are more easily interpreted, many fewer variables are identified as important, and $\hat{A}$ has better properties. Additionally, the computational time for the Virtual Twins method was substantially better than that for the logistic method.

7. Discussion

Clinical trial research often includes reports of subgroup analyses in order to explore differential treatment effects, oftentimes as post hoc or exploratory analyses. With the advent of tailored therapeutics, much has been published on genetic or other biomarkers that may be predictive of patient outcomes or differential treatment effects and thus may be useful to define subgroups. There are many dimensions to this problem (effect size in both the overall populations and in the subgroup of interest, subgroup size, study sample size, the number of covariates of interest, the number of covariates that define the subgroup, the correlation between covariates, etc.) that give it considerable complexity. Although many have commented on the dangers of subgroup analyses, whether planned or unplanned, there has been little serious investigation of methodologies for proper identification of subgroups and assessment of their reliability other than routine, conservative multiplicity adjustments on the number of treatment-by-subgroup interaction tests done.

In clinical drug development programs that are exploring an array of potential biomarkers, phase 2 trials usually include dozens or a few hundred patients at most. Our simulations of $n = 400$ patients indicate that single trials are unlikely to find the right biomarkers when the effect size is important but the subgroup size is modest. For satisfactory results, larger studies are needed, as would occur with phase 3 trials. Good results are within reach when the sample sizes are $n = 2000$, but such sample sizes are not always used in clinical drug development except for the largest event-driven trials in a few disease indications such as cardiovascular disease or osteoporosis. In trials of $n = 1000$ patients, the method we propose demonstrates reasonably good properties for identification of meaningful clinical effect sizes. Although in principle the methods we present could be viewed as providing definitive evidence for a subgroup, in practice the methods are more useful for giving leads and suggestions for future work and better than what one could achieve by simply looking for interactions. In reality, an actual new trial would be required for the results to be confirmed and accepted.

As we show, identifying subgroups of enhanced treatment effect is a challenging problem that would generally require large data sets. Verifying that the subgroups are real and are likely to be useful in future data is even more challenging. The Virtual Twins method appears to be a promising approach and better than a simple alternative for identifying a subgroup. For the cases we consider, the identified subgroup had reasonable properties as measured by sensitivity, specificity, and magnitude of the enhanced treatment effect; however, the method is less good at identifying the correct covariates. One reassuring finding from the simulations is that increasing the number of potential covariates, making the covariates correlated, or introducing between-subject heterogeneity does not appear to have much of an effect on the properties of the estimated subgroup, although we doubt this finding would hold if there were hundreds of covariates instead of the tens we had. One drawback of the Virtual Twins method is that it has a tendency to identify a subgroup even when there should not be one. One strategy to use to mitigate any consequence of this is to accept that an aggressive strategy may find subgroups too often but then to accompany the subgroup with an estimate of how good it is and an associated measure of uncertainty. To obtain an honest estimate of the enhanced treatment effect in the identified subgroup, we find that a bias-corrected bootstrap procedure gives good but not perfect results.

In any real situation, there are likely to be a number of logistical and practical considerations. For example, in the methods we have treated all the Xs equally; however, in reality they will not be equivalent. For example, there may be an *a priori* biological rationale why some of the Xs are more likely to have an interaction with treatment in a model. Another issue is that some Xs may be cheaper or easier to measure than others. Thus, it could be desirable to have the region $\hat{A}$ based on such Xs, provided this

does not substantially harm the properties of $\hat{A}$. Another practical concern is the size of $\hat{A}$. Depending on the context, it may be desirable to have a large $\hat{A}$ with a modest enhanced treatment effect or a much smaller $\hat{A}$ with a larger treatment effect.

In the simulations, we report whether the methods find the 'right' Xs; however, in some contexts this may not be such an important issue. The Xs are likely to be correlated with each other, so it is quite plausible that a 'wrong' set of Xs could be quite effective at defining the subjects who do have an enhanced treatment effect. The very concept of a set of 'right' Xs is itself too idealized. Although for some treatments it may be possible to hypothesize an all-or-nothing situation where the treatment will only work if the person has a certain set of attributes, it may be difficult to develop technology that can accurately and reliably measure these attributes. On the other hand, if a subgroup is to be based on a set of Xs, then to be accepted by the scientific community, these Xs would have to have at least some biological plausibility.

We have chosen to summarize the properties of the estimated subgroup using a metric $Q(\hat{A})$, which is the treatment effect in the subgroup minus the marginal treatment effect. An alternative metric is the treatment effect in $\hat{A}$ minus the treatment effect in the complement of $\hat{A}$. From the perspective of the drug developer, neither metric is totally satisfactory because they do not incorporate the size of $\hat{A}$. For a given enhanced treatment effect, a larger $\hat{A}$ is more desirable than a small $\hat{A}$. It is unlikely that a single summary measure can capture all the costs and benefits of the identified subgroup from the drug developer's or society's point of view.

Although we demonstrate that the Virtual Twins method appears promising, many variations could be contemplated. For example, in the first step the random forest could be replaced by other non-parametric regression methods, such as multivariate adaptive regression splines. This could give smoother forms for the estimate of $P(Y = 1|X, T)$ as a function of each continuous X, which could be plausible in many applications. In the second step, a regression tree is appealing because of its simplicity and its ability to select a small number of features. Other feature selection methods could also be considered. For instance, Kehl and Ulm [13] developed a method for censored data in which the second step involves the use of stabilized bump hunting rather than a regression tree. Variations on cross-validation and the bootstrap could be used to assess the properties of the estimated subgroup.

We consider the situation of a binary outcome variable, which gives rise to some issues, specifically whether one should consider treatment effects as measured on an absolute probability scale or on a logit scale. Whereas we simulated data using models on a logit scale, we assessed the enhanced treatment effect on the probability scale. It is likely that the methodology we describe would have looked better if we had used the same scale for both. What scale one uses to assess enhanced treatment effect depends on the specific application. The reason to use the probability scale for assessing the properties of the region is because this would translate directly into a number of people who would benefit from the treatment. If the outcome were a continuous measure, then the Virtual Twins methodology would still be applicable, and for this situation there would be no reason to have any scale of measurement other than linear. We hypothesize that the methods described in this paper would perform better for continuous outcome data, not only because of the uniform definition of the scale but also because there is inherently more information in continuous data than in binary data.

A subtle concern with the goal of finding subgroups using data from a specific randomized trial is that the trial may have had exclusion criteria. For example, there may have been a strong suspicion that the treatment would not be effective for low values of a certain covariate, and hence people with such low values were excluded from the trial. Criteria such as these may hinder the ability to find a meaningful interaction, either with this covariate or with others, that would be present in a broader population.

We have formulated the goal as finding a subgroup of patients based on a small number of Xs that have an enhanced treatment effect. A related goal, which in the long run might be more useful for patients, is giving each patient their predicted probability of response under each possible treatment, i.e., giving them $\hat{P}_{1i}$ and $\hat{P}_{0i}$.

Acknowledgements

The authors acknowledge the valuable suggestions of Yuefeng Lu. This research was partially supported by a grant from the Eli Lilly Corporation to the University of Michigan and by National Institutes of Health Grant T32 CA083654.

References

1. Ruberg SJ, Chen L, Wang Y. The mean doesn't mean as much anymore: finding sub-groups for tailored therapeutics. *Clinical Trials* 2010; **7**:574–583.
2. Assmann SF, Pocock SJ, Enos LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet* 2000; **355**:1064–1069.
3. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Smith GD. Subgroup analyses in randomized controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technology Assessment* 2001; **5**(33):1–56.
4. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *Journal of Clinical Epidemiology* 2004; **57**:229–236.
5. Cui L, Hung HMJ, Wang SJ, Tsong Y. Issues related to subgroup analysis in clinical trials. *Journal of Biopharmaceutical Statistics* 2002; **12**(3):347-358.
6. Lagakos SW. The challenge of subgroup analyses—reporting without distorting. *New England Journal of Medicine* 2006; **354**(16):1667–1669.
7. Peto R, Collins R, Gray R. Large-scale randomized evidence: large simple trials and overviews of trials. *Journal of Clinical Epidemiology* 1995; **48**:23–40.
8. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine* 2002; **21**:2917–2930.
9. Rothwell PM. Subgroup analysis in randomized controlled trials: importance, indications and interpretation. *The Lancet* 2005; **365**:176–186.
10. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *The Journal Of The American Medical Association* 1991; **266**(1):93–98.
11. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine* 2007; **357**(21):2189–2194.
12. Feinstein AR. The problem of cogent subgroups: a clinicostatistical tragedy. *Journal of Clinical Epidemiology* 1998; **51**(4):297–299.
13. Kehl V, Ulm K. Responder identification in clinical trials with censored data. *Computational Statistics & Data Analysis* 2006; **50**(5):1338–1355. http://dx.doi.org/10.1016/j.csda.2004.11.015.
14. Dixon DO, Simon R. Bayesian subset analysis. *Biometrics* 1991; **47**:871–881.
15. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 1985; **41**:361–372.
16. Simon R. Bayesian subset analysis: application to studying treatment-by-gender interactions. *Statistics in Medicine* 2002; **21**:2909–2916.
17. Negassa A, Ciampi A, Abrahamowicz M, Shapiro S, Boivin JR. Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria. *Statistics and Computing* 2005; **15**:231–239.
18. Su X, Tsai CL, Wang H, Nickerson DM, Bogong L. Subgroup analysis via recursive partitioning. *The Journal of Machine Learning Research* 2009; **10**:141–158.
19. Su X, Zhou T, Yan X, Fan J, Yang S. Interaction trees with censored survival data. *The International Journal of Biostatistics* 2008; **4**(1): Article 2.
20. Freidlin B, Jiang W, Simon R. The cross-validated adaptive signature design. *Clinical Cancer Research* 2010; **16**(2):691–698.
21. Breiman L. Random forests. *Machine Learning* 2001; **45**(1):5–32.