A permutation procedure to detect heterogeneous treatment effects in randomized clinical trials while controlling the Type-I error rate

Jack M. Wolf[1], Joseph S. Koopmeiners[1], David M. Vock[1]

Abstract

Background/Aims: Secondary analyses of randomized clinical trials often seek to identify subgroups with differential treatment effects. These discoveries can help guide individual treatment decisions based on patient characteristics and identify populations for which additional treatments are needed. Traditional analyses require researchers to pre-specify potential subgroups to reduce the risk of reporting spurious results. There is a need for methods that can detect such subgroups without *a priori* specification while allowing researchers to control the probability of falsely detecting heterogeneous subgroups when treatment effects are uniform across the study population.

Methods: We propose a permutation procedure for tuning parameter selection that allows for Type-I error control when testing for heterogeneous treatment effects framed within the Virtual Twins procedure for subgroup identification. We verify that the Type-I error rate can be controlled at the nominal rate and investigate the power for detecting heterogeneous effects when present through extensive simulation studies. We apply our method to a secondary analysis of data from a randomized trial of very low nicotine content cigarettes.

Results: In the absence of Type-I error control, the observed Type-I error rate for Virtual Twins was between 99 and 100%. In contrast, models tuned via the proposed permutation were able to control the Type-I error rate and detect heterogeneous effects when present. An application of our approach to a recently completed trial of very low nicotine content cigarettes identified several variables with potentially heterogeneous treatment effects.

Conclusions: The proposed permutation procedure allows researchers to engage in secondary analyses of clinical trials for treatment effect heterogeneity while maintaining the Type-I error rate without pre-specifying subgroups.

Keywords

permutation test, subgroup identification, treatment effect heterogeneity, Type-I error, Virtual Twins

Introduction

The primary objective of a randomized controlled trial (RCT) is typically to estimate and test the marginal treatment effect (i.e., the average treatment effect aggregated across the entire population). However, identifying subgroups with a differential response to treatment has long been an important scientific and secondary aim, which has grown in importance in the era of personalized medicine. This focus is motivated by the idea that the treatment effect may vary from individual to individual across the population, commonly referred to as treatment effect heterogeneity. Towards this aim, researchers are interested in identifying characteristics that can explain differences in the treatment effect and subgroups for which the treatment effect is different than the effect in the broader population. Characterizing treatment effect heterogeneity can help effectuate personalized medicine, deepen our understanding of possible treatment mechanisms, and suggest subgroups which may benefit from different or more intensive intervention. For example, in studies evaluating proposed regulations that would affect all members of a population, such as trials of very low nicotine content cigarettes under the broader umbrella of tobacco regulatory science[1] which motivate our work here, it is important to identify subgroups that may not have an ideal response to the policy so that additional targeted interventions can be developed in support of these populations[2].

Traditionally, subgroup analyses in an RCT must be pre-specified in the statistical analysis plan. Although pre-specification permits easier control of the family-wise error rate, the number of pre-specified subgroups is typically limited, involving only a small number of covariates, and any categorization of a continuous variable must be pre-specified as well.

Given the limitations of pre-specifying subgroups, many statistical methods for evaluating effect heterogeneity and discovering subgroups using flexible, data-adaptive methods have been proposed and studied. Existing tree-based methods include interaction trees[3], honest causal trees[4], and GUIDE[5]. Moreover there exist ensemble methods that leverage the estimates of many models such as random forests of interaction trees[6] and STIMA[7], which combines a multiple linear regression model with a regression tree to detect interaction effects. Bayesian approaches have also been proposed[8,9].

[1]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States

Corresponding author:
Jack M. Wolf
Email: wolfx681@umn.edu

While identifying heterogeneous subgroups is an important scientific aim, there is a long history of finding spurious subgroups that cannot be replicated in subsequent trials [10–14]. Thus, researchers are naturally concerned with the *a priori* probability of incorrectly detecting subgroups with differential treatment effects when there is a uniform treatment effect across the study population and whether the aforementioned probability can be controlled while maintaining sufficient power to detect heterogeneity when present. This is a particular concern for many data-adaptive approaches for which such control may be difficult to implement.

In this paper, we propose a permutation procedure for identifying treatment effect heterogeneity with Type-I error rate control. We frame our approach in the context of Virtual Twins [15], which is a popular two stage approach to subgroup detection that has been widely used and discussed since its original publication [16–20]. Despite the method's wide usage, there is currently little guidance on how to select penalty parameters needed to fit such models. Many researchers are left using default software settings in their applications which are typically selected with predictive performance in mind. We address this limitation by conceptualizing Virtual Twins as a hypothesis testing procedure and showing how parameters can be tuned to accurately control the associated Type-I error rate.

The rest of the paper proceeds as follows. First, we establish notation and review Virtual Twins as originally proposed. Next, we propose a permutation procedure to assist tuning parameter selection and Type-I error control. Then, we detail several simulation studies to assess our proposed method's performance in a variety of scenarios. Finally, we apply this method to data from an RCT of very low nicotine content cigarettes to describe patient characteristics that may impact smokers' individual responses to the intervention.

Notation and preliminaries

Notation

Consider the data $(Y_i, T_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$ from an RCT with response Y_i, binary treatment indicator T_i, and covariates $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{pi})$ (which may be continuous or categorical). We write the conditional mean of Y_i given T_i and $\boldsymbol{X}_i$ as

$$\mathrm{E}(Y_i|T_i, \boldsymbol{X}_i) = h(\boldsymbol{X}_i) + T_i g(\boldsymbol{X}_i) \tag{1}$$

We assume that T_i is associated with Y_i only through its first moment and that it does not affect the conditional variance or any higher moments of the response. However, we do not require such assumptions about the relationship between all other covariates and the response. The conditional average treatment effect for each subject is $\mathrm{E}(Y_i|T_i = 1, \boldsymbol{X}_i) - \mathrm{E}(Y_i|T_i = 0, \boldsymbol{X}_i) = g(\boldsymbol{X}_i)$, which we denote as Z_i.

Virtual Twins

Virtual Twins is a two-step approach that first estimates Z_i, typically using flexible regression techniques, and then models the estimated $\hat{Z}_i$ using parsimonious and interpretable models. Although in principle an analyst could fit the main effect and interpretable treatment model simultaneously, the two-step procedure provides maximum flexibility to explain variability in the main effects in Step 1, while providing an interpretable model in Step 2.

Step 1 Step 1 consists of estimating subjects' outcomes under both the control and treatment arms. This is accomplished by splitting the data based on the value of T_i and independently fitting flexible regression models $\hat{f}_0(\boldsymbol{X}_i)$ and $\hat{f}_1(\boldsymbol{X}_i)$ to estimate $\mathrm{E}(Y_i|T_i = 0, \boldsymbol{X}_i)$ and $\mathrm{E}(Y_i|T_i = 1, \boldsymbol{X}_i)$ respectively. Each subject's estimated conditional average treatment effect is then given by $\hat{Z}_i = \hat{f}_1(\boldsymbol{X}_i) - \hat{f}_0(\boldsymbol{X}_i)$. The original paper proposed using random forests [21] to estimate these quantities but other authors [22] have investigated the use of additional approaches to estimating the response surface in Step 1 including linear models fit using the lasso [23], MARS [24] and super learner [25].

Step 2 In Step 2 the analyst uses a simple and interpretable model such as a regression tree to model the estimated conditional average treatment effect $\hat{Z}_i$ as a function of the covariates $\boldsymbol{X}_i$. Variables included in this model are used to determine which covariates modify the treatment effect and identify subgroups of patients with homogeneous treatment effects. The original paper supports both regression and classification trees. Others [22] have explored using the lasso and conditional inference trees as possible Step 2 methods.

Type-I error rate control

The original presentation of Virtual Twins encourages fitting the Step 2 models using a fixed list of tuning parameters. While this approach has shown acceptable performance, data-adaptive methods for parameter selection based on performance metrics may be advantageous. However, standard data-adaptive methods typically select tuning parameters to maximize predictive performance which may not be optimal in this context. First, such an approach is not guaranteed to preserve any Type-I error of detecting heterogeneous treatment effects. Second, the estimated conditional average treatment effect (i.e., $\hat{Z}_i$) is a deterministic function of the features and, therefore, data-adaptive methods are likely to overfit the data. To address these limitations, we propose a permutation based framework to identify appropriate penalty parameters for a variety of Step 2 methods to maintain the Type-I error rate for concluding heterogeneity.

Framing as a hypothesis test

Controlling the Type-I error rate requires that we first frame Virtual Twins as a hypothesis test with a null hypothesis of a homogeneous treatment effect. This null hypothesis implies that each subject's treatment effect is equal and $g(\boldsymbol{X}_i) = \Delta$ for all i, so an individual's conditional average response can be simplified to:

$$\mathrm{E}(Y_i|T_i, \boldsymbol{X}_i) = h(\boldsymbol{X}_i) + T_i\Delta \tag{2}$$

We will reject the null hypothesis if the Step 2 model estimating $g(\boldsymbol{X}_i)$ includes any covariates (e.g., a tree with at least one split). Thus, a Type-I error corresponds to rejecting the null hypothesis when $g(\boldsymbol{X}_i)$ is constant.

Permutation procedure

We consider a class of methods for Step 2 which are fit by specifying one penalty parameter where for any fixed data set there exists a sufficiently large penalty parameter such that the fitted model is constant for all inputs. Examples of such methods include regression trees[26], conditional inference trees[27], and the lasso[23]. We will henceforth refer to the smallest penalty parameter that achieves this constant model for a fixed data set as the minimal null penalty parameter. Formally, we let $\hat{\theta}_N = \min\{\theta : \hat{g}(\boldsymbol{X}_i, \theta) = d \text{ for all } i\}$ be the minimal null penalty parameter of a given data set where $\hat{g}(\boldsymbol{X}_i, \theta)$ is the fitted Step 2 model given penalty parameter θ for a given data set.

We wish to identify the penalty parameter θ_α such that the procedure's Type-I error rate is at most α for any fixed $0 \leq \alpha \leq 1$. We estimate this parameter using a permutation procedure. Permutation tests are typically used to achieve exact inference in small sample sizes by deriving the null distribution of the test statistic without model-based assumptions. Our procedure slightly departs from this framework to describe the null distribution of the minimal null penalty parameter, $\hat{\theta}_N$, which is then used for parameter selection.

The procedure can be summarized as first permuting the treatment indicators to preserve the covariate main effects while eliminating potential treatment by covariate interactions, then fitting the Step 1 model to this permuted data to estimate the conditional average treatment effect under the null model, and finally fitting the Step 2 model to calculate $\hat{\theta}_N$ for the permuted data.

The proposed algorithm is as follows. First calculate the estimated mean treatment effect $\hat{\Delta}$ and obtain $\tilde{Y}_i = Y_i - \hat{\Delta} I(T_i = 1)$ to set the mean treatment effect to zero before permuting the treatment indicators. Note that testing for heterogeneity with Y_i is equivalent to doing so with $\tilde{Y}_i$. Then, for each j where $j = 1, \ldots, m$ for some large m,

1. Randomly permute the treatment indicator variables of the original data set to obtain $(\tilde{Y}_i, T_i^{*(j)}, \boldsymbol{X}_i)$.
2. Fit the Step 1 model on the permuted data to estimate $\hat{Z}_i^{*(j)}$ for each i.
3. Let $\hat{\theta}_N^{(j)}$ be the minimal null penalty parameter for the Step 2 model fit for $\hat{Z}_i^{*(j)}$.

Then, let $\hat{\theta}_\alpha$ be the $1-\alpha$ percentile of $\hat{\theta}_N^{(1)}, \ldots, \hat{\theta}_N^{(m)}$ and fit the Step 2 model for $\hat{Z}_i$ estimated through Step 1 on the original data using the penalty parameter $\hat{\theta}_\alpha$.

This permutation procedure can be reframed as a permutation test to leverage existing theory. Consider the test statistic $\hat{\theta}_N$ and the accompanying hypothesis test that rejects the null hypothesis of no effect heterogeneity if $\hat{\theta}_N > \theta_\alpha$. We wish to identify the largest critical value θ_α such that $\Pr(\hat{\theta}_N > \theta_\alpha | H_0) \leq \alpha$. This can be accomplished by taking the $1-\alpha$ percentile of the null distribution of $\hat{\theta}_N$ which we generate via the permutation procedure. This permutation test is valid because after permuting we have data generated equivalent in distribution under the null hypothesis that $\mathrm{E}(\tilde{Y}_i | T_i, \boldsymbol{X}_i) = h(\boldsymbol{X}_i)$ (recall the the main effect of treatment is 0 with $\tilde{Y}_i$).

Simulation studies

Simulation study design

Data were generated from the model $Y_i = h(\boldsymbol{X}_i) + T_i g(\boldsymbol{X}_i) + \epsilon_i$ where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ for $i = 1, \ldots, 1000$. Covariate vectors $\boldsymbol{X}_i$ consisted of p covariates with $p = 10, 20, 50$ with continuous and binary variables in a 4:1 ratio. Continuous covariates were generated from a multivariate normal distribution with an $\mathrm{AR}(1)$ correlation structure, and binary covariates were simulated from independent Bernoulli distributions. Treatment indicators were randomly assigned to give a 1:1 allocation ratio with 500 patients per arm.

Each subject's conditional average treatment effect was given via $g(\boldsymbol{X}_i)$. We considered a null scenario where $g(\boldsymbol{X}_i)$ was constant for all $\boldsymbol{X}_i$. Under this scenario we expected our permutation procedure to falsely detect heterogeneity with probability α. We also simulated scenarios where g was a linear and nonlinear function of $\boldsymbol{X}_i$ to assess our permutation procedure's power. The nonlinear g determined the conditional average treatment effect by partitioning the covariate space through several splits at the covariates' true mean values along with one or two interaction terms, depending on the number of covariates. The function $h(\boldsymbol{X}_i)$ can be viewed as a patient's expected outcome under the control. We examined scenarios where h was both a linear and nonlinear function of the covariates. The nonlinear h consisted of linear and quadratic terms, binary indicators for whether a covariate was above its true mean value, and covariate by covariate interactions.

We completed 2000 simulated trials for every combination of p, g, and h. Full simulation details are available in Table 1. The average R^2 when ignoring any treatment by covariate interactions was about 0.7 in each simulation to resemble our motivating data. Cohen's f^2, which measures the effect size of the interaction, ranged from 0.01 to 0.03 under simulations with effect heterogeneity. Additional details are provided in the supplementary materials. Table S1 displays the average R^2 for each scenario both when ignoring and including treatment by covariate interactions as well as Cohen's f^2. In addition, we ran a similar simulation study with increased residual variance to achieve an R^2 of about 0.2 in all simulated datasets.

Methods Considered

We implemented our permutation procedure using various methods for fitting models for Steps 1 and 2, the performance of which have been evaluated in the context of Virtual Twins[22].

In Step 1 we considered using random forests and super learner[25]. Random forests consisted of 1000 trees. The library for the super learner models used a linear model tuned using a lasso (L1) penalty, MARS[24], and a random forest.

We considered the lasso[23], regression trees[26], and conditional inference trees[27] as Step 2 methods, all of which require the specification of some penalty parameter before they can be fit, which we tuned to control the Type-I error rate. The lasso is a linear regression method that penalizes the L^1-norm of the non-intercept regression coefficients to perform variable selection and regularization. Our permutation procedure tuned the penalty term's weight.

Regression trees recursively partition the covariate space to identify subgroups of the data with similar response values. A dense tree is fit and then the optimal subtree that minimizes a weighted loss function that combines the mean squared error and the number of terminal nodes is selected. The weight assigned to the number of terminal nodes is the complexity parameter which we tuned through our procedure. Like regression trees, conditional inference trees also recursively partition the covariate space to identify groups with similar responses. Unlike regression trees, which are fit using measures of information, conditional inference trees use a significance test for variable selection; the tree only will split if a significance test comparing the mean outcome between both considered subgroups has test statistic greater than some pre-specified threshold, which we tuned through our permutation procedure.

We tested all combinations of the discussed Step 1 and Step 2 models using our permutation procedure with $\alpha = 0.2$ and $\alpha = 0.05$ and using standard parameter selection techniques. Additional details are available in Table S2.

Summaries of performance

We evaluated the performance of our permutation procedure using the following metrics:

Type-I error rate and power For each simulated trial we recorded whether the Step 2 model included any covariates or not. If the model included at least one covariate, we concluded that the model detected treatment effect heterogeneity. The average of this value across many simulations corresponds to the Type-I error rate or the power, depending on the absence or presence of treatment effect heterogeneity, respectively.

Sensitivity and specificity Consider the partition $\boldsymbol{X}_i = (\boldsymbol{X}_i^{\text{Heterogeneous}}, \boldsymbol{X}_i^{\text{Constant}})$ such that $g(\boldsymbol{X}_i) = g(\boldsymbol{X}_i^{\text{Heterogeneous}})$ for all i. We calculated the proportion of covariates in $\boldsymbol{X}_i^{\text{Heterogeneous}}$ that were included in the Step 2 model (sensitivity) as well as the proportion of covariates in $\boldsymbol{X}_i^{\text{Constant}}$ that were not included (specificity).

Individual treatment effect mean squared error We assessed accuracy in modeling the conditional average treatment effect by calculating the mean squared error: $\sum_{i=1}^{n}[\hat{g}(\boldsymbol{X}_i) - Z_i]^2/n$ where $\hat{g}(\boldsymbol{X}_i)$ is the fitted Step 2 model.

Simulation Results

Table 2 summarizes the proportion of simulated trials in which at least one covariate was included in the Step 2 model across different data generating models for $g(\boldsymbol{X}_i), h(\boldsymbol{X}_i)$, and number of covariates. In scenarios with a homogeneous treatment effect, this corresponds to the Type-I error rate. Implementations using standard approaches to selecting tuning parameters had Type-I error rates of nearly 100%. In contrast, when using our permutation procedure, we observed empirical Type-I error rates approximately equal to the targeted values. This metric is a model's empirical power in scenarios with effect heterogeneity. Across all such scenarios, the highest power was obtained when using a super learner model in Step 1 and the lasso to fit the Step 2 model (regardless of whether $g(\boldsymbol{X}_i)$ was linear or nonlinear). These trends held regardless of the number of covariates but the power for a specific combination of methods tended to increase with the number of covariates (See Table S3 for results when $p = 20$). We observed similar trends with less power in supplemental simulations with a lower overall R^2 value (Table S4).

Table S5 shows the sensitivity and specificity of each combination of Step 1 and Step 2 methods for each scenario. When the conditional average treatment effect was linear, the controlled lasso had the highest sensitivity of all controlled methods. When modeling a nonlinear conditional average treatment effect with Type-I error control the lasso tended to have the highest sensitivity of all methods for a given error rate. The sensitivity was relatively constant regardless of the number of covariates for all combinations of methods. Nearly all methods that controlled the Type-I error rate demonstrated near perfect specificity for all scenarios. When the Type-I error rate was not controlled, the specificity ranged from 0.01 to 0.96 and was lowest when the lasso was used in Step 2.

Table S6 displays the estimated mean squared error for the subject-specific conditional average treatment effect for all combinations of Step 1 and Step 2 methods. In the absence of treatment effect heterogeneity, the methods that did not attempt to control the Type-I error rate had mean squared errors substantially higher than their counterparts which controlled the error rate. When the treatment effect was heterogeneous, methods that fixed the Type-I error rate at $\alpha = 0.2$ tended to have the lowest mean squared error when compared to models without Type-I error control and with $\alpha = 0.05$ for all Step 1 methods. The mean squared error increased with the number of covariates except for models controlling the Type-I error when the treatment effect was homogeneous, for which the mean squared error remained constant as the number of covariates increased. Across all 18 simulation scenarios, using super learner and the lasso to fit the Step 1 and 2 models, respectively resulted in the smallest mean squared error out of all method combinations which control the Type-I error in Step 2.

Application

Smoking remains the leading cause of preventable death in the United States. Currently, researchers are considering the impact of multiple regulatory interventions to reduce the negative health effects of cigarette smoking, including reducing the nicotine content of cigarettes [1,2,28–31]. Reducing the nicotine content of cigarettes would impact all smokers in the United States. While RCTs have investigated the benefit of such regulations on the U.S. smoking population, *on average*, it is also important to identify potential subgroups that receive less benefit from or are potentially harmed by such regulations to design additional targeted interventions to reduce smoking or minimize unintended consequences.

A recent RCT [1] evaluated the impact of nicotine reduction in a randomized, double-blind trial that assigned subjects to one of three interventions: 1) immediate reduction in nicotine content, 2) gradual reduction in nicotine content, and 3) maintenance of standard tobacco cigarettes (i.e., the control condition) following a 2:2:1 allocation ratio. Subjects were

provided with cigarettes with nicotine content matching their treatment assignment for a 20-week intervention period, and the impact of nicotine reduction was evaluated by comparing the change in average number of cigarettes smoked per day from baseline to the last four weeks of the intervention.

Our application focused on comparisons between the gradual nicotine reduction group and the immediate nicotine reduction group as well as between the immediate reduction group and the control. We used the same modeling approaches in Step 1 and Step 2 as in our simulation study using 40 covariates that included demographic information and baseline smoking characteristics (Table S7 summarizes the study population along these covariates). We recorded which covariates were identified as having differential treatment effects in the Step 2 model.

Application results

Table 3 displays the covariates included in each model with Type-I error control when comparing the immediate and gradual groups and the immediate and control groups. The number of covariates included in each Step 2 model when the Type-I error rate was not controlled is presented in Table S8.

Immediate versus control Models exploring the effect of immediate reduction compared to the control detected at most one covariate (age) when controlling the Type-I error at 20%. Traditional approaches detected as many as 27 covariates. We note that due to the study design, this model had fewer observations ($n = 538$) than when comparing gradual to immediate reduction ($n = 723$) and had less power to detect differential effects.

Immediate versus gradual When modeling the effect of immediate versus gradual reduction using a random forest for Step 1 and a regression tree for Step 2 (as done in the original Virtual Twins paper) the covariates included are dependent on whether the model was tuned to control the Type-I error or not. When the error rate was not controlled six covariates were found to modify the the treatment effect. However, when controlling the error rate at either 20% or 5%, only total nicotine equivalents was found to modify the treatment effect. All other method combinations found at least one covariate in all but one case.

Given the combination's superior performance in simulation studies, we report the results found when when a super learner model in Step 1 was paired with the lasso in Step 2. We observed statistically significant treatment effect heterogeneity at the 0.05 significance level with total nicotine equivalents (nmol/ml) and cyanoethyl mercapturic acid/creatinine urine (nmol/mg) identified as likely treatment effect modifiers with estimated conditional average treatment effect associated with immediate nicotine reduction of $\hat{g}(\boldsymbol{X}_i) = 5.78 + 0.187X_{i1} + 0.009X_{i2}$, where X_{i1} and X_{i2} are centered and scaled (by the IQR) measures of total nicotine equivalents and cyanoethyl mercapturic acid, respectively. The results have important implications for tobacco regulatory science. While we observed significant treatment effect heterogeneity, the average treatment effect is 5.78 cigarettes per day. In contrast, the difference in the treatment effect associated with a difference equivalent to the interquartile range for total nicotine equivalents and cyanoethyl mercapturic acid is less than one cigarettes per day, which implies that the heterogeneity is small relative to the average treatment effect and that all smokers are likely to benefit from the intervention.

Discussion

We developed a permutation procedure that selects a tuning parameter which simultaneously regularizes the treatment effect heterogeneity and controls the Type-I error rate for detecting treatment effect heterogeneity in the Virtual Twins framework. This method both tests for heterogeneity and fits an estimated model for the conditional average treatment effect if there is heterogeneity. Our simulation results indicate that this procedure can control the Type-I error under a variety of null scenarios (e.g., with both linear and nonlinear covariate main effects, different numbers of measured covariates, etc.) and can detect treatment effect heterogeneity when it is present. Application to data from a recent RCT shows that when the Type-I error is not controlled, models tend to include far too many covariates to have face-validity or be useful to construct policy. Conversely, when controlling the error rate, our approach is able to detect covariates that are likely to modify the treatment effect based on our biological understanding of the intervention.

While many methods such as GUIDE, STIMA, and interaction trees have been proposed to detect subgroups with heterogeneous treatment effects or model the treatment effect, most if not all require the selection of some penalty parameter *a priori*. Proper specification of this parameter can control the Type-I error rate to a desired level; however, there is little to no guidance on how to select this parameter, and existing guidance often focuses on the model's mean squared error and not its Type-I error rate.

Our approach differs by offering explicit guidelines on how to select this parameter and control the overall Type-I error rate. Moreover, we note that our permutation procedure could be adjusted to aid parameter selection for GUIDE, STIMA, and interaction trees to facilitate Type-I error control.

Other methods control the Type-I error rate when testing the existence of treatment by covariate interactions. Some only offer a test for treatment effect heterogeneity and do not offer a method for describing the treatment by covariate interaction[32,33]. Additional work developed sophisticated permutation procedures to obscure treatment by covariate interactions while maintaining all other effects[34]. Although our proposed permutation procedure correctly controls the Type-I error of interest, these alternative permutation strategies may be more efficient and are worth investigating within our framework. Additionally, some have leveraged these permutation procedures to propose a method with the aim of identifying the appropriate complexity parameter for regression trees for the conditional average treatment effect estimated through propensity score matching such that the Type-I error rate is maintained and the conditional average treatment effect can be modeled if detected[35]. While our proposed method shares the goal of tuning regression trees' complexity parameters, it is far more general and can support any Step 2 method with a single tuning parameter.

While our method showed strong performance on both simulated and real data, some drawbacks are worth noting. First, although our approach controlled the Type-I error rate regardless of the methods used in Steps 1 and 2, the power depends on how well the model was matched to the true data generative process. For example, although using the lasso for Step 2 yielded the highest power in a majority of scenarios, the power of tree-based methods tended to be closer to the (optimal) power of methods using the lasso when the true treatment effect was a nonlinear rather than linear function of the covariate space. Additionally, we note that permutation based tests are often not efficient and that there is space to develop more powerful tests that can maintain the Type-I error rate. Moreover, while our approach maintains the Type-I error rate when testing for treatment effect heterogeneity, it does not offer any probabilistic statements about the specific variables included in the model. That is, while the probability of including any covariates under the null hypothesis is controlled, the *a priori* probability of selecting a specific covariate when it does not contribute to effect heterogeneity is unknown. Future work could develop a hierarchical testing procedure to first assess if there are heterogeneous effects via our proposed method and then, if that null hypothesis is rejected, devise a way to individually test the candidate covariates detected in the final model for treatment interactions while controlling the family-wise error rate. Additionally, the permutation process itself imposes moderate computational costs. Because the data must be permuted *before* the Step 1 model is fit (which is often a dense ensemble method that requires multiple fittings such as a random forest), the model must be refit for each permutation, which can lead to nontrivial computational demands. Finally, while our proposed method is situated within the context of RCTs, future work exploring the method's performance when applied to observational studies and extending the method to address any limitations in that context would be beneficial.

Although traditional guidelines recommend identifying potential subgroups *a priori* and testing for interactions to control the family-wise Type-I error rate[36], there is space for data-driven discovery. Our proposed approach allows researchers to control the Type-I error rate of an overall test for treatment effect heterogeneity and identify covariates and/or subgroups possibly associated with differential effects for future investigation. We note that this does not alleviate the need to account for conducting multiple hypothesis tests (e.g., also testing for the marginal treatment effect). However, our approach moves us towards a principled yet data-driven approach to discovery.

Software and code

The R package `tehtuner` implements our proposed method. The package and code to replicate the simulation studies can be downloaded at https://github.com/jackmwolf/tehtuner.

Acknowledgements

We would like to thank our collaborator, Dr. Dorothy Hatsukami, for providing access to the data used to illustrate our method.

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by the National Cancer Institute (Award Numbers R01CA214825 and R01CA225190), the National Institute on Drug Abuse (Award Numbers R01DA046320, and U54-DA031659) and National Center for Advancing Translational Science (Award Number UL1TR002494). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or Food and Drug Administration.

References

1. Dorothy K. Hatsukami, Xianghua Luo, Joni A. Jensen, Mustafa al'Absi, Sharon S. Allen, Steven G. Carmella, Menglan Chen, Paul M. Cinciripini, Rachel Denlinger-Apte, David J. Drobes, Joseph S. Koopmeiners, Tonya Lane, Chap T. Le, Scott Leischow, Kai Luo, F. Joseph McClernon, Sharon E. Murphy, Viviana Paiano, Jason D. Robinson, Herbert Severson, Christopher Sipe, Andrew A. Strasser, Lori G. Strayer, Mei Kuen Tang, Ryan Vandrey, Stephen S. Hecht, Neal L. Benowitz, and Eric C. Donny. Effect of Immediate vs Gradual Reduction in Nicotine Content of Cigarettes on Biomarkers of Smoke Exposure: A Randomized Clinical Trial. *JAMA*, 320(9):880–891, September 2018.
2. Dana M. Carroll, Bruce R. Lindgren, Sarah S. Dermody, Rachel Denlinger-Apte, Andrew Egbert, Rachel N. Cassidy, Tracy T. Smith, Lauren R. Pacek, Alicia M. Allen, Jennifer W. Tidey, Michael J. Parks, Joseph S. Koopmeiners, Eric C. Donny, and Dorothy K. Hatsukami. Impact of nicotine reduction in cigarettes on smoking behavior and exposure: Are there differences by race/ethnicity, educational attainment, or gender? *Drug and Alcohol Dependence*, 225:108756, August 2021.
3. Xiaogang Su, Chih-ling Tsai, Hansheng Wang, David M. Nickerson, Bogong Li, and Saharon Rosset. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 2009.
4. Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, July 2016.
5. Wei-Yin Loh, Xu He, and Michael Man. A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, 34(11):1818–1833, May 2015.
6. Xiaogang Su, Annette T. Peña, Lei Liu, and Richard A. Levine. Random Forests of Interaction Trees for Estimating Individualized Treatment Effects in Randomized Trials. *arXiv:1709.04862 [stat]*, September 2017. arXiv: 1709.04862.

7. Elise Dusseldorp, Claudio Conversano, and Bart Jan Van Os. Combining an Additive and Tree-Based Regression Model Simultaneously: STIMA. *Journal of Computational and Graphical Statistics*, 19(3):514–530, 2010. Publisher: [American Statistical Association, Taylor & Francis, Ltd., Institute of Mathematical Statistics, Interface Foundation of America].
8. Ciara Nugent, Wentian Guo, Peter Müller, and Yuan Ji. Bayesian Approaches to Subgroup Analysis and Related Adaptive Clinical Trial Designs. *JCO precision oncology*, 3:PO.19.00003, 2019.
9. Shonosuke Sugasawa and Hisashi Noma. Efficient screening of predictive biomarkers for individual treatment selection. *Biometrics*, 77(1):249–257, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.13279.
10. Xin Sun, Matthias Briel, Jason W Busse, John J You, Elie A Akl, Filip Mejza, Malgorzata M Bala, Dirk Bassler, Dominik Mertz, Natalia Diaz-Granados, Per Olav Vandvik, German Malaga, Sadeesh K Srinathan, Philipp Dahm, Bradley C Johnston, Pablo Alonso-Coello, Basil Hassouneh, Jessica Truong, Neil D Dattani, Stephen D Walter, Diane Heels-Ansdell, Neera Bhatnagar, Douglas G Altman, and Gordon H Guyatt. The influence of study characteristics on reporting of subgroup analyses in randomised controlled trials: systematic review. *BMJ*, 342, 2011.
11. Xin Sun, Matthias Briel, Jason W Busse, John J You, Elie A Akl, Filip Mejza, Malgorzata M Bala, Dirk Bassler, Dominik Mertz, Natalia Diaz-Granados, Per Olav Vandvik, German Malaga, Sadeesh K Srinathan, Philipp Dahm, Bradley C Johnston, Pablo Alonso-Coello, Basil Hassouneh, Stephen D Walter, Diane Heels-Ansdell, Neera Bhatnagar, Douglas G Altman, and Gordon H Guyatt. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ*, 344, 2012.
12. Benjamin Kasenda, Stefan Schandelmaier, Xin Sun, Erik von Elm, John You, Anette Blümle, Yuki Tomonaga, Ramon Saccilotto, Alain Amstutz, Theresa Bengough, Joerg J Meerpohl, Mihaela Stegert, Kelechi K Olu, Kari A O Tikkinen, Ignacio Neumann, Alonso Carrasco-Labra, Markus Faulhaber, Sohail M Mulla, Dominik Mertz, Elie A Akl, Dirk Bassler, Jason W Busse, Ignacio Ferreira-González, Francois Lamontagne, Alain Nordmann, Viktoria Gloy, Heike Raatz, Lorenzo Moja, Rachel Rosenthal, Shanil Ebrahim, Per O Vandvik, Bradley C Johnston, Martin A Walter, Bernard Burnand, Matthias Schwenkglenks, Lars G Hemkens, Heiner C Bucher, Gordon H Guyatt, and Matthias Briel. Subgroup analyses in randomised controlled trials: cohort study on trial protocols and journal publications. *BMJ*, 349, 2014.
13. Joshua D. Wallach, Patrick G. Sullivan, John F. Trepanowski, Kristin L. Sainani, Ewout W. Steyerberg, and John P. A. Ioannidis. Evaluation of Evidence of Statistical Support and Corroboration of Subgroup Claims in Randomized Clinical Trials. *JAMA Internal Medicine*, 177(4):554–560, 04 2017.
14. Sridharan Raghavan, Kevin Josey, Gideon Bahn, Domenic Reda, Sanjay Basu, Seth A. Berkowitz, Nicholas Emanuele, Peter Reaven, and Debashis Ghosh. Generalizability of heterogeneous treatment effects based on causal forests applied to two randomized clinical trials of intensive glycemic control. *Annals of Epidemiology*, July 2021.
15. Jared C. Foster, Jeremy M.G. Taylor, and Stephen J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880, October 2011.
16. Ralph van Hoorn, Marcia Tummers, Andrew Booth, Ansgar Gerhardus, Eva Rehfuess, Daniel Hind, Patrick M. Bossuyt, Vivian Welch, Thomas P. A. Debray, Martin Underwood, Pim Cuijpers, Helena Kraemer, Gert Jan van der Wilt, and Wietkse Kievit. The development of champ: a checklist for the appraisal of moderators and predictors. *BMC Medical Research Methodology*, 17(1):173, Dec 2017.
17. David M. Kent, Jessica K. Paulus, David van Klaveren, Ralph D'Agostino, Steve Goodman, Rodney Hayward, John P.A. Ioannidis, Bray Patrick-Lake, Sally Morton, Michael Pencina, Gowri Raman, Joseph S. Ross, Harry P. Selker, Ravi Varadhan, Andrew Vickers, John B. Wong, and Ewout W. Steyerberg. The predictive approaches to treatment effect heterogeneity (path) statement. *Annals of Internal Medicine*, 172(1):35–45, 2020. PMID: 31711134.
18. Kim C. M. Bul, Lisa L. Doove, Ingmar H. A. Franken, Saskia van der Oord, Pamela M. Kato, and Athanasios Maras. A serious game for children with attention deficit hyperactivity disorder: Who benefits the most? *PLoS ONE*, 13, 2018.
19. A. B. Apolo, J. A. Ellerton, J. R. Infante, M. Agrawal, M. S. Gordon, R. Aljumaily, T. Gourdin, L. Dirix, K. W. Lee, M. H. Taylor, P. Schöffski, D. Wang, A. Ravaud, J. Manitz, G. Pennock, M. Ruisi, J. L. Gulley, and M. R. Patel. Avelumab as second-line therapy for metastatic, platinum-treated urothelial carcinoma in the phase Ib JAVELIN Solid Tumor study: 2-year updated efficacy and safety analysis. *J Immunother Cancer*, 8(2), 10 2020.
20. Chirag Nagpal, Dennis Wei, Bhanukiran Vinzamuri, Monica Shekhar, Sara E. Berger, Subhro Das, and Kush R. Varshney. Interpretable subgroup discovery in treatment effect estimation with application to opioid prescribing guidelines. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 19–29, New York, NY, USA, 2020. Association for Computing Machinery.
21. Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.
22. Chuyu Deng, David M. Vock, Dana M. Carroll, Jeffrey A. Boatman, Dorothy K. Hatsukami, Ning Leng, and Joseph S. Koopmeiners. Practical Guidance on Modeling Choices for the Virtual Twins Method. *arXiv:2111.08741 [stat]*, November 2021. arXiv: 2111.08741.
23. Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. Publisher: [Royal Statistical Society, Wiley].
24. Jerome H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–67, March 1991. Publisher: Institute of Mathematical Statistics.
25. Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), September 2007. Publisher: De Gruyter Section: Statistical Applications in Genetics and Molecular Biology.
26. Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification And Regression Trees*. Routledge, Boca Raton, October 2017.
27. Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, September 2006. Publisher: Taylor & Francis _eprint:

https://doi.org/10.1198/106186006X133933.

28. Dorothy K. Hatsukami, Michael Kotlyar, Louise A. Hertsgaard, Yan Zhang, Steven G. Carmella, Joni A. Jensen, Sharon S. Allen, Peter G. Shields, Sharon E. Murphy, Irina Stepanov, and Stephen S. Hecht. Reduced nicotine content cigarettes: effects on toxicant exposure, dependence and cessation. *Addiction (Abingdon, England)*, 105(2):343–355, February 2010.
29. Neal L. Benowitz, Katherine M. Dains, Sharon M. Hall, Susan Stewart, Margaret Wilson, Delia Dempsey, and Peyton Jacob. Smoking behavior and exposure to tobacco toxicants during 6 months of smoking progressively reduced nicotine content cigarettes. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, 21(5):761–769, May 2012.
30. Dorothy K. Hatsukami, Louise A. Hertsgaard, Rachel I. Vogel, Joni A. Jensen, Sharon E. Murphy, Stephen S. Hecht, Steven G. Carmella, Mustafa al'Absi, Anne M. Joseph, and Sharon S. Allen. Reduced Nicotine Content Cigarettes and Nicotine Patch. *Cancer Epidemiology and Prevention Biomarkers*, 22(6):1015–1024, June 2013. Publisher: American Association for Cancer Research Section: Research Articles.
31. Neal L. Benowitz, Natalie Nardone, Katherine M. Dains, Sharon M. Hall, Susan Stewart, Delia Dempsey, and Peyton Jacob. Effect of Reducing the Nicotine Content of Cigarettes on Cigarette Smoking Behavior and Tobacco Smoke Toxicant Exposure: Two Year Follow Up. *Addiction (Abingdon, England)*, 110(10):1667–1675, October 2015.
32. Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Nonparametric Tests for Treatment Effect Heterogeneity. *Review of Economics and Statistics*, 90(3):389–405, August 2008.
33. Chi Chang, Thomas Jaki, Muhammad Saad Sadiq, Alena Kuhlemeier, Daniel Feaster, Natalie Cole, Andrea Lamont, Daniel Oberski, Yasin Desai, M. Lee Van Horn, and The Pooled Resource Open-Access ALS Clinical Trials Consortium. A permutation test for assessing the presence of individual differences in treatment effects. *Statistical Methods in Medical Research*, 30(11):2369–2381, 2021. PMID: 34570622.
34. Jared C. Foster, Bin Nan, Lei Shen, Niko Kaciroti, and Jeremy M. G. Taylor. Permutation Testing for Treatment–Covariate Interactions and Subgroup Identification. *Statistics in Biosciences*, 8(1):77–98, June 2016.
35. Jianshen Chen and Bryan Keller. Heterogeneous Subgroup Identification in Observational Studies. *Journal of Research on Educational Effectiveness*, 12(3):578–596, July 2019. Publisher: Routledge _eprint: https://doi.org/10.1080/19345747.2019.1615159.
36. ICH Expert Working Group. ICH harmonised tripartite guideline: Statistical principles for clinical trials E9. https://www.ich.org/page/efficacy-guidelines, 1998. Accessed: 2021-10-26.

Tables

Table 1. Simulation study details. We carried out 2000 simulations under every combination of $h(\boldsymbol{X}_i)$, $g(\boldsymbol{X}_i)$ and number of covariates p. Patient outcomes were generated through the model $Y_i = h(\boldsymbol{X}_i) + T_i g(\boldsymbol{X}_i) + \epsilon_i$ where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 4)$ and $n = 1000$.

	$p = 10$	$p = 20$	$p = 50$
Covariates			
Continuous	$(X_{i1}, \dots X_{i8})^T \overset{\text{iid}}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $\boldsymbol{\mu} \sim N(\boldsymbol{0}, 3\boldsymbol{I}_8)$ $\boldsymbol{\Sigma} = \text{AR}(1, \rho = 0.7)$	$(X_{i1}, \dots X_{i16})^T \overset{\text{iid}}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $\boldsymbol{\mu} \sim N(\boldsymbol{0}, 3\boldsymbol{I}_{16})$ $\boldsymbol{\Sigma} = \text{AR}(1, \rho = 0.7)$	$(X_{i1}, \dots X_{i40})^T \overset{\text{iid}}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $\boldsymbol{\mu} \sim N(\boldsymbol{0}, 3\boldsymbol{I}_{40})$ $\boldsymbol{\Sigma} = \text{AR}(1, \rho = 0.7)$
Binary	$X_{i9}, X_{i10} \overset{\text{iid}}{\sim} \text{Bernoulli}(0.7)$	$X_{i17}, \dots, X_{i20} \overset{\text{iid}}{\sim} \text{Bernoulli}(0.7)$	$X_{i41}, \dots, X_{i50} \overset{\text{iid}}{\sim} \text{Bernoulli}(0.7)$
Covariate Main Effects			
Linear	$h_1(\boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ $\beta_j = 1.25$ for $j = 1, \dots, 10$	$h_1(\boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ $\beta_j = \begin{cases} 1 & \text{for } j = 1, \dots, 10, 17, 18 \\ 0 & \text{otherwise} \end{cases}$	$h_1(\boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ $\beta_j = \begin{cases} 1 & \text{for } j = 1, \dots 12, 41, 42, 43 \\ 0 & \text{otherwise} \end{cases}$
Nonlinear	$h_2(\boldsymbol{X}_i) = X_{i1} + X_{i2} + X_{i9} + X_{i10} +$ $2\sum_{j=3}^{5}(X_{ij} - \mu_j)^2 +$ $2\sum_{j=6}^{8} I(X_{ij} > \mu_j) +$ $1/2(X_{i1} - \mu_1)(X_{i2} - \mu_2) -$ $(X_{i1} - \mu_1)X_{i9}$	$h_2(\boldsymbol{X}_i) = X_{i1} + X_{i2} + X_{i17} + X_{i18} +$ $5/4\sum_{j=3}^{6}(X_{ij} - \mu_j)^2 +$ $5/4\sum_{j=7}^{10} I(X_{ij} > \mu_j) +$ $1/2(X_{i1} - \mu_1)(X_{i2} - \mu_2) -$ $(X_{i1} - \mu_1)X_{i17}$	$h_2(\boldsymbol{X}_i) = X_{i1} + X_{i2} + X_{i41} + X_{i42} + X_{i43} +$ $5/4\sum_{j=3}^{7}(X_{ij} - \mu_j)^2 +$ $5/4\sum_{j=8}^{12} I(X_{ij} > \mu_j) +$ $1/2(X_{i1} - \mu_1)(X_{i2} - \mu_2) -$ $(X_{i1} - \mu_1)X_{i42}$
Conditional Average Treatment Effect			
Null	$g_0(\boldsymbol{X}_i) = 2$	$g_0(\boldsymbol{X}_i) = 2$	$g_0(\boldsymbol{X}_i) = 2$
Linear	$g_1(\boldsymbol{X}_i) = m + \boldsymbol{X}_i\boldsymbol{\beta}$ $\beta_j = \begin{cases} 1/2 & \text{for } j = 1, 9 \\ 0 & \text{otherwise} \end{cases}$ $m = 2 - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{\beta}$	$g_1(\boldsymbol{X}_i) = m + \boldsymbol{X}_i\boldsymbol{\beta}$ $\beta_j = \begin{cases} 1/2 & \text{for } j = 1, 2, 10, 17 \\ 0 & \text{otherwise} \end{cases}$ $m = 2 - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{\beta}$	$g_1(\boldsymbol{X}_i) = m + \boldsymbol{X}_i\boldsymbol{\beta}$ $\beta_j = \begin{cases} 1/2 & \text{for } j = 1, 2, 10, 41 \\ 0 & \text{otherwise} \end{cases}$ $m = 2 - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{\beta}$
Nonlinear	$g_2(\boldsymbol{X}_i) = m + \gamma(\boldsymbol{X}_i)$ $\gamma(\boldsymbol{X}_i) = 1/2X_{i9} + I(X_{i1} > \mu_1) +$ $1/8(X_{i1} - \mu_1)X_{i9}$ $m = 2 - \frac{1}{n}\sum_{i=1}^{n}\gamma(\boldsymbol{X}_i)$	$g_2(\boldsymbol{X}_i) = m + \gamma(\boldsymbol{X}_i)$ $\gamma(\boldsymbol{X}_i) = 1/2X_{i17} + I(X_{i1} > \mu_1) +$ $1/2I(X_{i2} > \mu_2) + 1/4I(X_{i10} > \mu_{10}) +$ $1/4(X_{i1} - \mu_1)X_{i17}$ $m = 2 - \frac{1}{n}\sum_{i=1}^{n}\gamma(\boldsymbol{X}_i)$	$g_2(\boldsymbol{X}_i) = m + \gamma(\boldsymbol{X}_i)$ $\gamma(\boldsymbol{X}_i) = 1/2X_{i41} + I(X_{i1} > \mu_1) +$ $1/2I(X_{i2} > \mu_2) + 1/4I(X_{i10} > \mu_{10}) +$ $1/4(X_{i1} - \mu_1)X_{i41} +$ $1/8(X_{i2} - \mu_2)(X_{i10} - \mu_{10})$ $m = 2 - \frac{1}{n}\sum_{i=1}^{n}\gamma(\boldsymbol{X}_i)$

Table 2. Proportion of simulations in which at least one covariate was included in the Step 2 model across all tested combinations of Step 1 (columns) and Step 2 (rows) methods. This corresponds to the Type I error rate for scenarios with homogeneous treatment effects and the power otherwise.

	$p = 10$		$p = 50$	
	RF	SL	RF	SL
Homogeneous Treatment Effects; Linear Main Effects				
LASSO($\alpha = 0.2$)	0.22*	0.21	0.19	0.18*
R.Tree($\alpha = 0.2$)	0.21	0.21	0.20	0.21
C.Tree($\alpha = 0.2$)	0.22*	0.22*	0.19	0.20
LASSO($\alpha = 0.05$)	0.05	0.06	0.05	0.05
R.Tree($\alpha = 0.05$)	0.05	0.06	0.06	0.06
C.Tree($\alpha = 0.05$)	0.05	0.06	0.05	0.04
Homogeneous Treatment Effects; Nonlinear Main Effects				
LASSO($\alpha = 0.2$)	0.19	0.22	0.21	0.20
R.Tree($\alpha = 0.2$)	0.19	0.20	0.21	0.20
C.Tree($\alpha = 0.2$)	0.19	0.21	0.21	0.19
LASSO($\alpha = 0.05$)	0.06	0.05	0.06*	0.06
R.Tree($\alpha = 0.05$)	0.06	0.06	0.06	0.06
C.Tree($\alpha = 0.05$)	0.06	0.06	0.06*	0.05
Linear Treatment Effects; Linear Main Effects				
LASSO($\alpha = 0.2$)	0.45	0.55	0.76	0.92
R.Tree($\alpha = 0.2$)	0.39	0.34	0.68	0.71
C.Tree($\alpha = 0.2$)	0.44	0.39	0.74	0.79
LASSO($\alpha = 0.05$)	0.20	0.30	0.48	0.77
R.Tree($\alpha = 0.05$)	0.15	0.09	0.40	0.33
C.Tree($\alpha = 0.05$)	0.19	0.15	0.48	0.47
Linear Treatment Effects; Nonlinear Main Effects				
LASSO($\alpha = 0.2$)	0.32	0.43	0.66	0.83
R.Tree($\alpha = 0.2$)	0.22	0.25	0.46	0.59
C.Tree($\alpha = 0.2$)	0.31	0.41	0.64	0.80
LASSO($\alpha = 0.05$)	0.09	0.19	0.34	0.64
R.Tree($\alpha = 0.05$)	0.06	0.08	0.18	0.29
C.Tree($\alpha = 0.05$)	0.09	0.16	0.32	0.57
Nonlinear Treatment Effects; Linear Main Effects				
LASSO($\alpha = 0.2$)	0.57	0.71	0.57	0.77
R.Tree($\alpha = 0.2$)	0.52	0.46	0.54	0.61
C.Tree($\alpha = 0.2$)	0.55	0.42	0.56	0.60
LASSO($\alpha = 0.05$)	0.30	0.44	0.31	0.52
R.Tree($\alpha = 0.05$)	0.28	0.16	0.28	0.29
C.Tree($\alpha = 0.05$)	0.27	0.15	0.30	0.29
Nonlinear Treatment Effects; Nonlinear Main Effects				
LASSO($\alpha = 0.2$)	0.39	0.65	0.46	0.68
R.Tree($\alpha = 0.2$)	0.25	0.36	0.34	0.45
C.Tree($\alpha = 0.2$)	0.38	0.59	0.45	0.64
LASSO($\alpha = 0.05$)	0.12	0.36	0.19	0.41
R.Tree($\alpha = 0.05$)	0.06	0.11	0.11	0.18
C.Tree($\alpha = 0.05$)	0.12	0.28	0.19	0.35

* For simulations with homogeneous treatment effects, 95% CI does not include α

Abbreviations: RF: random forest; SL: super learner; R.Tree: regression tree; C.Tree: conditional inference tree

Table 3. Variables determined to contribute to treatment effect heterogeneity for a given comparison of treatments on change in cigarettes per day. Covariates which had nonzero effects in the Step 2 Virtual Twins model fit for subjects' estimated individual treatment effects with a given Step 1 model are marked with a ✓. Variables with no estimated effect for all presented models are omitted.

		Immediate vs. Gradual					Immediate vs. Control	
Step 1	Step 2	TNE	CEMA	CPD	NNAL	Total	Age	Total
RF	LASSO($\alpha = 0.2$)	✓		✓		2		0
	R.Tree($\alpha = 0.2$)	✓				1	✓	1
	C.Tree($\alpha = 0.2$)	✓				1		0
	LASSO($\alpha = 0.05$)	✓		✓		2		0
	R.Tree($\alpha = 0.05$)	✓				1		0
	C.Tree($\alpha = 0.05$)	✓				1		0
SL	LASSO($\alpha = 0.2$)	✓	✓		✓	3		0
	R.Tree($\alpha = 0.2$)		✓			1		0
	C.Tree($\alpha = 0.2$)	✓				1		0
	LASSO($\alpha = 0.05$)	✓	✓			2		0
	R.Tree($\alpha = 0.05$)		✓			1		0
	C.Tree($\alpha = 0.05$)					0		0

Abbreviations: RF: random forest; SL: super learner; R.Tree: regression tree; C.Tree: conditional inference tree; TNE: total nicotine equivalents; CEMA: cyanoethyl mercapturic acid; CPD: cigarettes per day; NNAL: 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol

Supplemental Tables

Table S1. Simulations' average R^2 for the response both when assuming no covariate by treatment interaction (R^2_0) and when using the including the interaction (R^2_A) calculated using the true data generative model. Cohen's f^2, given by $(R^2_A - R^2_0)/(1 - R^2_A)$, measures the additional effect of the treatment by covariate interaction. Values are R^2_0 / R^2_A/ f^2.

		Treatment Effect		
Main Effects	p	Homogeneous	Linear	Nonlinear
Linear	10	0.71 / – / –	0.72 / 0.72 / 0.01	0.72 / 0.73 / 0.02
	20	0.68 / – / –	0.70 / 0.71 / 0.03	0.69 / 0.70 / 0.03
	50	0.73 / – / –	0.74 / 0.75 / 0.03	0.73 / 0.74 / 0.03
Nonlinear	10	0.75 / – / –	0.75 / 0.75 / 0.01	0.75 / 0.75 / 0.02
	20	0.64 / – / –	0.64 / 0.66 / 0.03	0.64 / 0.65 / 0.03
	50	0.69 / – / –	0.69 / 0.70 / 0.03	0.69 / 0.70 / 0.03

Table S2. Considered Step 1 and Step 2 methods for Virtual Twins. Step 1 models were fit on the subset of the data where $T_i = 0$ and its complement where $T_i = 1$ to estimate $\mathrm{E}(Y_i|T_i = 0, \boldsymbol{X}_i)$ and $\mathrm{E}(Y_i|T_i = 1, \boldsymbol{X}_i)$. A Step 2 model was then fit for the estimate of $\mathrm{E}(Y_i|T_i = 1, \boldsymbol{X}_i) - \mathrm{E}(Y_i|T_i = 1, \boldsymbol{X}_i)$ obtained by plugging in both Step 1 models' estimates.

Method	R Package::Function()	Details
Step 1		
Random Forest	`randomForestSRC::rfsrc()`	Fit with 1000 trees
SuperLearner	`SuperLearner::Superlearner()`	Consisted of a lasso (10-fold CV), random forest (250 trees), and MARS (default settings) Larger SuperLearner weights tuned via 3-fold CV.
Step 2		
Lasso	`glmnet::glmnet()`	Penalty parameter `lambda`
Regression Tree	`rpart::rpart()`	Penalty parameter `cp`
Conditional Inference Tree	`party::ctree()`	Penalty parameter `mincriterion` passed to `ctree.control` with `(testtype = "Teststatistic")`

Table S3. Proportion of simulations in which at least one covariate was included in the Step 2 model across all tested combinations of Step 1 (columns) and Step 2 (rows) methods. This corresponds to the Type I error rate for scenarios with homogeneous treatment effects and the power otherwise.

	$p = 20$	
	RF	SL
Homogeneous Treatment Effects; Linear Main Effects		
LASSO(α = 0.2)	0.20	0.21
R.Tree(α = 0.2)	0.21	0.21
C.Tree(α = 0.2)	0.20	0.20
LASSO(α = 0.05)	0.06	0.05
R.Tree(α = 0.05)	0.05	0.06*
C.Tree(α = 0.05)	0.06*	0.05
Homogeneous Treatment Effects; Nonlinear Main Effects		
LASSO(α = 0.2)	0.20	0.20
R.Tree(α = 0.2)	0.20	0.20
C.Tree(α = 0.2)	0.20	0.19
LASSO(α = 0.05)	0.05	0.06
R.Tree(α = 0.05)	0.05	0.05
C.Tree(α = 0.05)	0.05	0.05
Linear Treatment Effects; Linear Main Effects		
LASSO(α = 0.2)	0.85	0.93
R.Tree(α = 0.2)	0.76	0.66
C.Tree(α = 0.2)	0.84	0.76
LASSO(α = 0.05)	0.62	0.80
R.Tree(α = 0.05)	0.48	0.28
C.Tree(α = 0.05)	0.60	0.43
Linear Treatment Effects; Nonlinear Main Effects		
LASSO(α = 0.2)	0.78	0.90
R.Tree(α = 0.2)	0.58	0.70
C.Tree(α = 0.2)	0.77	0.86
LASSO(α = 0.05)	0.48	0.70
R.Tree(α = 0.05)	0.27	0.38
C.Tree(α = 0.05)	0.46	0.66
Nonlinear Treatment Effects; Linear Main Effects		
LASSO(α = 0.2)	0.69	0.80
R.Tree(α = 0.2)	0.66	0.62
C.Tree(α = 0.2)	0.68	0.59
LASSO(α = 0.05)	0.42	0.56
R.Tree(α = 0.05)	0.39	0.29
C.Tree(α = 0.05)	0.40	0.29
Nonlinear Treatment Effects; Nonlinear Main Effects		
LASSO(α = 0.2)	0.58	0.77
R.Tree(α = 0.2)	0.45	0.57
C.Tree(α = 0.2)	0.56	0.74
LASSO(α = 0.05)	0.29	0.51
R.Tree(α = 0.05)	0.18	0.28
C.Tree(α = 0.05)	0.27	0.45

* For simulations with homogeneous treatment effects, 95% CI does not include α.

Abbreviations: RF: random forest; SL: super learner; R.Tree: regression tree; C.Tree: conditional inference tree

Table S4. Proportion of simulations in which at least one covariate was included in the Step 2 model across simulations with R-squared values around 20%. This corresponds to the Type I error rate for scenarios with homogeneous treatment effects and the power otherwise.

	$p = 10$		$p = 20$		$p = 50$	
	RF	SL	RF	SL	RF	SL
Homogeneous Treatment Effects; Linear Main Effects						
LASSO($\alpha = 0.2$)	0.21	0.21	0.20	0.19	0.20	0.19
R.Tree($\alpha = 0.2$)	0.20	0.21	0.19	0.18*	0.20	0.21
C.Tree($\alpha = 0.2$)	0.22	0.22*	0.20	0.18*	0.19	0.21
LASSO($\alpha = 0.05$)	0.06*	0.06	0.06*	0.05	0.05	0.05
R.Tree($\alpha = 0.05$)	0.06	0.05	0.06	0.05	0.05	0.06
C.Tree($\alpha = 0.05$)	0.07*	0.06	0.06	0.05	0.05	0.06
Homogeneous Treatment Effects; Nonlinear Main Effects						
LASSO($\alpha = 0.2$)	0.21	0.21	0.20	0.19	0.19	0.20
R.Tree($\alpha = 0.2$)	0.20	0.21	0.19	0.19	0.18	0.20
C.Tree($\alpha = 0.2$)	0.22	0.20	0.20	0.20	0.18	0.20
LASSO($\alpha = 0.05$)	0.06	0.06	0.05	0.06	0.06	0.06
R.Tree($\alpha = 0.05$)	0.06	0.06	0.05	0.05	0.06	0.05
C.Tree($\alpha = 0.05$)	0.06	0.05	0.05	0.05	0.06	0.06*
Linear Treatment Effects; Linear Main Effects						
LASSO($\alpha = 0.2$)	0.25	0.25	0.34	0.34	0.31	0.34
R.Tree($\alpha = 0.2$)	0.23	0.23	0.31	0.28	0.30	0.30
C.Tree($\alpha = 0.2$)	0.24	0.23	0.35	0.29	0.31	0.31
LASSO($\alpha = 0.05$)	0.07	0.08	0.13	0.14	0.11	0.13
R.Tree($\alpha = 0.05$)	0.07	0.06	0.12	0.09	0.11	0.09
C.Tree($\alpha = 0.05$)	0.07	0.06	0.13	0.09	0.11	0.09
Linear Treatment Effects; Nonlinear Main Effects						
LASSO($\alpha = 0.2$)	0.24	0.25	0.33	0.34	0.29	0.31
R.Tree($\alpha = 0.2$)	0.22	0.21	0.30	0.26	0.27	0.27
C.Tree($\alpha = 0.2$)	0.24	0.24	0.33	0.33	0.29	0.30
LASSO($\alpha = 0.05$)	0.07	0.07	0.11	0.12	0.10	0.12
R.Tree($\alpha = 0.05$)	0.06	0.07	0.10	0.08	0.09	0.08
C.Tree($\alpha = 0.05$)	0.07	0.07	0.11	0.11	0.10	0.11
Nonlinear Treatment Effects; Linear Main Effects						
LASSO($\alpha = 0.2$)	0.26	0.29	0.27	0.28	0.25	0.27
R.Tree($\alpha = 0.2$)	0.26	0.24	0.26	0.23	0.25	0.25
C.Tree($\alpha = 0.2$)	0.26	0.23	0.28	0.24	0.25	0.26
LASSO($\alpha = 0.05$)	0.09	0.09	0.10	0.11	0.08	0.09
R.Tree($\alpha = 0.05$)	0.07	0.06	0.09	0.07	0.08	0.08
C.Tree($\alpha = 0.05$)	0.08	0.07	0.09	0.06	0.08	0.07
Nonlinear Treatment Effects; Nonlinear Main Effects						
LASSO($\alpha = 0.2$)	0.25	0.26	0.27	0.28	0.22	0.24
R.Tree($\alpha = 0.2$)	0.24	0.21	0.25	0.23	0.23	0.23
C.Tree($\alpha = 0.2$)	0.25	0.24	0.26	0.26	0.22	0.23
LASSO($\alpha = 0.05$)	0.07	0.08	0.08	0.09	0.08	0.08
R.Tree($\alpha = 0.05$)	0.07	0.06	0.08	0.06	0.07	0.06
C.Tree($\alpha = 0.05$)	0.07	0.07	0.08	0.08	0.08	0.08

* For simulations with homogeneous treatment effects, 95% CI does not include α

Table S5. Mean sensitivity/specificity across all tested combinations of Step 1 (columns) and Step 2 (rows) methods.

	$p = 10$		$p = 20$		$p = 50$	
	RF	SL	RF	SL	RF	SL
Homogeneous Treatment Effects; Linear Main Effects						
LASSO	– / 0.13	– / 0.01	– / 0.25	– / 0.05	– / 0.48	– / 0.20
R.Tree	– / 0.48	– / 0.65	– / 0.75	– / 0.79	– / 0.91	– / 0.91
C.Tree	– / 0.64	– / 0.65	– / 0.83	– / 0.80	– / 0.94	– / 0.91
LASSO($\alpha = 0.2$)	– / 0.97	– / 0.98	– / 0.99	– / 0.99	– / 0.99	– / 1.00
R.Tree($\alpha = 0.2$)	– / 0.98	– / 0.98	– / 0.99	– / 0.99	– / 1.00	– / 1.00
C.Tree($\alpha = 0.2$)	– / 0.98	– / 0.98	– / 0.99	– / 0.99	– / 1.00	– / 1.00
LASSO($\alpha = 0.05$)	– / 0.99	– / 0.99	– / 1.00	– / 1.00	– / 1.00	– / 1.00
R.Tree($\alpha = 0.05$)	– / 1.00	– / 0.99	– / 1.00	– / 1.00	– / 1.00	– / 1.00
C.Tree($\alpha = 0.05$)	– / 0.99	– / 0.99	– / 1.00	– / 1.00	– / 1.00	– / 1.00
Homogeneous Treatment Effects; Nonlinear Main Effects						
LASSO	– / 0.16	– / 0.09	– / 0.23	– / 0.14	– / 0.45	– / 0.29
R.Tree	– / 0.60	– / 0.61	– / 0.79	– / 0.78	– / 0.93	– / 0.92
C.Tree	– / 0.68	– / 0.63	– / 0.85	– / 0.81	– / 0.95	– / 0.94
LASSO($\alpha = 0.2$)	– / 0.97	– / 0.97	– / 0.99	– / 0.99	– / 0.99	– / 0.99
R.Tree($\alpha = 0.2$)	– / 0.98	– / 0.98	– / 0.99	– / 0.99	– / 1.00	– / 1.00
C.Tree($\alpha = 0.2$)	– / 0.98	– / 0.98	– / 0.99	– / 0.99	– / 1.00	– / 1.00
LASSO($\alpha = 0.05$)	– / 0.99	– / 0.99	– / 1.00	– / 1.00	– / 1.00	– / 1.00
R.Tree($\alpha = 0.05$)	– / 0.99	– / 0.99	– / 1.00	– / 1.00	– / 1.00	– / 1.00
C.Tree($\alpha = 0.05$)	– / 0.99	– / 0.99	– / 1.00	– / 1.00	– / 1.00	– / 1.00
Linear Treatment Effects; Linear Main Effects						
LASSO	0.92 / 0.12	1.00 / 0.01	0.89 / 0.24	1.00 / 0.07	0.79 / 0.48	0.92 / 0.23
R.Tree	0.39 / 0.46	0.62 / 0.75	0.47 / 0.80	0.50 / 0.91	0.43 / 0.93	0.47 / 0.96
C.Tree	0.42 / 0.63	0.61 / 0.75	0.46 / 0.86	0.51 / 0.92	0.41 / 0.96	0.47 / 0.96
LASSO($\alpha = 0.2$)	0.14 / 0.96	0.24 / 0.96	0.30 / 0.98	0.40 / 0.97	0.25 / 0.99	0.39 / 0.99
R.Tree($\alpha = 0.2$)	0.08 / 0.97	0.13 / 0.99	0.16 / 0.99	0.15 / 1.00	0.13 / 1.00	0.16 / 1.00
C.Tree($\alpha = 0.2$)	0.12 / 0.97	0.14 / 0.99	0.20 / 0.99	0.18 / 1.00	0.16 / 1.00	0.18 / 1.00
LASSO($\alpha = 0.05$)	0.06 / 0.98	0.13 / 0.99	0.19 / 0.99	0.29 / 0.99	0.13 / 1.00	0.28 / 0.99
R.Tree($\alpha = 0.05$)	0.04 / 0.99	0.04 / 1.00	0.10 / 1.00	0.07 / 1.00	0.08 / 1.00	0.08 / 1.00
C.Tree($\alpha = 0.05$)	0.06 / 0.99	0.06 / 1.00	0.13 / 1.00	0.10 / 1.00	0.10 / 1.00	0.11 / 1.00
Linear Treatment Effects; Nonlinear Main Effects						
LASSO	0.92 / 0.15	0.94 / 0.07	0.90 / 0.23	0.96 / 0.16	0.78 / 0.45	0.86 / 0.31
R.Tree	0.26 / 0.54	0.42 / 0.59	0.42 / 0.81	0.47 / 0.83	0.33 / 0.93	0.38 / 0.93
C.Tree	0.39 / 0.66	0.53 / 0.66	0.45 / 0.86	0.52 / 0.88	0.34 / 0.96	0.41 / 0.95
LASSO($\alpha = 0.2$)	0.06 / 0.96	0.18 / 0.97	0.27 / 0.97	0.36 / 0.97	0.19 / 0.99	0.30 / 0.99
R.Tree($\alpha = 0.2$)	0.00 / 0.97	0.04 / 0.98	0.10 / 0.99	0.14 / 0.99	0.07 / 1.00	0.11 / 1.00
C.Tree($\alpha = 0.2$)	0.04 / 0.97	0.14 / 0.98	0.17 / 0.99	0.22 / 0.99	0.13 / 0.99	0.19 / 1.00
LASSO($\alpha = 0.05$)	0.01 / 0.99	0.07 / 0.99	0.14 / 0.99	0.24 / 0.99	0.08 / 1.00	0.20 / 1.00
R.Tree($\alpha = 0.05$)	0.00 / 0.99	0.01 / 0.99	0.05 / 1.00	0.08 / 1.00	0.03 / 1.00	0.05 / 1.00
C.Tree($\alpha = 0.05$)	0.01 / 0.99	0.05 / 0.99	0.10 / 0.99	0.16 / 1.00	0.06 / 1.00	0.13 / 1.00
Nonlinear Treatment Effects; Linear Main Effects						
LASSO	0.97 / 0.12	1.00 / 0.01	0.85 / 0.24	0.99 / 0.06	0.74 / 0.47	0.92 / 0.21
R.Tree	0.49 / 0.48	0.79 / 0.79	0.41 / 0.78	0.38 / 0.87	0.36 / 0.93	0.33 / 0.94
C.Tree	0.56 / 0.64	0.78 / 0.79	0.37 / 0.85	0.38 / 0.87	0.32 / 0.95	0.32 / 0.94
LASSO($\alpha = 0.2$)	0.22 / 0.95	0.41 / 0.96	0.21 / 0.98	0.22 / 0.97	0.16 / 0.99	0.20 / 0.99
R.Tree($\alpha = 0.2$)	0.16 / 0.97	0.21 / 0.99	0.14 / 0.99	0.10 / 0.99	0.10 / 1.00	0.09 / 0.99
C.Tree($\alpha = 0.2$)	0.19 / 0.97	0.18 / 0.99	0.15 / 0.99	0.12 / 0.99	0.11 / 1.00	0.12 / 1.00
LASSO($\alpha = 0.05$)	0.12 / 0.98	0.23 / 0.99	0.12 / 0.99	0.14 / 0.99	0.08 / 1.00	0.13 / 1.00
R.Tree($\alpha = 0.05$)	0.10 / 0.99	0.07 / 1.00	0.09 / 1.00	0.05 / 0.99	0.06 / 1.00	0.04 / 1.00
C.Tree($\alpha = 0.05$)	0.10 / 0.99	0.07 / 1.00	0.09 / 1.00	0.06 / 1.00	0.06 / 1.00	0.06 / 1.00
Nonlinear Treatment Effects; Nonlinear Main Effects						
LASSO	0.96 / 0.14	0.97 / 0.07	0.85 / 0.22	0.93 / 0.15	0.73 / 0.44	0.84 / 0.30
R.Tree	0.38 / 0.55	0.61 / 0.63	0.34 / 0.79	0.36 / 0.80	0.25 / 0.93	0.27 / 0.93
C.Tree	0.55 / 0.67	0.72 / 0.70	0.33 / 0.85	0.41 / 0.85	0.24 / 0.95	0.28 / 0.95
LASSO($\alpha = 0.2$)	0.11 / 0.95	0.36 / 0.97	0.15 / 0.98	0.21 / 0.97	0.11 / 0.99	0.16 / 0.99
R.Tree($\alpha = 0.2$)	0.03 / 0.98	0.10 / 0.98	0.07 / 0.99	0.09 / 0.99	0.04 / 1.00	0.07 / 1.00
C.Tree($\alpha = 0.2$)	0.09 / 0.97	0.26 / 0.99	0.11 / 0.99	0.14 / 0.98	0.07 / 1.00	0.11 / 0.99
LASSO($\alpha = 0.05$)	0.03 / 0.99	0.18 / 0.99	0.07 / 0.99	0.12 / 0.99	0.04 / 1.00	0.10 / 1.00
R.Tree($\alpha = 0.05$)	0.00 / 0.99	0.04 / 1.00	0.03 / 1.00	0.05 / 1.00	0.01 / 1.00	0.03 / 1.00
C.Tree($\alpha = 0.05$)	0.02 / 0.99	0.12 / 0.99	0.05 / 1.00	0.09 / 0.99	0.03 / 1.00	0.07 / 1.00

Abbreviations: RF: random forest; SL: super learner; R.Tree: regression tree; C.Tree: conditional inference tree

Table S6. Mean squared error across estimating subjects' conditional average treatment effects across across all tested combinations of Step 1 (columns) and Step 2 (rows) methods

	$p = 10$		$p = 20$		$p = 50$	
	RF	SL	RF	SL	RF	SL
Homogeneous Treatment Effects; Linear Main Effects						
LASSO	0.48	0.70	0.53	1.03	0.66	1.41
R.Tree	1.06	0.44	0.94	0.53	1.03	0.59
C.Tree	0.72	0.43	0.67	0.52	0.75	0.58
LASSO(α = 0.2)	0.08	0.07	0.08	0.07	0.09	0.07
R.Tree(α = 0.2)	0.18	0.12	0.18	0.14	0.21	0.14
C.Tree(α = 0.2)	0.17	0.12	0.17	0.13	0.19	0.13
LASSO(α = 0.05)	0.07	0.07	0.08	0.07	0.09	0.06
R.Tree(α = 0.05)	0.10	0.09	0.11	0.09	0.13	0.09
C.Tree(α = 0.05)	0.10	0.08	0.11	0.09	0.12	0.08
Homogeneous Treatment Effects; Nonlinear Main Effects						
LASSO	0.59	1.02	0.62	1.05	0.81	1.52
R.Tree	1.44	1.68	1.14	1.34	1.34	1.54
C.Tree	1.01	1.19	0.79	0.93	0.91	1.01
LASSO(α = 0.2)	0.11	0.08	0.10	0.08	0.12	0.08
R.Tree(α = 0.2)	0.28	0.26	0.23	0.22	0.30	0.26
C.Tree(α = 0.2)	0.24	0.21	0.20	0.18	0.26	0.21
LASSO(α = 0.05)	0.10	0.08	0.09	0.08	0.11	0.08
R.Tree(α = 0.05)	0.17	0.14	0.14	0.12	0.17	0.14
C.Tree(α = 0.05)	0.15	0.12	0.12	0.10	0.17	0.12
Linear Treatment Effects; Linear Main Effects						
LASSO	0.54	0.70	0.71	1.04	0.89	1.42
R.Tree	1.15	0.49	1.37	0.82	1.58	0.87
C.Tree	0.86	0.49	1.17	0.80	1.34	0.85
LASSO(α = 0.2)	0.34	0.31	0.91	0.77	0.98	0.78
R.Tree(α = 0.2)	0.47	0.38	1.06	1.00	1.16	0.99
C.Tree(α = 0.2)	0.45	0.38	1.02	0.96	1.11	0.95
LASSO(α = 0.05)	0.36	0.34	1.04	0.91	1.09	0.93
R.Tree(α = 0.05)	0.42	0.38	1.11	1.11	1.19	1.09
C.Tree(α = 0.05)	0.41	0.37	1.07	1.06	1.15	1.04
Linear Treatment Effects; Nonlinear Main Effects						
LASSO	0.66	1.16	0.78	1.09	1.05	1.60
R.Tree	1.61	1.80	1.65	1.68	2.02	2.00
C.Tree	1.14	1.28	1.32	1.32	1.57	1.52
LASSO(α = 0.2)	0.40	0.35	0.97	0.84	1.07	0.90
R.Tree(α = 0.2)	0.59	0.56	1.23	1.18	1.36	1.28
C.Tree(α = 0.2)	0.55	0.50	1.12	1.07	1.27	1.16
LASSO(α = 0.05)	0.40	0.37	1.09	0.98	1.17	1.02
R.Tree(α = 0.05)	0.47	0.45	1.23	1.19	1.30	1.24
C.Tree(α = 0.05)	0.46	0.44	1.16	1.10	1.26	1.16
Nonlinear Treatment Effects; Linear Main Effects						
LASSO	0.70	0.80	0.85	1.17	1.06	1.57
R.Tree	1.31	0.56	1.32	0.72	1.53	0.81
C.Tree	1.01	0.57	1.10	0.72	1.27	0.81
LASSO(α = 0.2)	0.65	0.59	0.82	0.74	0.88	0.77
R.Tree(α = 0.2)	0.71	0.63	0.84	0.77	0.97	0.80
C.Tree(α = 0.2)	0.70	0.64	0.83	0.76	0.95	0.78
LASSO(α = 0.05)	0.68	0.65	0.88	0.83	0.93	0.85
R.Tree(α = 0.05)	0.71	0.69	0.88	0.86	0.97	0.88
C.Tree(α = 0.05)	0.70	0.68	0.87	0.84	0.96	0.86
Nonlinear Treatment Effects; Nonlinear Main Effects						
LASSO	0.82	1.30	0.90	1.20	1.20	1.73
R.Tree	1.80	1.81	1.57	1.57	1.94	1.89
C.Tree	1.29	1.29	1.21	1.16	1.49	1.34
LASSO(α = 0.2)	0.72	0.64	0.86	0.77	0.94	0.84
R.Tree(α = 0.2)	0.91	0.86	1.02	0.94	1.15	1.05
C.Tree(α = 0.2)	0.84	0.75	0.93	0.85	1.09	0.93
LASSO(α = 0.05)	0.74	0.69	0.92	0.85	0.98	0.90
R.Tree(α = 0.05)	0.80	0.77	0.99	0.94	1.06	1.00
C.Tree(α = 0.05)	0.79	0.74	0.96	0.88	1.05	0.95

Abbreviations: RF: random forest; SL: super learner; R.Tree: regression tree; C.Tree: conditional inference tree

Table S7. Demographic summary of the study population at baseline. Values are mean (SD) for numeric covariates and N (%) for categorical covariates.

Covariate	Overall (N=921)	Control (N=198)	Gradual (N=383)	Immediate (N=340)
Age	46.0 (13.3)	45.5 (13.3)	45.4 (13.0)	47.0 (13.7)
Gender				
Female	421 (45.7%)	90 (45.5%)	173 (45.2%)	158 (46.5%)
Male	500 (54.3%)	108 (54.5%)	210 (54.8%)	182 (53.5%)
Race				
Black	279 (30.3%)	61 (30.8%)	115 (30.0%)	103 (30.3%)
Other	67 (7.3%)	16 (8.1%)	29 (7.6%)	22 (6.5%)
White	575 (62.4%)	121 (61.1%)	239 (62.4%)	215 (63.2%)
Education				
Less than High School	75 (8.1%)	18 (9.1%)	26 (6.8%)	31 (9.1%)
High School	294 (31.9%)	65 (32.8%)	124 (32.4%)	105 (30.9%)
More than High School	552 (59.9%)	115 (58.1%)	233 (60.8%)	204 (60.0%)
CES				
Satisfaction	4.7 (1.3)	4.7 (1.2)	4.7 (1.4)	4.6 (1.3)
Psych Reward	3.0 (1.4)	3.0 (1.3)	3.1 (1.4)	3.0 (1.4)
Aversion	1.3 (0.57)	1.3 (0.6)	1.2 (0.49)	1.3 (0.64)
Enjoy Sensation	3.4 (1.7)	3.5 (1.7)	3.4 (1.8)	3.4 (1.6)
Reduce Craving	4.6 (1.9)	4.5 (1.8)	4.5 (2.0)	4.6 (1.8)
WISDM				
Affiliative	2.5 (1.7)	2.5 (1.7)	2.5 (1.8)	2.4 (1.7)
Automaticity	3.8 (1.8)	3.8 (1.7)	3.8 (1.9)	3.8 (1.8)
Loss of Control	3.8 (1.6)	3.9 (1.6)	3.8 (1.6)	3.8 (1.6)
Cognitive	2.9 (1.7)	2.9 (1.8)	2.9 (1.8)	2.8 (1.6)
Craving	4.3 (1.6)	4.4 (1.6)	4.3 (1.7)	4.3 (1.6)
Cue	3.8 (1.6)	3.8 (1.5)	3.7 (1.7)	3.8 (1.6)
Social	3.7 (2.0)	3.8 (2.1)	3.8 (2.0)	3.6 (2.0)
Taste	4.4 (1.8)	4.5 (1.7)	4.3 (1.8)	4.4 (1.7)
Tolerance	4.6 (1.6)	4.6 (1.6)	4.6 (1.7)	4.6 (1.6)
Weight	2.0 (1.4)	2.2 (1.5)	2.0 (1.4)	2.0 (1.4)
Affective	3.3 (1.7)	3.4 (1.7)	3.2 (1.7)	3.2 (1.7)
QSU				
Factor 1	17.7 (9.2)	17.7 (9.4)	17.9 (9.3)	17.5 (9.0)
Factor 2	9.8 (6.4)	9.9 (6.3)	9.7 (6.4)	9.8 (6.3)
PANAS				
Positive	31.6 (8.8)	31.1 (8.7)	32.1 (9.3)	31.5 (8.2)
Negative	15.0 (5.3)	15.6 (6.1)	14.8 (5.1)	14.9 (5.1)
FTND	4.2 (1.7)	4.0 (1.8)	4.2 (1.7)	4.3 (1.7)
CESD	6.0 (5.6)	6.4 (5.7)	6.0 (5.9)	5.8 (5.2)
SMAST	2.9 (2.1)	3.2 (2.3)	2.8 (2.1)	2.9 (2.0)
DAST	1.7 (2.0)	1.7 (2.0)	1.7 (1.9)	1.8 (2.0)
CO	19.1 (9.4)	19.6 (9.7)	18.9 (9.6)	19.0 (9.1)
PSS	4.2 (2.9)	4.3 (3.0)	4.1 (2.9)	4.2 (2.7)
CPD	17.1 (8.6)	17.0 (8.4)	16.8 (8.2)	17.5 (8.9)
TNE	68.8 (38.0)	70.0 (39.2)	68.0 (38.8)	69.0 (36.4)
NNAL	1.8 (1.8)	2.1 (2.5)	1.7 (1.4)	1.9 (1.7)
PHET	3.0 (3.0)	3.1 (2.5)	2.7 (2.1)	3.4 (3.8)
CEMA	0.88 (0.73)	0.92 (0.83)	0.86 (0.7)	0.89 (0.69)
PGEM	72.5 (152.8)	66.3 (61.4)	80.0 (216.6)	67.7 (90.3)
ISO	1.3 (0.76)	1.3 (0.86)	1.3 (0.73)	1.3 (0.73)
Weight	86.9 (22.5)	82.2 (19.9)	88.9 (23.2)	87.5 (22.9)

Table S8. Number of variables determined to contribute to treatment effect heterogeneity for a given comaprison of treatents on change in cigarettes per day. The number of covariates which had a nonzero effect in the Step 2 Virtual Twins model (rows) fit for subjects' estimated conditional average treamtent effects with a given Step 1 model (columns) is reported.

	Immediate vs. Gradual		Immediate vs. Control	
	RF	SL	RF	SL
LASSO	28	33	27	23
R.Tree	6	5	3	4
C.Tree	5	6	3	3
LASSO($\alpha = 0.2$)	2	3	0	0
R.Tree($\alpha = 0.2$)	1	1	1	0
C.Tree($\alpha = 0.2$)	1	1	0	0
LASSO($\alpha = 0.05$)	2	2	0	0
R.Tree($\alpha = 0.05$)	1	1	0	0
C.Tree($\alpha = 0.05$)	1	0	0	0

Abbreviations: RF: random forest; SL: super learner; R.Tree: regression tree; C.Tree: conditional inference tree