Published in final edited form as:

J R Stat Soc Ser C Appl Stat. 2019 November ; 68(5): 1371–1391. doi:10.1111/rssc.12366.

Case-only trees and random forests for exploring genotype-specific treatment effects in randomized clinical trials with dichotomous endpoints

James Y. Dai,

Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, USA.

Michael LeBlanc

Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.

Summary.

Discovering gene-treatment interactions in clinical trials is of rising interest in the era of precision medicine. Nonparametric statistical learning methods such as trees and random forests are useful tools for building prediction rules. In this article, we introduce trees and random forests to the recently proposed case-only approach for discovering gene-treatment interactions and estimating marker-specific treatment effects for a dichotomous trial endpoints. The motivational example is a case-control genetic association study in the Prostate Cancer Prevention Trial (PCPT), which tested the hypothesis whether finasteride can prevent prostate cancer. We compare this novel approach to the interaction tree method previously proposed. Because of the modeling simplicity – directly targeting at interaction – and the statistical efficiency of the case-only approach, case-only trees and random forests yield more accurate prediction of heterogeneous treatment effects and better measure of variable importance, relative to the interaction tree method which uses data from both cases and controls. Application of the proposed case-only trees and random forests to the PCPT study yielded a discovery of genotypes that may influence the prevention effect of finasteride.

Keywords

case-only methods; gene-treatment interaction; individualized treatment effects; precision medicine; subgroups

1. Introduction

In clinical trials, the primary goal is to evaluate the efficacy of an investigational treatment, typically relative to a standard treatment. A key component of secondary correlative analyses is to determine whether there are biomarker-based subgroups with heterogeneous treatment effects. For example, pharmacogenetics studies aim to identify genotypes that influence

jdai@fredhutch.org.

Conflict of interest

The authors declare no potential conflict of interests.

treatment effects, also known as gene-treatment interactions (Evans and McLeod, 2003; Weinshilboum and Wang, 2004). Identification of such predictive markers of differential treatment effects may guide individualized treatment selection, a critical step for precision medicine. The advent of high-throughput genomic assays such as microarrays facilitates biomarker discoveries, yet at the same time drastically increases the chance of false discoveries due to the high-dimensional interrogation. As noted extensively in the clinical literature, agnostic subgroup analyses for gene-treatment interactions without a prior hypothesis are highly exploratory with limited power (Assmann et al., 2000; Lagakos, 2006).

The motivational example of this article is the Prostate Cancer Prevention Trial (PCPT). This large phase-3 trial tested the hypothesis whether finasteride, an inhibitor of the enzyme 5 α -reductase for androgen metabolism, can prevent prostate cancer (Thompson et al., 2003). Between 1994 and 1997, 18,882 men were randomized to finasteride or placebo and followed for 7 years. One unique feature of this trial is that the primary endpoint was the prevalence of prostate cancer in the seven-year period and cancer cases were defined to be biopsy-proven presence of prostate cancer, either by a for-cause biopsy for cancer diagnosis during follow-up, or end-of-study biopsy for all men without a cancer diagnosis. The primary result was mixed: though finasteride reduced the risk of prostate cancer by 25%, the risk of high-grade prostate cancer (Gleason score >6) was elevated in the finasteride arm. Because of this result, finasteride was not approved as a chemoprevention agent for prostate cancer. Since then, substantial efforts have been devoted to understanding the biology of finasteride and prostate cancer, including a P01 research program with five projects and a number of single nucleotide polymorphisms (SNPs) in candidate genes genotyped for a case-control subset of participants (Goodman et al., 2010). We are interested to discover whether there were genotype-specific subgroups in which finasteride may reduce the risk of high-grade prostate cancer.

As in the PCPT trial, outcome-dependent sampling such as case-control sampling is commonly used to increase cost-efficiency when biomarkers are expensive to measure, or specimen collection is difficult. The full set of association parameters including gene-treatment interactions can be estimated without losing too much efficiency, compared to a full-cohort analysis. One of the more cost-efficient approaches to discovering baseline markers that modify treatment effect in randomized trials is the case-only approach (Vittinghoff and Bauer, 2006; Dai et al., 2012, 2014, 2016, 2017), exploiting the gene-treatment independence dictated by randomization. Only samples from cases are needed for measuring markers, a huge cost reduction compared to the traditional case-control or case-cohort approach. This method was originally proposed for estimating gene-environment interactions in epidemiologic studies (Piegorsch et al., 1994; Umbach and Weinberg, 1997), subject to limitations that gene-environment independence does not always hold (Albert et al., 2001), and that the disease of interest has to be rare (Schmidt and Schaid, 1999).

In a series of works, case-only methods have been investigated and extended to genetic studies in randomized clinical trials, large-scale prevention trials in particular. The gene-environment dependence in observational studies is no longer a concern because of randomization. Briefly, for clinical trials with binary endpoints (Dai et al., 2012, 2017), and

rare failure events(Vittinghoff and Bauer, 2006; Dai et al., 2014, 2016), case-only data are sufficient to estimate the subgroup-specific treatment effects and gene-treatment interactions in the relative-risk scale and in the odds-ratio scale, the latter of which can be an approximation for rare disease endpoints, which has been investigated in simulation experiments(Vittinghoff and Bauer, 2006; Dai et al., 2017). In particular, for trials with binary outcomes, the rarity of the study endpoint is no longer needed, and the case-only estimator is consistent for estimating treatment effects and interactions on the relative risk scale(Dai et al., 2017). Another statistical advantage of the case-only estimator is that it can be implemented by any parametric or nonparametric learning method for delineating high-dimensional gene-treatment interactions(Dai et al., 2017). While efficient for estimating gene-treatment interaction, using data in cases only have two consequences: first, we are not able to estimate the main (genetic) effect without control data; second, the estimand has to be in risk ratio (relative risk) but not odds ratio except in the rare disease scenarios.

In this paper, we take the case-only approach to one of the most widely used nonparametric learning methods - trees and random forests - to explore gene-treatment interactions and gene-specific treatment effects in the relative risk scale. By recursively partitioning the predictor space, the hierarchical tree method makes piecewise approximation of the underlying response-predictor relationship. The applications of tree models have been greatly advanced since the development of CART (classification and regression trees) (Breiman et al., 1984). More recently, the random forests method builds an ensemble classifier with a smooth response-predictor surface by averaging a large number of trees, thereby delivering much improved prediction(Breiman, 2001). The goal of this paper is to elucidate the mechanics and the benefits of extending the case-only approach to trees and random forests, and make comparison to the interaction tree method for subgroup analyses(Su et al., 2008, 2009). The key statistical advantage of the proposed case-only approach, beyond the reduction of the cost and the resource by using cases only, is that it enables straightforward adaptation of existing random forests algorithms for predicting individualized treatment effects, the quantities of keen interest in the precision medicine era.

There are several existing methods and algorithms for tree-based, subgroup-finding methods. However, most of them model continuous trial outcomes, and none of existing methods was introduced to account for case-control sampling. Therefore these methods are not directly applicable to the PCPT study. For example, the interaction trees method(Su et al., 2009) closely follows the goodness of splitting algorithm developed for survival trees(LeBlanc and Crowley, 1993), using the likelihood ratio test statistic for the interaction as the splitting statistic, pruning by the weakest link method as in CART, and selecting the best-sized tree by a testing sample or bootstrap samples. Though originally coded for continuous trial outcomes, the interaction trees can be adapted to account for dichotomous outcomes and case-control sampling, as we implemented in our simulation study for comparison. A potential problem of the interaction tree method is, because splitting is determined by interactions, it forces the main marker effects to have the same tree structure as the gene-treatment interactions, which may cost statistical efficiency in finite samples. Connected to this issue, the permutation step in evaluating variable importance by the interaction random forests method nullifies both the main genetic effect and the interaction(Su et al., 2008, 2009), therefore can be problematic to rank markers that are influential to gene-treatment

interactions. More recently, the interaction trees method has been extended to random forests of interaction trees (RFIT)(Su et al., 2018), in order to estimate the individualized treatment effects (ITEs). However the implementation of RFIT is limited to continuous outcomes, perhaps due to the computational complexity of building a versatile algorithm for random forests, and it does not address case-control sampling. In contrast, the proposed case-only approach directly models gene-treatment interaction structures, removing the need for main-effect modeling. It allows exploitation and adaptation of the existing powerful random forests algorithm in R package randomForest for estimating ITEs. The goal of this paper is to illustrate the advantages of case-only trees and random forests, through a side-by-side comparison to the interaction trees method.

Other existing methods include the virtual twins (VT) method(Foster et al., 2011), the subgroup identification based on differential effect search (SIDES) method(Lipkovich et al., 2011), and the generalized unbiased interaction detection and estimation (GUIDE)(Loh, 2002; Loh et al., 2015). Like the interaction trees method of Su et al. (2008), these methods present adaptations of CART to better achieve the goal of discovering gene-treatment interaction or subgroup treatment effects in randomized clinical trials. None of these methods can directly account for case-control sampling that is commonly used in cancer trials such as PCPT, nor has been extended to random forests to enhance prediction of individualized treatment effects. Therefore these methods are not included in comparison to case-only methods.

The remainder of the paper is organized as follows. In Section 2, the case-only trees and random forests are presented side-by-side with the interaction trees and random forests procedure. Section 3 contains simulation studies designed for assessing the case-only trees and random forests relative to the interaction trees and random forests. In Section 4, we apply the proposed case-only procedure to analyze the genetic study from PCPT for identifying gene-treatment interactions. Section 5 concludes the paper with a brief discussion.

2. Methods

Consider a randomized clinical trial with n participants, each with the data vector (Y_i, Z_i, G_i) for $i = 1, \dots, n$, where Y_i is a binary disease outcome (e.g., developing prostate cancer during the 7-year follow-up period in PCPT), Z_i is the treatment assignment for the active drug or a placebo with probability π or $1-\pi$, and G_i is a vector of p covariates $(G_{i1}, \dots, G_{ip})$. The goal is to discover subgroups with heterogeneous treatment effects that are determined by some of or all p covariates in the risk ratio scale, formally expressed as

$$\frac{\Pr(Y_i = 1 | Z_i = 1, G_{i1}, \dots, G_{ip})}{\Pr(Y_i = 1 | Z_i = 0, G_{i1}, \dots, G_{ip})}.$$

A classification tree procedure for identifying gene-treatment interaction is appealing, because recursive partitioning in tree-based algorithms will naturally result in terminal nodes as mutually exclusive subgroups(Breiman et al., 1984). There are prior efforts in developing tree based procedures for subgroup discovery, for censored survival outcomes and for binary

outcomes(Negassa et al., 2005; Su et al., 2008, 2009). Using a splitting statistic for interaction, the classical CART algorithm can be adapted for growing a large initial tree, pruning back sequentially, and determining the best-size tree by a validation method with honest estimates of prediction errors(Su et al., 2009). Variable importance in the random forests method can be extracted for ordering covariates that modify treatment effect(Breiman, 2001; Su et al., 2009). This tree-based method for subgroup analysis was called “interaction trees”, which will serve as the comparator to the case-only tree next.

For trials with binary endpoints, a case-only tree procedure is much more efficient than the standard case-control analysis for identifying genotype-specific treatment effects(Dai et al., 2012, 2014, 2016, 2017), because of exploiting the independence of germline genetic variants and the treatment dictated by randomization. Only genotypes in cases are needed, therefore substantially reducing the cost of the planned study for predictive biomarkers. As we will show next, the algorithm is quite simple, adapting the standard CART and the random forests algorithms for growing a large initial tree, determining the best-size tree and assembling a large number of trees to improve prediction. Compared to the standard case-control method or the full cohort analysis, the case-only method cannot estimate any of the main genetic effects, defined for G_{i1} for example as

$$\frac{\Pr(Y_i = 1 | Z_i, G_{i1} = c, \dots, G_{ip})}{\Pr(Y_i = 1 | Z_i, G_{i1} = 0, \dots, G_{ip})}.$$

If the goal is to find the predictive genetic markers among a set of candidates that modified treatment effects, the genetic main effects are less of interest and therefore can be neglected for the discovery stage.

2.1. The splitting statistic and an initial large tree

As an illustrative example for a binary genetic feature G_1 , one of many genotypes being evaluated, Figure 1 shows a single split for the case-only tree and for the interaction tree. The interaction estimands and corresponding statistics for the case-only tree and the interaction tree are quite different. Important intuition about operating characteristics of the two methods can be gained by inspecting the respective splitting statistics.

2.1.1. Interaction Tree—The interaction tree method was originally proposed for continuous trial outcomes(Su et al., 2009). It is straightforward to extend it to trials with dichotomous outcomes and case-control sampling. In Figure 1, the standard odds-ratio (OR) interaction estimate between Z and G_1 , which involves all eight cell counts stratified by Y , Z and G_1 , is expressed as

$$\begin{aligned} \text{Interaction(OR)} &= \frac{\Pr(Y = 1 | G_1 = 1, Z = 1)}{\Pr(Y = 0 | G_1 = 1, Z = 1)} \bigg/ \frac{\Pr(Y = 1 | G_1 = 1, Z = 0)}{\Pr(Y = 0 | G_1 = 1, Z = 0)} \\ &= \frac{\Pr(Y = 1 | G_1 = 0, Z = 1)}{\Pr(Y = 0 | G_1 = 0, Z = 1)} \bigg/ \frac{\Pr(Y = 1 | G_1 = 0, Z = 0)}{\Pr(Y = 0 | G_1 = 0, Z = 0)} \\ &= \frac{eh/fg}{ad/bc}. \end{aligned}$$

The best split in a node by the interaction tree method is determined by the 1-df likelihood ratio statistic for the interaction parameter, comparing the full model with the treatment main effect, the genetic main effect, and the gene-treatment interaction to the reduced model without the gene-treatment interaction. A Wald test can be similarly used with the variance of the logarithm of this OR interaction estimate $\frac{1}{a} + \frac{1}{e} + \frac{1}{g} + \frac{1}{c} + \frac{1}{b} + \frac{1}{h} + \frac{1}{f} + \frac{1}{d}$. The case-control sampling scheme can be accounted for by inverse-probability weighting. From the model-fitting point of view, this way of splitting may not be the best exercise as it enforces tree growing based on the 1-df interaction solely, yet the main genetic effects may have a tree structure that is completely different from the interaction tree. This drawback will cost the performance of the interaction tree method, as we show later in simulations.

2.1.2. Case-only Tree—A single split in growing the case-only tree will be based on the case-only relative-risk interaction estimate. The standard relative-risk interaction estimate is expressed as

$$\begin{aligned} \text{Interaction(RR)} &= \frac{\Pr(Y = 1 | G_1 = 1, Z = 1)/\Pr(Y = 1 | G_1 = 1, Z = 0)}{\Pr(Y = 1 | G_1 = 0, Z = 1)/\Pr(Y = 1 | G_1 = 0, Z = 0)} \\ &= \frac{\frac{h/(g+h)}{f/(e+f)}}{\frac{d/(c+d)}{b/(a+b)}}. \end{aligned} \quad (1)$$

Both OR interaction and RR interaction should be compared to 1 to assess heterogeneous treatment. As we have shown previously (Dai et al., 2017), this relative-risk (RR) interaction can be reduced to the case-only interaction due to randomization of Z with respect to G_1 . Specifically, for any parental node the randomization between G_1 and Z holds in probability so that $\Pr(Z | G_1) = \Pr(Z)$. Observe that by the Bayes theorem and for a subgroup with $G_1 = g$

$$\begin{aligned} \frac{\Pr(Z = 1 | G_1 = g, Y = 1)}{\Pr(Z = 0 | G_1 = g, Y = 1)} &= \frac{\Pr(Y = 1 | Z = 1, G_1 = g)\Pr(Z = 1 | G_1 = g)}{\Pr(Y = 1 | Z = 0, G_1 = g)\Pr(Z = 0 | G_1 = g)} \\ &= \frac{\Pr(Y = 1 | Z = 1, G_1 = g)}{\Pr(Y = 1 | Z = 0, G_1 = g)} \frac{\pi}{1 - \pi}, \end{aligned} \quad (2)$$

where π is the randomization fraction to the active treatment arm. This equation states that the odds of $Z = 1$ among cases with genotype g is proportional to the treatment effect (on the relative risk scale) in this subgroup. From (1) - (2), the case-only interaction estimate in the RR scale is

$$\frac{\Pr(Z = 1 | G_1 = 1, Y = 1)}{\Pr(Z = 0 | G_1 = 1, Y = 1)} \bigg/ \frac{\Pr(Z = 1 | G_1 = 0, Y = 1)}{\Pr(Z = 0 | G_1 = 0, Y = 1)} = \frac{bh}{df}.$$

More generally, a logistic regression model using the case-only data, with the offset $\log\{\pi/(1 - \pi)\}$, yields the subgroup treatment effect in the relative risk scale (Dai et al., 2017). The case-only relative risk interaction estimate is simply bh/fd , requiring only the four cell counts in the cases. The variance of the logarithm of this case-only relative risk interaction estimate is $\frac{1}{b} + \frac{1}{h} + \frac{1}{f} + \frac{1}{d}$. In comparison to the variability of the logarithm of the

standard relative-risk interaction estimator that involves all eight cell counts, this represents a huge reduction of the variability.

2.1.3. Growing the case-only tree—A large initial case-only tree is grown by splitting cases according to their genotype data. For each parental node, we choose the split resulting in the minimal deviance statistic that describes the distribution of treatment assignment in cases, which corresponds to maximal likelihood estimation of the aforementioned case-only interaction parameter via the logistic regression.

Specifically for an ordered covariate G_{1i} with observed covariate values at g_{1i} , splits are of the form $g_{1i} < c$ or $g_{1i} \geq c$ and for categorical variables $g_{1i} \in S_h$ or $g_{1i} \notin S_h$, where S_h is a set of covariate values at node h . The improvement for a split at node h into left and right daughter nodes $l(h)$ and $r(h)$ is

$$D(h) - [D(l(h)) + D(r(h))]$$

where $D(h)$ is the case-only logistic deviance, summed over cases at the node h among cases, $i \in \{Y_i = 1\}$,

$$D(h) = - \sum_{i \in \{Y_i = 1\}, h} d(\hat{P}(S_h, Z_i = 1))$$

and where d is the logistic deviance for an individual and estimated probability of observing $Z_i = 1$ in the node, i.e., $\hat{P}(S_h, Z_i = 1) = P(Z_i = 1 | g_i \in S_h, Y_i = 1)$. For eligible genotypes (g_{1i} for example) within a node S_h , this entails fitting a logistic regression model $Z \sim I(g_{1i} > c | S_h)$, where $c = 0, 1$, and minimizing the deviance of the fitted model.

This is implemented by using the deviance statistic as the splitting statistic in standard CART algorithms, e.g. the R package rpart. Modifications are required to fit CART into the case-only estimation. When growing a large initial tree T_0 with the deviance splitting statistic in CART, a terminal node will be declared when either of two following conditions is met. First, the size of the node is less than a preset threshold, e.g., 20. This threshold is set to be bigger than the usual minimal node size for a standard CART, because the case-only method is based on randomization being balanced in the covariate space (for example the root node in Figure 1 where splitting is optimized). A small node size does not necessarily guarantee the balance of the two treatment assignments given a genotype. Second, the minimum of the number of placebo recipients in a node is less than a preset threshold, e.g., 5. This is because as the number of placebo recipients decreases, there is an increasing probability that splitting may result in a terminal node with few or no placebo recipients. If there is no placebo recipient in a node, the treatment effect in relative risk is infinity, as the case-only estimation entails

$$\text{Treatment effect (RR)} = \frac{\text{Observed probability of treatment assignment in cases}}{\text{Observed probability of placebo assignment in cases}}.$$

2.2. Pruning

Both the case-only tree and the interaction tree use the efficient “weakest link” pruning algorithm. Let $\alpha \geq 0$ denote the complexity parameter and define the split-complexity measure for an interaction tree to be

$$R_\alpha(T) = R(T) - \alpha |S(T)|,$$

where $R(T)$ is the sum over the splitting statistics in the tree T , and $|S(T)|$ is the number of internal nodes. Because the splitting statistic is a likelihood ratio test statistic, $R(T)$ will increase even if there is no tree structure or any outcome association in the data, and $R(T)$ is evaluated in a test sample. This is in contrast to a classical CART algorithm where the cost of the tree (e.g., impurity, deviance and Gini index) evaluated in a test sample will decrease once splits representing random noise were pursued. To prune the initial large tree to a sequence of nested tree, the “weakest link” method in CART was applied in the interaction tree for any given α . For any internal node h , calculate $R(h) = R(T_h)/|S(T_h)|$, where T_h is the branch with h as a root and $|S(T_h)|$ denotes the number of internal nodes of T_h . The node with the minimal $R(h)$ will be pruned. This procedure applies iteratively to generate a sequence of nested subtrees.

The pruning procedure of the case-only tree is simpler, proceeding with the weakest link method in CART, using the classical cost-complexity measure. The cost of a tree is the sum of the deviance statistics over the tree.

2.3. Selecting the best-sized tree

Suppose a sequence of nested interaction trees have been resulted from the pruning procedure, namely $T_M < \dots < T_m < \dots < T_0$, where T_M is the tree with a root node only and T_0 is the largest tree, $<$ means “is a subtree of”.

Because the interaction tree method uses between-node dissimilarity in treatment effects to measure tree-fitting performance, a more sophisticated procedure is needed to select a pruned subtree. First, to correct for optimism in evaluating the best splitting statistic for each tree size, a testing sample is needed to yield the unbiased estimate of prediction error. Sometimes a large testing sample set may not be available, bootstrap samples can be used instead (LeBlanc and Crowley, 1993). Second, the complexity parameter α has to be specified a priori. The sum of splitting statistics for a tree will increase in a test sample even when there is no tree structure, since, under the null hypothesis of no interaction, the likelihood ratio statistic for testing the interaction follows a χ^2 distribution. As noted before (Su et al., 2009), the choice of $\alpha = 2$ is in the spirit of the Akaike information criterion (AIC) and $\alpha = 4$ corresponds roughly to the 0.05 significance level on the $\chi^2(1)$ distribution. Another choice is $\alpha = \log(n)$ as in the Bayesian information criterion (BIC).

The case-only tree adopts the CART algorithm, in which in-node purity drives the fitness of the tree. The selection of the best-sized tree hinges on the selection of the optimal cost-complexity parameter through cross-validation, which is computationally much simpler than using the bootstrap sample in the interaction tree.

2.4. Random forest and variable importance

Ensemble learning methods such as random forests often can improve prediction of a single learner. This is particularly of interest for trees-based methods because a single tree can be unstable because of its binary, deterministic splitting. As shown previously, the prediction performance of the IT method can be improved by random forests (Su et al., 2008). One useful feature of random forests is that, through permutation, it outputs a variable importance measure ranking predictive features and enabling selection of the most influential ones for a replication or validation study. However, it is much harder to assess the importance of a predictive marker to delineating treatment effect heterogeneity. As we explain next, the variable importance measure proposed by the interaction random forests may not be correct for evaluating solely an individual marker's contribution to interaction and treatment effect heterogeneity.

Algorithm 1 describes in our notation the steps to compute the variable importance measure in the interaction random forest (Su et al., 2009). Briefly let V_j denote the importance measure of the j th marker G_j . The random forests of the interaction tree method are constructed by taking B bootstrap samples $\mathcal{L}_b$, $b = 1, \dots, B$. This is done by searching over only a subset of randomly selected m_0 covariates at each split. For each IT tree T_b , the b th out-of-bag sample (denoted as $\mathcal{L} - \mathcal{L}_b$), which contains all observations that are not in $\mathcal{L}_b$ is sent down T_b to compute the interaction measure $R(T_b)$. The problem of Algorithm 1 lies in the permutation step for G_j : permuting the values of G_j not only nullifies the targeted gene-treatment interaction, but also removes the main genetic effect. The reduction of $R(T_b)$ is therefore a mixed outcome of removing both the main genetic effect and the interaction. In a regression setting, the permutation test with reshuffled genotypes does not provide a valid null distribution for testing the interaction.

Algorithm 1 Computing the variable importance measure via the standard interaction random forests (Su et al., 2009)

```

Initialize all  $V_j$ s to 0.
for  $b = 1, 2, \dots, B$  do
    Generate bootstrap sample  $\mathcal{L}_b$  and obtain the out-of-bag sample  $\mathcal{L} - \mathcal{L}_b$ .
    Grow a large IT tree  $T_b$  in  $\mathcal{L}_b$  by searching over  $m_0$  randomly selected covariates
    at each split.
    Send  $\mathcal{L} - \mathcal{L}_b$  down  $T_b$  to compute  $R(T_b)$ 
    for all genetic markers  $G_j$ ,  $j = 1, \dots, G_p$  do
        permute the values of  $G_j$  in  $\mathcal{L} - \mathcal{L}_b$ 
        Send the permuted  $\mathcal{L} - \mathcal{L}_b$  down  $T_b$  to compute  $R_j(T_b)$ 

        Update  $V_j \leftarrow V_j + \frac{R(T_b) - R_j(T_b)}{R(T_b)}$ .

Average  $V_j \leftarrow V_j / B$ .

```

Like other tree methods, the prediction performance of case-only tree methods can be improved by random forests. Case-only random forest can be implemented by the R package `randomForest`, using the treatment assignment as the outcome and genotypes as

predictors. Two modifications are required for unbiased prediction. First, the individual treatment effects in relative risk have to be estimated from each tree in random forests using expression (2) as odds of treatment assignment being 1 in each terminal node, then averaged across trees to obtain the prediction of treatment effect. This is different from the standard prediction from random forests by averaging the 0–1 votes across trees. Otherwise the estimated individualized treatment effects are biased because odds of the averaged probabilities for the treatment are not equal to the average of the odds of the treatment (case-only treatment effect), a result following Jensen's inequality. Second, the standard random forests method will grow a very large tree. Here the case-only random forests will need to stop splitting when 20–30 cases left in the node. This is again because case-only estimation requires balanced randomization in any baseline stratum and some placebo recipients for stable estimation. The randomForest package outputs variable importance measures computed by permutation algorithms similar to Algorithm 1, except that the permutation step in the case-only random forests is a valid test for the interaction, because case-only methods estimate the interaction parameter without having to estimate the main genetic effect at the same time. So case-only random forests generate a more principled variable importance measure that is useful for ranking predictive features and selecting the most influential ones for a replication study.

3. Results

3.1. Simulations

3.1.1. Simulation set-up—Hereafter the interaction tree method will be called “IT”, the case-only tree method will be called “CT”. The performance of CT and IT in prediction and variable importance was compared in simulated experiments. Briefly, a randomized clinical trial with either 1000 or 5000 participants was generated with probability 0.5 assigned to either the investigational treatment or the control treatment. To evaluate the performance of the two methods in a variety of settings, we designed the following four interaction models. In Model 1–3, eight SNPs ($G_1, \dots, G_8$) were generated with minor allele frequency ranging from 0.3 to 0.5, evaluated at the genotypic value (0, 1, 2), and two of the eight SNPs modify the treatment effect in different ways and the rest of six SNPs do not interact with the treatment. In model 4, three of eight predictors are continuous and interact with the treatment. In model 5, the data-generating model is similar to model 4 but in a logistic model. The goal is to evaluate the sensitivity and robustness of the risk-ratio estimation to model mis-specification.

Similar to the PCPT study, case-control sampling was employed to obtain all cases and a random subset of controls with the same number as cases. The proportion of cases is 10%, so that on average there are 100 cases in each simulated dataset. The goal of case-control sampling is to reduce the cost of genotyping while preserving statistical efficiency in estimation. The comparison was made between IT and CT for estimating the relative risk treatment effect and gene-treatment interaction, based on the case-control data and the case-only data respectively. The IT method was implemented by modifying the R code provided by Dr. Xiaogang Su (also available at <http://biopharmnet.com/subgroup-analysis-software/>) to account for dichotomous endpoints and case-control sampling. Due to complexity of

programming for random forests, we are not able to modify the code for RFIT to account for dichotomous endpoints and case-control sampling (Su et al., 2018). The CT method was implemented by R package `rpart`. The case-only random forest method was implemented by R package `randomForest`.

- **Model 1:** three subgroups were defined by G_1 and G_2 with fractions of participants (0.36, 0.54, 0.1), shown in Figure 2. The logarithm of the subgroup treatment RR is $(-2, -0.5, 1)$. This represents a setting where there are three subgroups with substantial differences in treatment effects, shown as the following data generating model for $\Pr(Y = 1|Z, G_1, G_2)$

$$\exp\{-2 - 2I(G_1 = 0)Z - 0.5I(G_2 \leq 1, G_1 > 0)Z + 1I(G_2 = 1, G_1 > 0)Z\}.$$

The treatment effect among subgroups can be represented by a tree structure. This data-generating model is a log-linear model and there is no main genetic effect for SNPs and the treatment.

- **Model 2:** three subgroups were defined by G_1 and G_2 with fractions of participants (0.91, 0.08, 0.01). The logarithm of the subgroup treatment RR is $(-2, -0.5, 1.5)$. This represents a setting where one of the three groups is rare but with substantial differences in treatment effects, shown in the following data-generating model for $\Pr(Y = 1|Z, G_1, G_2)$

$$\exp\{-1.5 - 2I(G_1 \leq 1)Z - 0.5I(G_1 = 2, G_2 \leq 1)Z + 1.5I(G_1 = 2, G_2 = 2)Z\}$$

This treatment effects among subgroups can be represented by a tree structure. The data-generating model is a log-linear model and there is no main genetic effect for SNPs and the treatment.

- **Model 3:** three subgroups were defined by G_1 and G_2 with fractions of participants (0.36, 0.54, 0.1). The logarithm of the subgroup treatment RR $(-2, -0.5, 1)$. The treatment effect model is identical as Model 1, however there are main genetic effects associated with G_1 and G_2 , shown in the following data generating model for $\Pr(Y = 1|Z, G_1, G_2)$

$$\exp\{-2 + 0.3I(G_1 > 0) - 0.5I(G_1 = 0, G_2 = 2) - 0.8I(G_1 = 0, G_2 < 2) - 2I(G_1 = 0)Z - 0.5I(G_2 \leq 1, G_1 > 0)Z + 1I(G_2 = 2, G_1 > 0)Z\}.$$

There is no main effect for the treatment. Note that the main effects take a tree structure that is different from the interaction tree at different cut points.

- **Model 4:** Three continuous biomarker G_6, G_7, G_8 were generated by a uniform distribution $[0, 1]$, all of which were discretized at $(0, 1/100, 2/100, \dots, 1)$. The rest of 5 markers are SNP genotypes generated as in Model 1–3. There is no main effect for any of eight markers, nor for the treatment. This represents a setting where the gene-treatment interactions were smooth and continuous, as in the model below for $\Pr(Y = 1|Z, G_1, G_2)$

$$\exp\{-2 - 0.75G_6Z - 0.75G_7Z + 0.75G_8Z\}.$$

- **Model 5:** We created one model using the logistic model to test the sensitivity of methods toward model misspecification. Three continuous biomarker G_6 , G_7 , G_8 were generated by a uniform distribution $[0,1]$, all of which were discretized at $(0,1/100, 2/100, \dots, 1)$. The rest of 5 markers are SNP genotypes generated as in Model 1–3. There is no main effect for any of eight markers, nor for the treatment. This represents a setting where the gene-treatment interactions were smooth and continuous, as in the model below for $\Pr(Y = 1|Z, G_1, G_2)$

$$\frac{\exp\{-1 - 0.75G_6Z - 0.75G_7Z + 0.75G_8Z\}}{1 + \exp\{-1 - 0.75G_6Z - 0.75G_7Z + 0.75G_8Z\}}.$$

3.1.2. Simulation results—The goal of the case-only trees is to represent marker-defined subgroups by a parsimonious tree structure, so that each terminal node naturally represents a subgroup with a different treatment effect. The goal of the case-only random forests is to predict accurately the individual treatment effects based on baseline markers. We therefore summarize the simulation results by the following metrics: the number of terminal nodes (average in simulated datasets), the percentage of simulations when the true underlying tree structure was recovered (where splits in X_1 and X_2 are corrected specified), the correlation of estimated individual treatment effects and true individual treatment effects (average in simulated datasets), and the mean squared error of estimated individual treatment effects (average in simulated datasets). We define the identification of the true underlying tree model by both the correct order of splitting variables and the correct splitting points, as this is the most parsimonious way to represent subgroups.

Table 1 shows the four evaluation metrics in the testing set for three methods under investigation: the IT method, the CT method with the final tree selected by the minimum cross-validation deviance, and the case-only random forests. Again, we are not able to modify RFIT for binary outcomes and case-control sampling, so RFIT is not included in the comparison. Figure 2 shows the boxplots of these correlations for the three methods under the four models. The sample size of the simulated training clinical trial is set to be 1000. A separate testing set was generated by the same probabilistic model with the equal sample size. These performance metrics were averaged over 200 simulations.

In Table 1 we listed results for the IT method using $\alpha = 2$ or $\alpha = 4$ as the tuning parameter in the cost-complexity measure, the latter generally produced a smaller model (often no splitting at all with 1 node) and worse performance in correlation of predicted treatment effects with true treatment effects. Because smaller models are often have smaller variances, using $\alpha = 4$ for IT yields a smaller MSE in most of models. When $\alpha = 2$, the resulting IT have a generally larger number of terminal nodes compared to CT (Model 1, 2, 4). Despite the smaller final tree size, the case-only tree is more accurate in identifying correct variables and correct splits in Model 1–3, and produce better prediction of individual treatment effects for Model 1–4 (Table 1 and Figure 2, higher correlation and lower prediction MSE). The

advantage of case-only tree methods is more pronounced in Model 3–4. This is because Model 3 presents a scenario in which the main genetic association takes a tree structure that is different from the interaction tree, thus the IT method can not yield the correct model fitting for main genetic effects, which affects the accuracy in discovering interactions. CT does not model the main genetic associations and therefore outperform the interaction tree method in prediction. Model 4 has three additive, continuous predictors for interactions, and therefore neither of the two tree-based methods is able to obtain the correct model.

As expected, the case-only random forest method offers a substantial improvement in prediction over the case-only tree-based methods, even in a simple three-subgroup setting such as Model 1. The prediction performance of case-only random forests is much better in the latter two models, particularly in Model 4, where the model structure for interaction is much more continuous and therefore more difficult for single tree-based methods to capture. In Model 5, although the data generating model is logistic, the case-only random forests still deliver a much higher correlation.

Figure 2 shows the boxplots of the correlations of estimated and true individualized treatment effects in the testing set in 200 simulations. For IT the results of $\alpha = 2$ were plotted. In model 1–3, a correlation near zero means there isn't any interaction structure being found, and a correlation near 1 means that the true 3-subgroup structure was correctly recovered by the tree-based methods. Clearly case-only trees are more likely to discover the correct interaction structure and better predict individual treatment effects. As noted in Table 1, the case-only random forests method provides a substantial improvement in prediction performance in all models.

Table 2 shows the same set of performance metrics when the sample size of the simulated clinical trial is 5000. The goal of this set of simulations is to examine model selection and prediction performance when the power to detect the gene-treatment interactions increases. The size of the final tree tends to be similar in this setting, and the advantage of the prediction performance of the CT method is less pronounced when compared to the IT method. Take Model 2 for example, IT generated a smaller tree with the average number of terminal nodes around 2.17, while CT resulted in the average number of nodes 3.31. This result suggests that the interaction tree likely correctly identified the first split, but has difficulty to capture the second split correctly because of the low frequency for the third subgroup, therefore only 4% simulations have resulted in the correct model. This caused IT an inferior prediction performance (Table 2).

One of the major advantages of the case-only random forests method relative to the interaction random forests method is unbiased assessment of variable importance, since the permutation procedure used in a standard random forests method would nullify both main genetic effects and gene-treatment interactions. As discussed, the variable importance measure derived from the interaction random forests method may not solely reflect the contribution of gene-treatment interaction. The case-only random forests method does not suffer from this limitation since it does not need to model main genetic effects. Figure 4 shows the boxplots of the variable importance measures in 200 simulations ($n=1000$) for the interaction random forests method and the case-only random forests method. This advantage

of the case-only random forests method in assessing variable importance is clearly seen in Model 1 and 3, where the case-only random forests method correctly ranks the importance of 8 predictors. In contrast, the interaction random forests method fails to rank x_1 as one of top predictors in terms of variable importance for Model 1 and 3.

3.2. Data application

In the Prostate Cancer Prevention Trial (PCPT), 18,882 men 55 years of age or older with a normal digital rectal examination and a prostate-specific antigen (PSA) level of 3.0 ng per milliliter or lower were randomized to either finasteride, an inhibitor of 5 α -reductase, or placebo for preventing prostate cancer. The primary intent-to-treat analysis showed a 24.8% reduction in prevalence of prostate cancer in the finasteride arm during the seven-year follow-up period, yet a higher rate of high-grade prostate cancer (Gleason grade 7) was detected in the finasteride recipients (Thompson et al., 2003). Though the reason for the increased risk in the finasteride arm has yet to be determined, these results complicate the potential of finasteride as a preventative agent for prostate cancer. In an effort to understand the biological mechanism of the finasteride effect on the risk of prostate cancer (a UM1 project), several hundred of SNPs in genes in multiple androgen pathways were genotyped by case-control sampling of stored blood specimens. Here we use case-only tree methods to discover genotypes that modify the finasteride effect on high-grade prostate cancer and investigate whether there are subgroups who do not have an increased risk for high-grade prostate cancer.

We restricted our analysis to 306 high-grade prostate cancer cases and 357 SNPs which passed quality control (missing data less than 20% and Hardy-Weinberg equilibrium). The missing genotypic values were imputed by the mean genotypic value in cases or controls depending on the case-control status of the man missing the genotype. The tree algorithm in the R package `rpart` was used with 10-fold cross-validation to select the tree with the minimum deviance. The `randomForest` package was used to evaluate variable importance using 2000 bootstrap samples, each split to trees with terminal nodes size greater or equal to 30. Figure 5(a) shows the final case-only tree by cross-validation. Two SNPs were retained in the tree, and the 3 terminal nodes represent a subgroup with treatment relative risk 0.44 ($rs472402=0/1$), a subgroup with treatment relative risk 0.54 ($rs472402=2$ and $rs1052536=0/1$), and a subgroup with treatment relative risk 2.13 ($rs472402=2$ and $rs1052536=2$). In other words, while the overall intent-to-treat effect was hazardous, there were two subgroups (37% in the cases) in which finasteride was protective in reducing the risk of high-grade prostate cancer.

Figure 5(b) shows the variable importance plot for the top twenty SNPs generated by the `randomForest` package. Two measures were used to assess the importance of SNP predictors: prediction accuracy and Gini index. Both measures identified the top four important predictors, the first and the third of which were picked by `rpart`: $rs472402$, $rs11689432$, $rs1052536$, and $rs3756423$. The rest of SNPs show less separation in variable importance and less concordance between two measures. We used the case-only random forests to predict the individualized treatment effects for cases and controls. The out-of-bag error rate is 37.58%, though this does not connect to the accuracy of estimating

the heterogeneous individual treatments. The latter is not directly observable and it is an interesting research problem how to evaluate the accuracy of predicting individual treatment effects in real data. We included with different weights to adjust for case-control sampling and generated a distribution of individualized treatment effects corresponding to the population of men in the PCPT trial. Figure 5(c) shows the distribution of the estimated individualized treatment effects in PCPT. The vertical line indicates zero treatment effect ($RR=1$). There were estimated 37% of PCPT participants whose risk of high-grade prostate cancer were decreased by finasteride ($RR<1$).

Due to the limited sample size for cases, we did not split the samples to training and testing sets. Therefore the random forests may have the overfitting issue. As a partial internal validation of case-only analyses, for the top four SNPs ranked by the importance measures in case-only random forest, Table 3 shows the results of fitting a standard logistic model with the treatment, the genotype (one SNP at a time) and the interaction term to the case-control data. Except for rs11689432, the four SNPs have rare missing values. All SNPs have a moderate level of genetic association in the placebo arm, and all of them have a nominally significant interaction with the treatment when evaluated one SNP at a time and using both cases and controls.

4. Discussion

Discovering gene-treatment interactions and identifying predictive markers are central for precision medicine, treatment and prevention. CART, one of the most popular nonparametric methods, has been modified and adapted to identification of subgroups in randomized clinical trials, including IT(Su et al., 2008, 2009), VT(Foster et al., 2011), SIDES(Lipkovich et al., 2011), and GUIDE(Loh, 2002; Loh et al., 2015), each of which addresses different issues of CART algorithms. The proposed case-only trees and random forests methods offer three advantages over existing methods. First, case-only methods only need biomarker data from cases, a huge reduction of the cost when a large number of molecular features such as genotypes are to be investigated for useful predictive markers. This approach is perhaps ideal for early exploratory studies to generate a list of candidate predictive markers. Second, because of exploiting gene-treatment independence dictated by randomization, case-only methods generally deliver more efficient estimators of gene-treatment interactions(Dai et al., 2017). Third, case-only trees and random forests directly target gene-treatment interactions, circumventing the necessity of separately modeling main genetic effects as prognostic factors. See, for example, the partially additive linear model trees(Seibold et al., 2016). The latter method allows model parameters specified a priori to be estimated globally for all observations, or locally from subgroups (for example heterogeneous treatment effects). In particular, main effects in a gene-treatment interaction model can be specified as linear additive prognostic variables that act on the study outcome globally. We have carefully examined the features of case-only trees and random forests in algorithm steps, prediction performance, and variable importance, in comparison to the existing interaction tree method side-by-side. Our results suggest that the proposed case-only methods deliver superior prediction and better measures of variable importance, despite the much reduced cost and specimens being processed.

One of notable advantages of the proposed case-only method is that it allows a straightforward adaptation of the existing random forests algorithm in the R package `randomForest`. In the simulation experiment with a limited number of markers, case-only random forests deliver a convincing performance in predicting individualized treatment effects. Computationally, `rpart` and `randomForest` can handle hundreds of markers, though not all markers are useful for prediction. `randomForest` has a built-in function (`rfcv`) to sequentially reduce the number of predictors via cross validation. When there are a great number of markers for building a prediction signature, one can consider the univariate case-only interaction test to first select a subset with a soft threshold (nominal p-value < 0.2 for example), then employ the random forests method to improve the prediction.

The best scenario for case-only methods would be a high-dimensional study in prevention trials with a relatively rare disease endpoint (for example, the setting of PCPT trial in the data example) for potential effect-modifiers, e.g. GWAS for gene-treatment interaction. In this setting, case-only methods have the greatest cost reduction while having the efficiency of estimating gene-treatment interaction as if genotyping the entire trial population. On the flip side of case-only methods, if the interest is the ratio of the survival probability in the trial at a fixed time point, then the control-only analysis can be used to estimate interaction on the survival probability. This quantity may be of interest in many cancer therapeutic trials.

A weakness of the PCPT data example is that, it is difficult to establish a validation dataset for the effect-modifying SNPs in the PCPT data application, since PCPT is the only trial that tests the effect of finasteride for preventing prostate cancer. However the consistency between case-control interaction estimates and the case-only interaction estimates is reassuring on the credibility of the top effect-modifying SNPs.

Other than predicting treatment effects on primary efficacy endpoints, an important application area is prediction of adverse events for cancer treatment based on underlying patient genetics. For instance, the assessing severe immunotoxicity for checkpoint inhibitor-based therapies is an area of clinical interest where case-only techniques may allow a feasible and efficient path forward.

One limitation of the proposed case-only method is that the clinical endpoint has to be binary. This works for the PCPT study as we illustrated, since the primary endpoint is the prevalence of prostate cancer during the 7-year follow-up period. The majority of therapeutic and prevention trials have time-to-event endpoints. Existing case-only approaches have not been able to account for time-to-event endpoints and the popular Cox proportional hazards model. The future work includes extending case-only approaches to survival outcomes critical to answer many therapeutic treatment questions.

Acknowledgments

We thank Xiaogang Su for sharing R codes to implement the interaction tree and random forests method. We also thank Cathy Tangen for providing PCPT genetic data and helpful comments.

Financial disclosure

This work was supported by the National Institutes of Health grants R01 CA222833, P01 CA53996, R01 HL114901, U10 CA180819. The analysis of the PCPT data was supported by U01CA182883 (PCPT and SELECT cohorts: core infrastructure support for cancer research).

References

- Albert PS, Ratnasinghe D, Tangrea J and Wacholder S (2001) Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology*, 154, 587–693.
- Assmann S, Pocock S, Enos L and Kasten L (2000) Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*, 255, 1064–1069.
- Breiman L (2001) Random forests. *Machine Learning*, 45, 5–32.
- Breiman L, Friedman J, Olshen R and Stone C (1984) *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group.
- Dai J, Kooperberg C, LeBlanc M and Prentice R (2012) Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika*, 99, 929–944. [PubMed: 23843674]
- Dai J, Li S and Gilbert P (2014) Case-only methods for competing risks models with application to assessing differential vaccine efficacy by viral and host genetics. *Biostatistics*, 15, 196–203. [PubMed: 23813283]
- Dai J, Liang C, LeBlanc M, Prentice R and Janes H (2017) Case-only approach to identifying markers predicting treatment effects on the relative risk scale. *Biometrics*
- Dai J, Zhang X, Wang C and Kooperberg C (2016) Augmented case-only design for randomized clinical trials with failure time endpoints. *Biometrics*, In Press.
- Evans W and McLeod H (2003) Pharmacogenomics- drug disposition, drug targets, and side effects. *N Engl J Med*, 348(6), 538–549. [PubMed: 12571262]
- Foster JC, Taylor JMG and Buber SJ (2011) Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30, 2867–2880. [PubMed: 21815180]
- Goodman PJ, Tangen CM, Kristal A et al. (2010) Transition of a clinical trial into translational research: the prostate cancer prevention trial experience. *Cancer Prev Res (Phila)*, 3, 1523–1533. [PubMed: 21149329]
- Lagakos S (2006) The challenge of subgroup analyses- reporting without distorting. *The New England Journal of Medicine*, 354, 1667–1669. [PubMed: 16625007]
- LeBlanc M and Crowley J (1993) Survival trees by goodness of split. *JASA*, 88, 457–467.
- Lipkovich I, Dmitrienko A, Denne J and Enas G (2011) Subgroup identification based on differential effect search — a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30, 2601–2621. [PubMed: 21786278]
- Loh W-Y (2002) Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12, 361–386.
- Loh W-Y, He X and Man M (2015) A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, 34, 1818–1833. [PubMed: 25656439]
- Negassa A, Ciampi A, Abrahamowicz M, Shapiro S and Boivin J (2005) Tree-structured subgroup analysis for censored survival data: validation of computationally inexpensive model selection criteria. *Statistics and Computing*, 15, 231–239.
- Piegorsch WW, Weinberg CR and Taylor JA (1994) Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. *Statistics in Medicine*, 13, 153–162. [PubMed: 8122051]
- Schmidt S and Schaid DJ (1999) Potential misinterpretation of the case-only study to assess gene-environment interaction. *American Journal of Epidemiology*, 150, 878–885. [PubMed: 10522659]
- Seibold H, Hothorn T and Zeileis A (2016) Generalised linear model trees with global additive effects. *arXiv*, 1612.07498.
- Su X, Pena A, Liu L and Levine R (2018) Random forests of interaction trees for estimating individualized treatment effects in randomized trials. *Statistics in Medicine*, 37, 2547–2560. [PubMed: 29707855]

- Su X, Tsai C, Wang H and Nickerson D (2009) Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10, 141–158.
- Su X, Zhou T, Yan X, Fan J and Yang S (2008) Interaction trees with censored survival data. *The International Journal of Biostatistics*, 4, 2.
- Thompson I, Goodman P, Tangen C et al. (2003) The influence of finasteride on the development of prostate cancer. *The New England Journal of Medicine*, 349, 215–224. [PubMed: 12824459]
- Umbach DM and Weinberg CR (1997) Designing and analyzing case-control studies to exploit independence of genotype and exposure. *Statist. Med*, 16, 1731–1743.
- Vittinghoff E and Bauer DC (2006) Case-only analysis of treatment-covariate interactions in clinical trials. *Biometrics*, 62, 769–776. [PubMed: 16984319]
- Weinshilboum R and Wang L (2004) Pharmacogenomics: bench to bedside. *Nat Rev Drug Discov*, 3(9), 739–748. [PubMed: 15340384]

Fig. 1.

A single split for the case-only tree method and the interaction tree method. For a binary Z , a binary G_1 and a binary Y , the interaction tree requires all 8 cell counts for splitting, whereas the case-only tree only requires four cell counts in cases.

Fig. 2.
The subgroup tree structure for the simulation model 1.

Fig. 3. Correlation of predicted individual treatment effects and true individual treatment effects for the three tree-based methods (Interaction tree, case-only tree, and case-only random forest) in 200 simulated trial datasets, each with 1000 participants. The correlation was evaluated in the testing set. The comparisons are made for five models: (a) Model 1; (b) Model 2; (c) Model 3; (d) Model 4. (e) Model 5.

Fig. 4. Variable importance measures for the eight predictors in Models 1–5. The boxplots shows the median and 25%, 75% quartiles for the variable importance measures in 200 simulated datasets each having 1000 participants. (a) Case-only random forest. (b) Interaction random forest.

Fig. 5.

Applying case-only tree and random forest methods to the genetic data (416 SNPs) in 306 high-grade prostate cancer patients in PCPT. (a) the final tree with the minimal prediction error evaluated by 10-fold cross validation. (b) SNPs with top variable importance measures from the case-only random forests. (c) the distribution of individualized treatment effects in case-control samples.

Table 1.

Comparison of tree-based methods in 200 simulated datasets (n=1000).

Method	# nodes	% simulations with correct models	Correlation (testing set)	Prediction MSE (testing set)
Model 1				
Interaction Tree - $\alpha = 2$	2.92	0.12	0.21	0.97
Interaction Tree - $\alpha = 4$	1.56	0.025	0.15	0.61
Case-only Tree	2.49	0.26	0.34	0.98
Case-only random forest	–	–	0.66	0.22
Model 2				
Interaction Tree - $\alpha = 2$	2.69	0.14	0.38	0.28
Interaction Tree - $\alpha = 4$	1.83	0.02	0.35	0.27
Case-only Tree	2.11	0.35	0.46	0.29
Case-only random forest	–	–	0.56	0.21
Model 3				
Interaction Tree - $\alpha = 2$	2.67	0.07	0.16	1.03
Interaction Tree - $\alpha = 4$	1.31	0	0.08	0.63
Case-only Tree	2.75	0.18	0.39	0.93
Case-only random forest	–	–	0.63	0.24
Model 4				
Interaction Tree - $\alpha = 2$	4.06	–	0.16	3.27
Interaction Tree - $\alpha = 4$	1.36	–	0.05	0.90
Case-only Tree	2.43	–	0.15	1.06
Case-only random forest	–	–	0.57	0.22
Model 5				
Interaction Tree - $\alpha = 2$	2.94	–	0.04	0.69
Interaction Tree - $\alpha = 4$	1.03	–	0	0.07
Case-only Tree	2.66	–	0.15	0.61
Case-only random forest	–	–	0.54	0.19

Table 2.

Comparison of tree-based methods in 200 simulated datasets (n=5000).

Method	# nodes	% simulations with correct	Correlation testing set	Prediction MSE testing set
Model 1				
Interaction Tree - $\alpha = 2$	3.19	0.51	0.72	0.26
Interaction Tree - $\alpha = 4$	2.44	0.4	0.69	0.22
Case-only Tree	3.34	0.98	0.96	0.09
Case-only random forest	–	–	0.92	0.07
Model 2				
Interaction Tree - $\alpha = 2$	2.17	0.04	0.59	0.19
Interaction Tree - $\alpha = 4$	2.02	0.01	0.59	0.19
Case-only Tree	3.31	0.93	0.94	0.15
Case-only random forest	–	–	0.89	0.09
Model 3				
Interaction Tree - $\alpha = 2$	2.73	0.31	0.66	0.33
Interaction Tree - $\alpha = 4$	2.1	0.19	0.63	0.28
Case-only Tree	2.72	0.51	0.80	0.17
Case-only random forest	–	–	0.92	0.08
Model 4				
Interaction Tree - $\alpha = 2$	1.97	–	0.02	0.27
Interaction Tree - $\alpha = 4$	1.03	–	0.01	0.29
Case-only Tree	3.27	–	0.21	0.34
Case-only random forest	–	–	0.65	0.16
Model 5				
Interaction Tree - $\alpha = 2$	1.28	–	0.03	0.06
Interaction Tree - $\alpha = 4$	1.08	–	0.02	0.06
Case-only Tree	2.58	–	0.23	0.09
Case-only random forest	–	–	0.62	0.11

Table 3.

The top four SNPs ranked by variable importance in case-only random forests.

SNP	% miss	Minor allele frequency	log(Genetic OR) in placebo		log(Case-control interaction)	
			estimate	p-value	estimate	p-value
rs472402	0.5	0.48	0.39	0.003	−0.57	0.002
rs11689432	15.1	0.43	0.25	0.07	−0.51	0.01
rs1052536	1.2	0.43	0.30	0.02	−0.47	0.01
rs3756423	0.5	0.20	0.25	0.11	0.66	0.004