Received: 25 July 2017 | Revised: 20 December 2017 | Accepted: 7 February 2018

DOI: 10.1002/sim.7660

RESEARCH ARTICLE

WILEY Statistics in Medicine

Random forests of interaction trees for estimating individualized treatment effects in randomized trials

Xiaogang Su[1] | Annette T. Peña[1] | Lei Liu[2] | Richard A. Levine[3]

[1] Department of Mathematical Sciences, University of Texas at El Paso, El Paso, TX 79968-0514, USA

[2] Division of Biostatistics, Washington University in St. Louis, St. Louis, MO 63130, USA

[3] Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182, USA

Correspondence
Xiaogang Su, Department of Mathematical Sciences, University of Texas at El Paso, El Paso, TX 79968-0514, USA.
Email: xiaogangsu@gmail.com

Funding information
NIMHD, Grant/Award Number: 2G12MD007592; AHRQ, Grant/Award Number: HS 020263; NSF, Grant/Award Number: 1633130

Assessing heterogeneous treatment effects is a growing interest in advancing precision medicine. Individualized treatment effects (ITEs) play a critical role in such an endeavor. Concerning experimental data collected from randomized trials, we put forward a method, termed random forests of interaction trees (RFIT), for estimating ITE on the basis of interaction trees. To this end, we propose a smooth sigmoid surrogate method, as an alternative to greedy search, to speed up tree construction. The RFIT outperforms the "separate regression" approach in estimating ITE. Furthermore, standard errors for the estimated ITE via RFIT are obtained with the infinitesimal jackknife method. We assess and illustrate the use of RFIT via both simulation and the analysis of data from an acupuncture headache trial.

KEYWORDS
individualized treatment effects, infinitesimal jackknife, precision medicine, random forests, treatment-by-covariate interaction

1 | INTRODUCTION

Precision medicine aims to optimize the delivery of stratified or individualized therapies by integrating comprehensive patient data. This emerging approach is a growing interest in biomedical applications. Precision medicine faces many statistical challenges before it can be broadly deployed in clinical practice. Many available statistical methods, driven by the "one-size-fits-all" conventional medicine, are primarily concerned about the overall main effect of a treatment over the entire population and rely heavily on the traditional significance testing. To advance precision medicine, 1 critical statistical challenge is to understand and quantify differential treatment effects.

There are many newly proposed approaches in this endeavor; see Lipkovich et al[1] for a recent survey. Among them, tree-based methods[2] are dominant for several reasons. Built simply on the basis of a 2-sample test statistic, trees facilitate a powerful comprehensive modeling scheme by recursively splitting data. Differential treatment effects essentially involve treatment-by-covariate interactions, which may be of nonlinear forms and of high orders. Trees excel in dealing with complex interactions. Furthermore, tree models are capable of handling high-dimensional covariates of mixed types and present as an off-the-shelf tool in the sense that minimal data preparation is required.

Interaction trees (ITs)[3] extend tree methods to subgroup analysis by explicitly assessing the treatment-by-covariate interactions. In the "virtual twins"[4] approach, subgroups are identified by first estimating the potential outcomes. Subgroup identification based on differential effect search (SIDES)[5] seeks subgroups with enhanced treatment effects, possibly taking into account both efficacy and toxicity. Qualitative ITs[6] focus on qualitative interactions. Loh et al[7] proposed a tree procedure for subgroup identification that is less prone to biased variable selection. The optimal treatment regime,[8] which aims to find the recommended treatment based on individual patient information, offers an alternative way of looking at the problem. Along this direction, tree-based approaches are also common.[9,10]

There are typically 2 types of precision medicine: stratified and personalized medicines. The aforementioned methods belong to the former type, primarily concerned about stratified treatment effects or regimes where groups of individuals showing homogeneous treatment effects are sought. Comparatively, individualized treatment effects (ITEs) are of key importance in deploying tailored treatment plans as part of personalized medicine. A model for ITE predicts the effect of treatment on a future patient. Individualized treatment effect assessment affords deeper study of treatment efficacy by quantifying how heterogeneous treatment effects are and whether directional or qualitative interactions exist. This information allows for estimation of the proportion of patients who benefit from the treatment and identification of those who may be harmed by the treatment. Besides, ITE models can pinpoint important predictive factors,[11] ie, patients' characteristics that moderate or modify the treatment effects, and offer insight for understanding the pharmacological mechanisms of a drug. Individualized treatment effect estimation is necessarily a first step for many methods[4,9,10] in stratified medicine and optimal treatment regime. The optimal choice of the treatments would be revealed once ITE is known.

Our focus is on the estimation of ITE with data collected from randomized trials. One available method for this task is separate regression (SR),[4,12] in which separate predictive models for the response variable are built using data in the treated group and data in the untreated group, respectively, and applied to each individual. The difference in predicted response from the 2 models supplies an estimator of ITE. The idea of SR is intuitive within the causal inference framework; we shall elaborate more in the ensuing sections. One major shortcoming of SR is that one has to deal with both prognostic and predictive factors in SR, although ITE assessment involves predictive factors only. Besides, there is no standard error (SE) formula available for the estimated ITE from SR.

To overcome the deficiencies of SR, we examine an ensemble-learning approach for ITE estimation using IT.[3] We coin the proposed method as random forests of interaction trees (RFIT) for RFs of IT. Our methodological contribution is threefold: First, we implement random forests (RFs) on the basis of ITs, which is different from ordinary RFs[13] of classification or regression trees;[2] second, a faster alternative splitting method, called smooth sigmoid surrogate (SSS), is introduced to speed up construction of ITs; and third, we extend the infinitesimal jackknife (IJ) method[14] to compute the SEs for ITE estimates. Compared with SR, RFIT is superior by focusing exclusively on predictive factors. We investigate the performance of RFIT via extensive numerical experiments.

The remainder of this article is organized as follows. In Section 2, we first introduce the concept of ITE within Rubin causal model framework. We then present RFIT with SSS splitting for estimating ITE and the SE formula for estimated ITE. Section 3 contains simulation experiments that are designed to compare RFIT with other methods and assess the proposed SE formulation. In Section 4, we illustrate our proposed RFIT approach with data from an acupuncture headache trial.

2 | RF OF INTERACTION TREES

Consider a randomized trial with data $\mathcal{D} = \{(y_i, T_i, \mathbf{x}_i):i = 1, \cdots, n\}$ consisting of n IID copies of $(Y, T, \mathbf{X})$, where y_i is the continuous response or outcome for the ith subject; T_i is the binary treatment assignment indicator, 1 for the treated group and 0 for control; and $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T \in \mathbb{R}^p$ is a p-dimensional covariate vector of mixed types.

The Neyman-Rubin causal model[15-17] provides a way of finely calibrating the causal effect of treatment T on the response via the concept of potential outcomes. Let $Y_1^{'}$ and $Y_0^{'}$ denote the response values for a subject when assigned to the treated and the control group, respectively. Either $Y_1^{'}$ or $Y_0^{'}$, but not both, can be observed, which is the so-called "fundamental problem of causal inference."[18] The observed outcome is given by $Y = Y_1^{'} T + Y_0^{'}(1-T)$. Within this framework, the treatment effect can be evaluated at 3 levels: the population level $E(Y_1^{'}-Y_0^{'})$ (referred to as the average treatment effect or ATE[18]), the subpopulation level $E(Y_1^{'}-Y_0^{'}|\mathbf{X}\in A)$ for a subset $A\subset\mathbb{R}^p$, and the unit or subject level $Y_1^{'}-Y_0^{'}$. These 3 levels form a hierarchy of causal inference in increasing order of strength, in the sense that ATE can be obtained from the knowledge of subpopulation-level inferences, which in turn can be obtained from the knowledge of unit-level inferences, but not vice versa. Let δ be a generic notation for treatment effect.

Definition 1. The ITE is defined as $\delta(\mathbf{x}) = E(Y_1^{'}-Y_0^{'}|\mathbf{X} = \mathbf{x})$.

Note that $\delta(\mathbf{x})$ is different from the (random) unit-level effect $(Y_1^{'} - Y_0^{'})$. Strictly speaking, $\delta(\mathbf{x})$ is a subpopulation-level effect among individuals with $\mathbf{X}=\mathbf{x}$. Nevertheless, $\delta(\mathbf{x})$ is the finest approximation to the unit-level effect that is possibly available in practice.

Causal inference is essentially concerned with estimating δ at different levels through the available data $\mathcal{D}$. The difficulty in causal inference stems primarily from the convoluted roles (eg, confounder, effect modifier or moderator, or mediator) played by each covariate in $\mathbf{X}$. For experimental data from trials with random treatment assignment mechanisms, T is independent of other variables. As a result, the unconfoundedness condition[17] $(Y_1, Y_0) \perp\!\!\!\perp T \mid \mathbf{X}$, being sufficient for obtaining population-level inference from $\mathcal{D}$, is trivially met. Randomization renders the confounding issue of little concern; however, covariate modification to the treatment effects may remain at both subpopulation and unit levels, referred to as the treatment-by-covariate interactions.

2.1 | SSS for identifying the best cutoff point

Interaction tree[3] seeks subgroups with heterogeneous treatment effects by following the paradigm of Classification and Regression Trees (CART)[2]; hence, IT supplies causal inference at the subpopulation level. Nevertheless, results from IT can be building blocks for inferences at other levels: One has the flexibility to move backward to the ATE estimation by integration and move forward to the ITE estimation via ensemble learning. The main objective of this article is to examine the use of RFIT in estimating $\delta(\mathbf{x})$. Random forests[13] are an ensemble-learning method, constructing a collection of tree models and integrating results. Among its many merits, RF is an off-the-shelf method and a top performer in predictive modeling.[19]

To extend RFs on the basis of ITs, one essential ingredient is the splitting statistic. In CART, one splits data so that the difference in response between 2 child nodes is maximized or, equivalently, the within-node impurity or variation is minimized. In IT, data are split so that the difference in treatment effects between 2 child nodes is maximized. A split on data is induced by a binary variable of general form $\Delta = \Delta(X_j;c) = I(X_j \leq c)$ that applies a threshold on covariate X_j at cutoff point c. When X_j is nominal or categorical, one common strategy is to sort the variable levels according to the treatment effect estimate at each level and treat it as if ordinal. A theoretical justification for doing so can be found in Su et al.[3, Appendix A]

In our setting, any binary split results in the following 2 × 2 table, where n_{1L} denotes the number of treated subjects in the left child node, $\bar{y}_{1L}$ denotes the sample mean response for treated subjects in the left child node, and so on for notation in the other cells.

	Child Node	
Treatment	**Left**	**Right**
0	$(\bar{y}_{0L}, n_{0L})$	$(\bar{y}_{0R}, n_{0R})$
1	$(\bar{y}_{1L}, n_{1L})$	$(\bar{y}_{1R}, n_{1R})$

The splitting statistic in IT can be based on the Wald test of $H_0 : \beta_3 = 0$ in the interaction model:

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 \Delta_i + \beta_3 T_i \cdot \Delta_i + \varepsilon_i \text{ with } \varepsilon_i \overset{IID}{\sim} N(0, \sigma^2), \tag{1}$$

where $\Delta_i = \Delta(x_{ij}; c)$. The least-squares (LS) estimate of β_3 is given by $\hat{\beta}_3 = (\bar{y}_{1L} - \bar{y}_{0L}) - (\bar{y}_{1R} - \bar{y}_{0R})$, corresponding to the concept of "difference in differences."[20] The resultant Wald test statistic amounts to

$$Q(c) = \frac{\{(\bar{y}_{1L} - \bar{y}_{0L}) - (\bar{y}_{1R} - \bar{y}_{0R})\}^2}{\hat{\sigma}^2 \left(1/n_{1L} + 1/n_{0L} + 1/n_{1R} + 1/n_{0R}\right)}, \tag{2}$$

where

$$\hat{\sigma}^2 = \frac{1}{n-4} \left(\sum_{i=1}^{n} y_i^2 - \sum_{k=0,1} \sum_{t \in \{L,R\}} n_{kt} \bar{y}_{kt}^2 \right) \tag{3}$$

is the pooled estimator of σ^2. $Q(c)$ measures the difference in treatment effects between the 2 child nodes. With the conventional greedy search (GS) approach, the best cutoff point $\hat{c}$ for X_j is $\hat{c} = \text{argmax}_c Q(c)$. It is worth noting that minimizing the LS criterion with model (1) does not serve well in IT. A cutoff point can yield the minimum LS criterion merely for its strong additive effect associated with β_2.

Greedy search evaluates the splitting measure at every possible cutoff point for X_j. This can be slow when the number of cutoff points to be evaluated is large, even though GS can be implemented by updating the computation of $Q(c)$ for neighboring c values. Furthermore, this discrete optimization procedure yields erratic fluctuations, as exemplified by the orange line in Figure 1B. As a result, GS may mistakenly select a local spike due to large variation. These deficiencies motivate us to consider a smooth alternative to GS. Our idea is to approximate the threshold indicator function Δ_i, involved in many components of the splitting statistic, with a smooth sigmoid function. For this reason, we call the method "smooth sigmoid surrogate" or SSS in short. While many sigmoid functions can be used, it is natural to consider the logistic or expit function:

$$s(x;a,c) = [1+\exp\{-a(x-c)\}]^{-1} = \frac{\exp\{a\,(x-c)\}}{1+\exp\{a\,(x-c)\}}, \tag{4}$$

with a shape or scale parameter $a > 0$. Figure 1A depicts the expit function for different a values, where $c = 0$ coincides with the mean of a standardized covariate. To approximate $Q(c)$, we start with approximating $n_{l\tau}$ with $\tilde{n}_{lt}$ for $l = 0{,}1$ and $t \in \{L{,}R\}$ as follows:

$$\begin{cases} n_{1L} = \sum_{i=1}^{n} T_i \Delta_i & \approx \quad \tilde{n}_{1L} = \sum_{i=1}^{n} T_i s_i, \\ n_{1R} = n_1 - n_{1L} & \approx \quad \tilde{n}_{1R} = n_1 - \tilde{n}_{1L}, \\ n_{0L} = \sum_{i=1}^{n} (1-T_i)\delta_i & \approx \quad \tilde{n}_{0L} = \sum_{i=1}^{n} (1-T_i) s_i, \\ n_{0R} = n_0 - n_{0L} & \approx \quad \tilde{n}_{0R} = n_0 - \tilde{n}_{0L}, \end{cases}$$

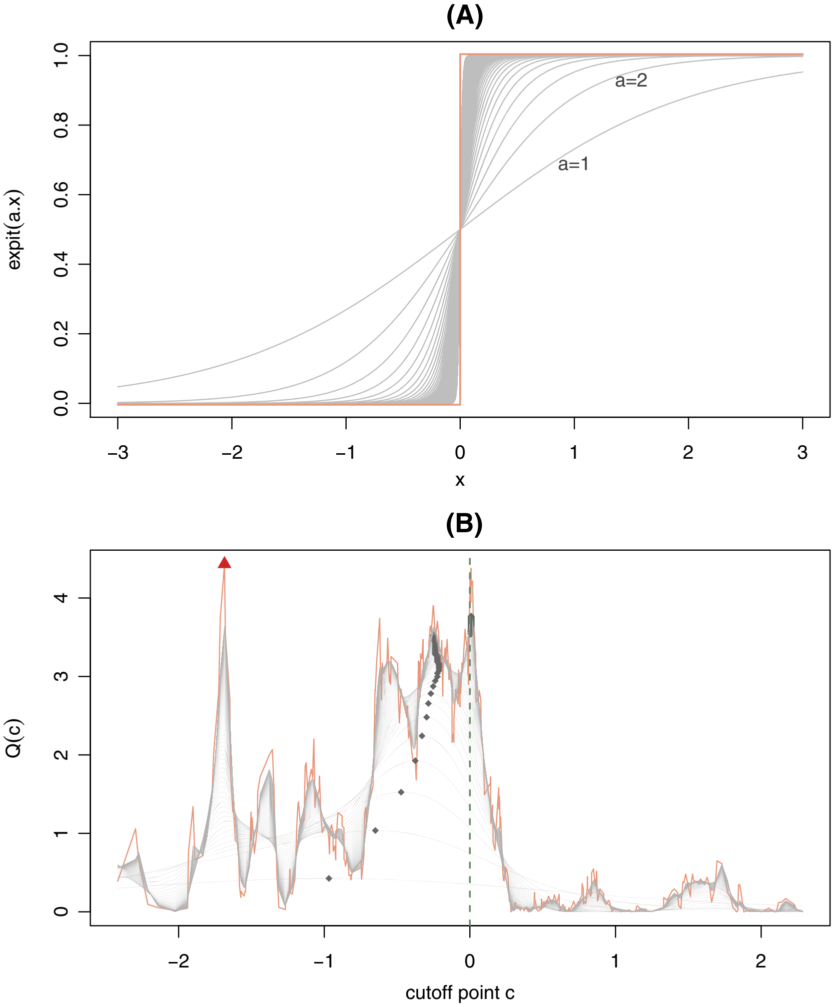

FIGURE 1 Illustration of smooth sigmoid surrogate (SSS) for splitting data: A, the discrete threshold function $\Delta(x;c) = I(x \geq c)$ with $c = 0$ (in orange) and its expit approximation $s\,(x;c) = \text{expit}\{a\,(x-c)\}$ (in gray); B, the splitting statistic $Q(c)$ computed at each cutoff point c in greedy search and its SSS approximations with a={1,2,…,100}. In panel B, data of size $n = 500$ are generated from model $y = 0.5 + 0.5T + 0.5\Delta + 0.5 \cdot T\Delta + \varepsilon$, where $\Delta = \Delta(x;c_0)$ with true cutoff point c_0=0 (indicated by the green dashed vertical line) and both x and ε are from $N(0{,}1)$. The best cutoff point found by greedy search is denoted by the red triangle, while the blackblack diamond dots indicate the best cutoff points found by SSS with different a values. [Colour figure can be viewed at wileyonlinelibrary.com]

where $s_i = s(x_{ij};a,c)$ approximates Δ_i, $n_1 = \sum_i T_i$ is the total number of treated individuals, and $n_0 = \sum_{i=1}^n (1-T_i)$ is the total number of untreated individuals. Let S_{lt} denote the associated sum of observed responses values in each cell. They can be approximated in a similar manner:

$$\begin{cases} S_{1L} = \sum_{i=1}^n y_i T_i \Delta_i & \approx \quad \tilde{S}_{1L} = \sum_{i=1}^n y_i T_i s_i, \\ S_{1R} = S_1 - S_{1L} & \approx \quad \tilde{S}_{1R} = S_1 - \tilde{S}_{1L}, \\ S_{0L} = \sum_{i=1}^n y_i (1-T_i) \Delta_i & \approx \quad \tilde{S}_{0L} = \sum_{i=1}^n y_i (1-T_i) s_i, \\ S_{0R} = S_0 - S_{0L} & \approx \quad \tilde{S}_{0R} = S_0 - \tilde{S}_{0L}, \end{cases}$$

where $S_1 = \sum_i T_i y_i$ is the sum of response values for all treated individuals and similarly S_0 for the untreated. Note that quantities n_1, $n_0 = n - n_1$, S_1, and $S_0 = \sum_i y_i - S_1$ do not involve the split variable Δ_i and can be computed beforehand. It follows that $\bar{y}_{lt} = S_{lt}/n_{lt} \quad \approx \quad \tilde{S}_{lt}/\tilde{n}_{lt} = \tilde{y}_{lt}$ for $l = 0,1$ and $t = \{L,R\}$. Next, bringing $(\tilde{n}_{lt}, \tilde{y}_{lt})$ into (3) gives its approximation $\tilde{\sigma}^2$. Finally, plugging all the approximated quantities into $Q(c)$ in (2) yields

$$\tilde{Q}(c) = \frac{\{(\tilde{y}_{1L} - \tilde{y}_{0L}) - (\tilde{y}_{1R} - \tilde{y}_{0R})\}^2}{\tilde{\sigma}^2 (1/\tilde{n}_{1L} + 1/\tilde{n}_{0L} + 1/\tilde{n}_{1R} + 1/\tilde{n}_{0R})}. \tag{5}$$

Now, $\tilde{Q}(c)$ is a smooth objective function for c only and can be directly maximized to obtain the best cutoff point $\hat{c}$.

Besides c, there is a scale parameter a involved in $\tilde{Q}(c)$ given by (5). As shown by simulation in Section 3, the performance of the SSS method is quite robust with respect to the choice of a for a wide range of values. Thus, a can be fixed a priori. To do so, we standardize the predictor $x_{ij} := (x_{ij} - \bar{x}_j)/\hat{\sigma}_j$, where $(\bar{x}_j, \hat{\sigma}_j)$ denote the sample mean and standard deviation (SD) of variable X_j, respectively. For standardized covariates, we recommend fixing a at a value in [10, 50]. With fixed a, the best cutoff point $\hat{c}$ can be obtained by maximizing $\tilde{Q}(c)$ with respect to c and then transformed back to the original data scale for interpretability. This 1-dimensional smooth optimization problem can be conveniently solved by many standard optimization routines. We use the Brent[21] method available in the R[22] function `optimize` in our implementation. Given the nonconcave nature of the maximization problem, techniques such as multistart or partitioning the search range may be used in combination with Brent method as further efforts to locate the global optimum. However, as shown in our numerical studies, a plain application of Brent method works quite effectively in estimating c.

Smooth sigmoid surrogate smooths out local spikes in GS splitting measures and hence helps identify the true cutoff point; see Figure 1B for one example. Additional simulation studies in Section II.1 in the Supporting Web Materials show that SSS outperforms GS in estimating c if there exists a true cutoff point. Another main advantage of SSS over GS is computational efficiency. The following proposition provides an asymptotic quantification of the computational complexity involved in GS and SSS splitting.

> **Proposition 1.** *Consider a typical data set of size* n *in the IT setting, where both GS and SSS are used to find the best cutoff point* $\hat{c}$ *for a continuous predictor* X *with* $O(n)$ *distinct values. In terms of computation complexity, GS is at best* $O\{\ln(n)n\}$ *with the updating scheme and* $O(n^2)$ *without the updating scheme. Comparatively, SSS is* $O(kn)$ *with* k *being the number of iterations in Brent method.*

A proof of Proposition 1 is relegated to the Supporting Information. Implementation of tree methods benefits from incremental updating.[2,23] We note that the GS splitting with updating is commonly mistaken to be of order $O(n)$. Updating the IT splitting statistic entails sorting the Y values according to the X values within each treatment group. It turns out that this sorting step would dominate the algorithm in complexity asymptotically with a rate of $O\{\ln(n)n\}$. Comparatively, SSS depends on the number of iterations in Brent method, k. Although the number of iterations is affected by the convergence criterion and the desired accuracy, k is generally small since Brent method has guaranteed convergence at a superlinear rate. Based on our numerical experience, k rarely gets over 15 even for large n. In other words, the $O(kn)$ rate for SSS essentially amounts to the linear rate $O(n)$. A empirical comparison of computing time between SSS and GS can be found in Section II.1 in the Supporting Web Materials.

2.2 | Estimating ITE via RFIT

The RFIT follows the standard paradigm of RF.[13] Take a bootstrap sample $\mathcal{D}_b$ from data $\mathcal{D}$ and construct an IT $\mathcal{T}_b$ using $\mathcal{D}_b$. To split a node, a subset of m covariates are randomly selected, and the optimal split for each covariate is identified and compared to determine the best split of the data. This step is iterated until a large tree $\mathcal{T}_b$ is grown. Each terminal node τ in $\mathcal{T}_b$ is summarized by an estimated treatment effect $\hat{\delta}_\tau$, which is simply the difference in mean response between treated and untreated individuals falling into τ, ie,

$$\hat{\delta}_\tau = \sum_{i:\ \mathbf{x}_i \in \mathcal{D}_b \cap \tau} \left\{ \frac{T_i y_i}{n_{1\tau}} - \frac{(1-T_i) y_i}{n_{0\tau}} \right\},$$

where $n_{1\tau} = \sum_{i:\mathbf{x}_i \in \mathcal{D}_b \cap \tau} T_i$ is the number of treated individuals in $\mathcal{D}_b$ that fall into τ and $n_{0\tau}$ for the untreated.

The entire tree construction procedure is then repeated on B bootstrap samples, which results in a sequence of bootstrap trees $\{\mathcal{T}_b : b = 1, 2, \cdots, B\}$. An individual with covariate vector $\mathbf{x}$ would fall into 1 and only 1 terminal node $\tau_b(\mathbf{x})$ of $\mathcal{T}_b$. Denoting $\hat{\delta}_b(\mathbf{x}) = \hat{\delta}_{\tau_b(\mathbf{x})}$, the ITE for this individual can then be estimated as

$$\hat{\delta}(\mathbf{x}) = \frac{1}{B} \sum_{i=1}^{B} \hat{\delta}_b(\mathbf{x}). \tag{6}$$

Efron[34,14] discusses methods for computing SEs for bootstrap-based estimators and advocates the use of IJ as a general approach. Infinitesimal jackknife is found preferable in RFs, as further explored by Wager et al.[24] Proposition 2 applies the IJ method to obtain a SE formula for estimated ITE $\hat{\delta}(\mathbf{x})$. Its proof is outlined in the Supporting Information.

Proposition 2. *The IJ estimate of variance of* $\hat{\delta}(\mathbf{x})$ *is given by*

$$\hat{V} = \sum_{i=1}^{n} \bar{Z}_i^2, \tag{7}$$

where $\bar{Z}_i = \sum_{b=1}^{B} Z_{bi}/B$ *and* $Z_{bi} = (N_{bi} - 1)\{\hat{\delta}_b(\mathbf{x}) - \hat{\delta}(\mathbf{x})\}$ *with* $\mathrm{N_{bi}}$ *being the number of times that the* i*th observation appears in the* b*th bootstrap sample. In other words, the quantity* $\bar{Z}_i$ *is the bootstrap covariance between* $\mathrm{N_{bi}}$ *and* $\hat{\delta}_{\mathrm{b}}(\mathbf{x})$*. In practice,* $\hat{V}$ *is biased upwards, especially for a small or moderate B. A bias-corrected version is given by*

$$\hat{V}_c = \hat{V} - \frac{1}{B^2} \sum_{i=1}^{n} \sum_{b=1}^{B} (Z_{bi} - \bar{Z}_i)^2. \tag{8}$$

Further assuming approximate independence of N_{bi} *and* $\hat{\delta}_b(\mathbf{x})$*, another bias-corrected version is given by*

$$\hat{V}_c = \hat{V} - \frac{n-1}{B^2} \sum_{b=1}^{B} \{\hat{\delta}_b(\mathbf{x}) - \hat{\delta}(\mathbf{x})\}^2, \tag{9}$$

which is easier to compute than (8).

The validity of these SE formulas will be investigated by simulation in Section 3. The bias-corrected SE formulas in (8) and (9) generally yield very similar results, both outperforming the uncorrected version (7). Note that computing (8) entails evaluation of the matrix $\mathbf{Z} = (Z_{bi})$ at each different $\mathbf{x}$. Therefore, the SE given in (9) is recommended for its enhanced computational efficiency.

2.3 | Comparison with SR

Under the potential outcome framework, SR is an intuitive approach for estimating $\delta(\mathbf{x})$.[4,12] Separate regression builds a model for $\mu_1(\mathbf{x}) = E(Y_1|\mathbf{X} = \mathbf{x})$ based on data of treated individuals only. This step essentially involves predictive modeling of the observed response Y on the covariates $\mathbf{X}$ using the treated group data; RFs[13] can be used for this purpose. Similarly, a model for $\mu_0(\mathbf{x}) = E(Y_0|\mathbf{X} = \mathbf{x})$ is built using data of untreated individuals only. For an individual with covariate vector $\mathbf{x}$, both models are applied to each individual to predict his or her mean potential outcomes. Let $\hat{\mu}_0(\mathbf{x})$ and $\hat{\mu}_1(\mathbf{x})$ denote the resultant estimates of $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$, respectively. Individualized treatment effects can be estimated as

$$\tilde{\delta}(\mathbf{x}) = \hat{\mu}_1(\mathbf{x}) - \hat{\mu}_0(\mathbf{x}). \tag{10}$$

It is worth noting that it is tempting to use the observed response Y of a treated (untreated) individual as an estimate for μ_1(or μ_0) directly. But this is not a good idea due to the potentially inflated variance.

We argue that RFIT is superior to SR, mainly because RFIT works on a simpler problem. To explain, consider the model form $Y = \mu_0(\mathbf{x}) + T\delta(\mathbf{x}) + \varepsilon$, where $\mu_1(\mathbf{x}) = \mu_0(\mathbf{x}) + \delta(\mathbf{x})$. Functions $\mu_0(\mathbf{x})$ and $\delta(\mathbf{x})$ may involve different sets of covariates. In the clinical setting, covariates showing up in $\mu_0(\mathbf{x})$ are called prognostic factors, while covariates showing up in $\delta(\mathbf{x})$ are called predictive factors.[11] In other words, predictive factors interact with the treatment and hence cause differential treatment effects. In SR, both $\mu_1(\mathbf{x})$ and $\mu_0(\mathbf{x})$ have to be estimated in order to estimate the difference $\delta(\mathbf{x})$; thus, it must take both prognostic and predictive factors into consideration. Comparatively, RFIT estimates $\delta(\mathbf{x})$ directly by focusing on predictive factors only. This is because a prognostic factor does not cause a difference in differences, referring to the splitting statistic in (2) for RFIT. In the following, we introduce a performance measure for RFIT and SR in estimating ITE $\delta(\mathbf{x})$, and a theoretical understanding of the measure is attempted.

Both RFIT and SR take the bootstrap-based ensemble-learning approach. The ITE estimates $\hat{\delta}(\mathbf{x})$ in (6) and $\tilde{\delta}(\mathbf{x})$ in (10) involve randomness owing to bootstrap resampling, the current data $\mathcal{D}$, and the point $\mathbf{x}$ at which the estimation is made. To compare RFIT with SR, we consider an average mean-squares error (AMSE) measure defined by

$$\text{AMSE} = E_{\mathbf{X},\mathcal{D},\mathcal{B}}\{\hat{\delta}(\mathbf{X}) - \delta(\mathbf{X})\}^2, \tag{11}$$

where the expectation is taken with respect to the bootstrap distribution $\mathcal{B}$ given the current data $\mathcal{D}$, the sampling distribution of data $\mathcal{D}$, and then the distribution of $\mathbf{X}$.

Define

$$\bar{\delta}(\mathbf{x};\mathcal{D}) = E_{\mathcal{B}}\{\hat{\delta}(\mathbf{x})\}, \quad \text{and} \quad \bar{\delta}(\mathbf{x}) = E_{\mathcal{D}}\{\bar{\delta}(\mathbf{x};\mathcal{D})\}, \tag{12}$$

where $\bar{\delta}(\mathbf{x};\mathcal{D})$ is the RFIT estimate of $\delta(\mathbf{x})$ obtained with perfect bootstrap or $B \to \infty$ and $\bar{\delta}(\mathbf{x})$ is the perfect bootstrap RFIT estimate if, furthermore, we are allowed to recollect data $\mathcal{D}$ freely. Similarly, we define $\{\bar{\mu}_0(\mathbf{x};\mathcal{D}), \bar{\mu}_0(\mathbf{x})\}$ on the basis of $\hat{\mu}_0(\mathbf{x})$ and $\{\bar{\mu}_1(\mathbf{x};\mathcal{D}), \bar{\mu}_1(\mathbf{x})\}$ on the basis of $\hat{\mu}_1(\mathbf{x})$ in SR. Proposition 3 provides a decomposition of the AMSE for the ITE estimate $\hat{\delta}(\mathbf{x})$ by RFIT and for $\tilde{\delta}(\mathbf{x})$ by SR.

Proposition 3. *For the RFIT estimate $\hat{\delta}(\mathbf{x})$ in (6),*

$$AMSE = E_{\mathbf{X},\mathcal{D},\mathcal{B}}\left\{\hat{\delta}(\mathbf{X}) - \bar{\delta}(\mathbf{X};\mathcal{D})\right\}^2 + E_{\mathbf{X},\mathcal{D}}\{\bar{\delta}(\mathbf{X};\mathcal{D}) - \bar{\delta}(\mathbf{X})\}^2 + E_{\mathbf{X}}\{\bar{\delta}(\mathbf{X}) - \delta(\mathbf{X})\}^2. \tag{13}$$

For the SR estimate $\tilde{\delta}(\mathbf{x})$ in (10),

$$\begin{aligned} AMSE \quad &= E_{\mathbf{X},\mathcal{D},\mathcal{B}}\{\hat{\mu}_1(\mathbf{X}) - \bar{\mu}_1(\mathbf{X};\mathcal{D})\}^2 + E_{\mathbf{X},\mathcal{D}}\{\bar{\mu}_1(\mathbf{X};\mathcal{D}) - \bar{\mu}_1(\mathbf{X})\}^2 + E_{\mathbf{X}}\{\bar{\mu}_1(\mathbf{X}) - \mu_1(\mathbf{X})\}^2 \\ &\quad + E_{\mathbf{X},\mathcal{D},\mathcal{B}}\{\hat{\mu}_0(\mathbf{X}) - \bar{\mu}_0(\mathbf{X};\mathcal{D})\}^2 + E_{\mathbf{X},\mathcal{D}}\{\bar{\mu}_0(\mathbf{X};\mathcal{D}) - \bar{\mu}_0(\mathbf{X})\}^2 + E_{\mathbf{X}}\{\bar{\mu}_0(\mathbf{X}) - \mu_0(\mathbf{X})\}^2 \\ &\quad - 2E_{\mathbf{X}}[\{\bar{\mu}_1(\mathbf{X}) - \mu_1(\mathbf{X})\}\{\bar{\mu}_0(\mathbf{X}) - \mu_0(\mathbf{X})\}] \qquad . \end{aligned} \tag{14}$$

The first term of the AMSE in (13) corresponds to Monte Carlo variation resulting from using a finite number of B bootstrap samples. The second term represents the sampling variation owing to the lack of an endless supply of training data in reality. The third term is the bias. An analogous interpretation applies to the terms in (14), yet with an additional covariance term $-2E_{\mathbf{X}}[\{\bar{\mu}_1(\mathbf{X}) - \mu_1(\mathbf{X})\}\{\bar{\mu}_0(\mathbf{X}) - \mu_0(\mathbf{X})\}]$. It is worth noting that such a decomposition holds true for general bootstrap-based ensemble predictions.

Ensemble learners such as RF and bagging aim for variance reduction by imitating the endless supply of replicate data via bootstrap resampling. This is why we have the additional decomposition

$$E_{\mathbf{X},\mathcal{D},\mathcal{B}}\{\hat{\delta}(\mathbf{X}) - \bar{\delta}(\mathbf{X})\}^2 = E_{\mathbf{X},\mathcal{D},\mathcal{B}}\left\{\hat{\delta}(\mathbf{X}) - \bar{\delta}(\mathbf{X};\mathcal{D})\right\}^2 + E_{\mathbf{X},\mathcal{D}}\{\bar{\delta}(\mathbf{X};\mathcal{D}) - \bar{\delta}(\mathbf{X})\}^2$$

in (13), similarly for $\hat{\mu}_1(\mathbf{X})$ and $\hat{\mu}_0(\mathbf{X})$ in (14). However, ensemble learning has little effect on the bias term

$E_{\mathbf{X}}\{\bar{\delta}(\mathbf{X})-\delta(\mathbf{X})\}^2$ in (13), similarly for the 2 bias terms in (14) as well as the covariance term $-2E_{\mathbf{X}}[\{\bar{\mu}_1(\mathbf{X})-\mu_1(\mathbf{X})\}\{\bar{\mu}_0(\mathbf{X})-\mu_0(\mathbf{X})\}]$. The bias problem for ensemble learners such as RFs has been noted by Breiman[25] and others. From another perspective, RF facilitates a smoothing procedure by averaging data over an adaptive neighborhood; as a result, it cuts the hill and fills the valley.

While both RFIT and SR would suffer from certain bias, the AMSE in SR tends to be larger than that of RFIT in general, as we shall demonstrate numerically in Section 3. Numerical evidence shows that SR is more prone to the bias problem because it tends to underestimate a large ITE and overestimate a small ITE. In fact, such a bias also has an effect on the last covariance term in (14). A large ITE $\delta(\mathbf{x})$ occurs when $\mu_1(\mathbf{x})$ is large and/or $\mu_0(\mathbf{x})$ is small. The smoothing effect yields $\bar{\mu}_1(\mathbf{X})-\mu_1(\mathbf{X})<0$ with cut hills and $\bar{\mu}_0(\mathbf{X})-\mu_0(\mathbf{X})>0$ with filled valleys. Thus, $\{\bar{\mu}_1(\mathbf{X})-\mu_1(\mathbf{X})\}\{\bar{\mu}_0(\mathbf{X})-\mu_0(\mathbf{X})\}$ tends to be negative. A similar observation holds for a small ITE, which occurs when $\mu_1(\mathbf{X})$ is small and/or $\mu_0(\mathbf{X})$ is large. As a result, the last term in (14) tends to be negative, leading to a more inflated AMSE for SR.

3 | SIMULATION STUDIES

This section presents results from simulation studies designed to compare RFIT with other methods in estimating the ITEs. We also investigate the SE formulas for the ITE estimates by RFIT.

3.1 | Comparison in estimating ITE

To compare RFIT with other methods, we generate data by the following scheme. First, simulate 5 (p = 5) predictors $x_j \sim$ uniform[0, 1] for j = 1,...,5 with a common correlation ρ. This is achieved by simulating multivariate normal vectors with the common correlation $\rho' = 2\sin(\rho\pi/6)$ and applying the probability integral transform.[26] Two correlations $\rho \in \{0, 0.5\}$ are considered. Then, we generate $y_0' = \mu_0(\mathbf{x})+\alpha+\varepsilon_0$ with a nonlinear polynomial mean function $\mu_0(\mathbf{x}) = -2-2x_1-2x_2^2+2x_3^3$, and α and ε_0 independently follow a $\mathcal{N}(0,1)$ distribution. Next, we generate $y_1' = \mu_1(\mathbf{x})+\alpha+\varepsilon_1$, where $\mu_1(\mathbf{x}) = \mu_0(\mathbf{x}) + \delta(\mathbf{x})$ and $\varepsilon_1 \sim \mathcal{N}(0,1)$ is independent of both α and ε_0. The random effect term α is introduced to mimic some common characteristics shared by repeated measures Y_0' and Y_1' taken from the same subject. The unit-level effect $Y_1'-Y_0'$ equals $\delta(\mathbf{x}) + (\varepsilon_1 - \varepsilon_0)$, where $(\varepsilon_1-\varepsilon_0)$ represents additional random errors that cannot be accounted for by covariates $\mathbf{x}$. Four models (I)-(IV) are considered for the ITE $\delta(\mathbf{x})$, as tabulated below:

Model	Form	var(μ_1(X))	var(δ(X))	var(α+ε)
I	$\delta(\mathbf{x})=5$	1.009	0.000	1.996
II	$\delta(\mathbf{x})=-5+5x_1+5x_2$	1.017	4.183	2.002
III	$\delta(\mathbf{x})=-5+5x_4+5x_5$	1.002	4.201	1.998
III	$\delta(\mathbf{x})=-2+2I(x_1\leq0.5)+2I(x_2\leq0.5)I(x_3\leq0.5)$	1.014	1.764	1.996
IV	$\delta(\mathbf{x}) = -6+0.1\exp(4x_1)+4\exp\{20(x_2-0.5)\}+3x_3+2x_4+x_5$	1.012	6.316	1.999
V	$\delta(\mathbf{x}) = -10+10\sin(\pi x_1 x_2)+20(x_3-0.5)^2+10x_4+5x_5$	1.009	23.837	1.990

Model I is a null model where the treatment does not have heterogeneous effects. Both models II and III exemplify a linear ITE, but X_1 and X_2 are both prognostic and predictive in model II, while model III contains different sets of covariates as prognostic factors (X_1, X_2, X_3) and predictive factors (X_4, X_5). Model III represents a tree-structured model. Models IV and V are 2 nonlinear models derived from Friedman.[27] Finally, we simulate the randomized treatment assignment variable T independently from a Bernoulli(0.5) distribution and hence the observed response $y = Ty_1' + (1-T)y_0'$. Also provided in the above table are the empirical variances (based on 100 000 realizations) of the additive effect $\mu_0(\mathbf{X})$, ITE $\delta(\mathbf{X})$, and the error term α+ε. These variance values inform us about the signal-to-noise ratio in each model.

For each training data set $\mathcal{D}$, 4 methods are used to learn a model on ITE: single IT analysis, SIDES, RFIT, and SR. In IT,[3] B = 30 bootstrap samples are used to determine the final tree structure. The default setting is used in SIDES. A total of B = 500 bootstrap samples are taken in RFIT and SR. In RFIT, we set a = 10 in SSS splitting. The number m of randomly chosen covariates to examine at each node splitting is set as $m = 2$ in both RFIT and SR. To evaluate performance, a test sample $\mathcal{D}'$ of size $n' = 2000$ is generated beforehand. The ITE models trained with different methods in each simulation are applied to estimate the ITE for $\mathcal{D}'$, and a mean-squared error (MSE) measure $\text{MSE} = \sum_{i=1}^{n'} \{\hat{\delta}(\mathbf{x}_i) - \delta(\mathbf{x}_i)\}^2 / n'$ is computed. Two sample sizes $n \in \{100, 500\}$ are considered for the training data $\mathcal{D}$, and a total of 100 simulation runs are used for each simulation setting.

Figure 2 presents parallel boxplots of the MSE measures when the covariates $\{X_1, \dots, X_5\}$ are independent ($\rho = 0$). The averaged MSE over 100 simulation runs is highlighted with blue bars in each boxplot, corresponding to estimates of the AMSE in (11). It can be seen that SIDES perform poorly in all scenarios. SIDES is not suitable for the task of ITE estimation since it essentially splits data into at most 2 groups: one subgroup containing individuals with enhanced treatment effects and the other group formed by the remaining individuals. Both RFIT and SR outperform IT for a great deal except the null case, model I, indicating an advantage of ensemble-learning methods over the single tree analysis. In comparison with SR, RFIT tends to have smaller MSE values consistently in nearly all scenarios except model I, where SR slightly outperforms RFIT. In this null model, RFIT forces superfluous partitions, while SR mainly accounts for the effects of prognostic factors. Again, the superiority of RFIT in nonnull cases can be explained by the fact that it works on an easier task than SR by examining predictive factors only. The amount of outperformance varies, depending on factors such as the sample size, the signal strength, and the level of nonlinearity. Some additional results are relegated to the Supporting Information. Numerical insight into the bias problem

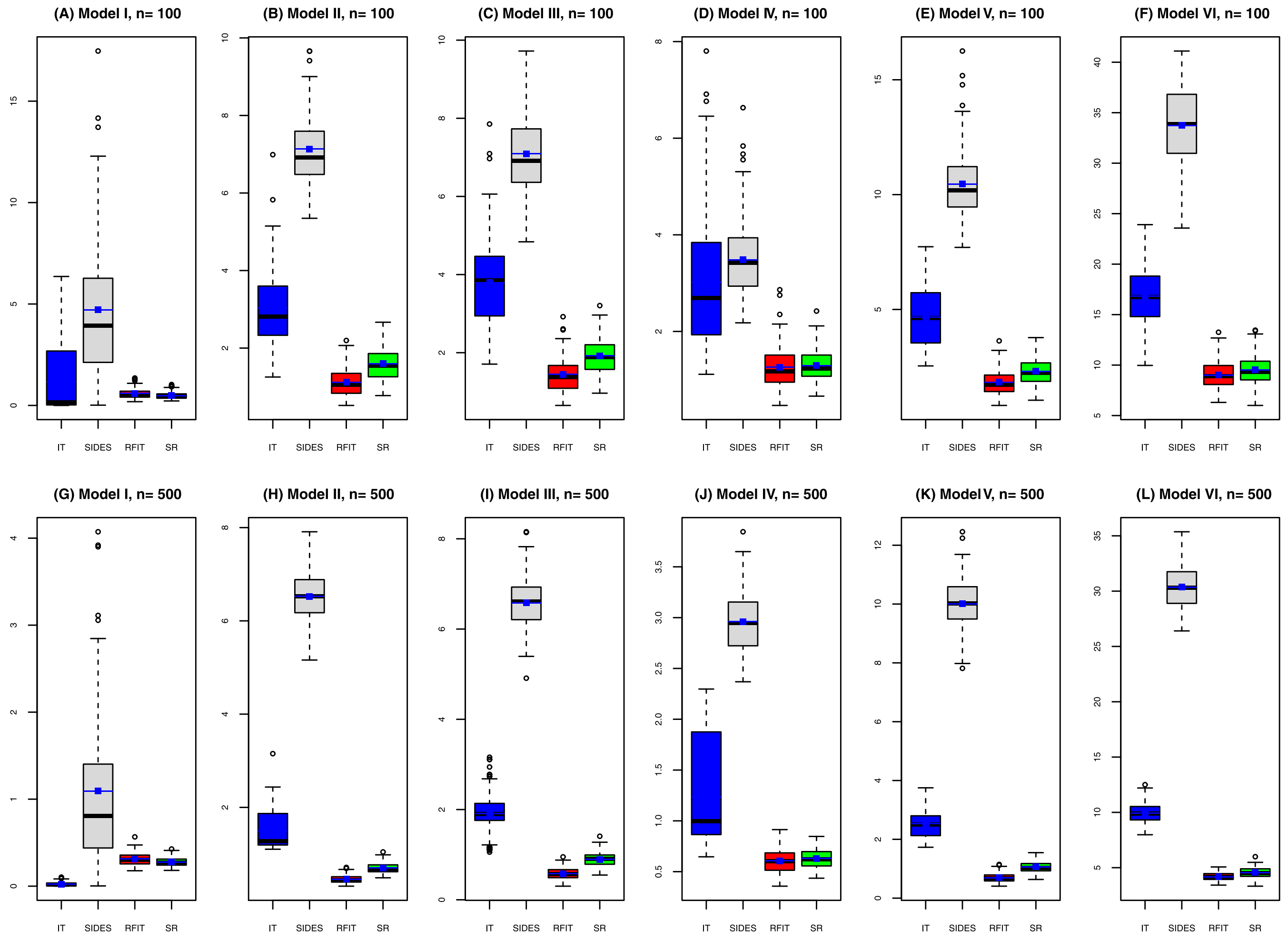

FIGURE 2 Comparison of interaction tree (IT), subgroup identification based on differential effect search (SIDES), random forests of interaction trees (RFIT), and random forest (SR) in estimating individualized treatment effect: the independent ($\rho = 0$) case. Parallel boxplots of mean-squared error (MSE) values are based on a test sample of size n'=2000 with 100 simulation runs. The blue middle bar indicates the average of MSE measures [Colour figure can be viewed at wileyonlinelibrary.com]

is provided by plotting the averaged ITE estimates $\hat{\delta}(\mathbf{x})$ versus the actual ITE $\delta(\mathbf{x})$. Having correlated covariates (with $\rho = 0.5$) does not seem to affect the results much.

3.2 | SE formulas

To investigate the validity and performance of the SE formulas, we generated training data sets of size n=500 from model III and 1 test data set $\mathcal{D}'$ of size n'=50. For each training data set $\mathcal{D}$, RFIT is trained with B=2000 bootstrap samples and applied to estimate ITE for each observation in $\mathcal{D}'$, together with SEs. We repeat the experiment for 200 simulation runs. At the end of the experiment, we have 200 predicted ITE $\hat{\delta}$ for each observation in $\mathcal{D}'$, together with 200 SEs. We compute the SD of these ITE estimates $\hat{\delta}$ and average the SE values. If the SE formula works well, the averaged SE values should be close to their corresponding SD values.

Figure 3 plots the average SE versus SD for each observation in the test sample $\mathcal{D}'$. It can be seen that the uncorrected SEs are overly conservative. After bias correction, the average SEs become reasonably close to the SD values. The bias-corrected SE presented here is computed from (9). The other version (8) that is somewhat harder to compute provides very similar results, which have been omitted from the plot.

We experimented with other models in section 3.1, and similar results were obtained. One issue pertains to the number B of bootstrap samples needed. According to Efron,[14] a large B, eg, $B = 2000$, is needed to guarantee the validity of IJ-based SEs. We experimented with different B values. Generally speaking, ITE estimation stabilizes quickly even with a small B, eg, $B = 100$; however, negative values may frequently occur for the bias-corrected variance estimates in both (8) and (9) when B is small or moderate, eg, $B = 500$. Thus, a large number of bootstrap samples are needed to have sensible results for the SE formula.

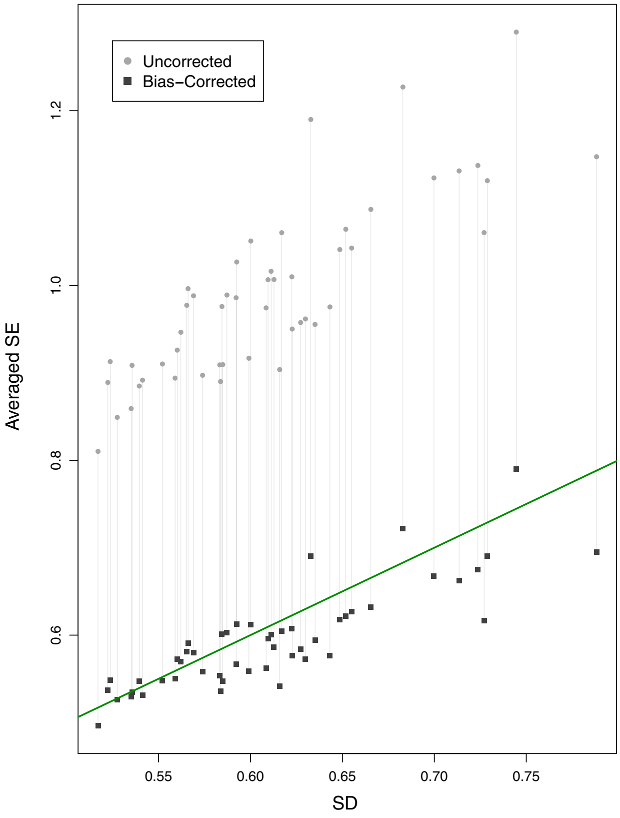

FIGURE 3 Plot of averaged standard errors (SE) versus sample standard deviation (SD) of predicted individualized treatment effect $\hat{\delta}(\mathbf{x})$ for n'=50 observations in a test sample. The SDs are computed based on 200 simulation runs, while the SEs are averaged over the 200 runs. In each simulation run, a training sample of size n=500 is generated from model III, and a bootstrap size $B = 2000$ is used to build random forests of interaction trees. The bias-corrected and uncorrected SE averages for the same observation are connected by a gray line segment. The reference line in green is $y = x$ [Colour figure can be viewed at wileyonlinelibrary.com]

4 | APPLICATION: ACUPUNCTURE TRIAL

For further illustration of RFIT, we consider data collected from an acupuncture headache trial,[28,29] available at https://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-7-15.

In this randomized study, 401 patients with chronic headache, predominantly migraine, were randomly assigned either to receive up to 12 acupuncture treatments over 3 months or to a control intervention offering usual care. Among many other measurements, the primary endpoint of the trial is the change in headache severity score from baseline to 12 months since study entry. The acupuncture treatment was concluded effective overall in bringing down the headache score significantly more than the control group. More details of the trial and its results are reported in Vickers et al.[28]

To apply RFIT, we consider only the 301 participants who completed the trial. The response variable is taken as the difference in headache severity score between baseline and 12 months, whereas the score at baseline is treated as a covariate. There are 3 subjects with some missing data, which are imputed with RFs (see R pacakge missForest[30]). A total of 18 covariates are included in the analysis; these are demographic, medical, or treatment variables measured at baseline. See Table 1 for a brief variable description.

A total of B=5000 trees are used to build RFIT, where the scale parameter a is set as a=10 in SSS splitting. Individualized treatment effect is estimated for each individual in the same data set, and the IJ-based SE with bias correction is also computed. Figure 4A provides a bar plot of the estimated ITE, plus and minus 1 SE, sorted by ITE. It can be seen that most (76.85%) ITEs are above 0, indicating the effectiveness of acupuncture in achieving a greater reduction in headache severity score from baseline to month 12 in comparison with the control group. Overall speaking, the treatment effects in this trial show certain heterogeneity, but not by much. It is interesting to note that the averaged ITE is 3.9. Comparatively, the unadjusted effect of acupuncture (ie, mean difference between acupuncture and control groups in headache severity score change from baseline to month 12) is estimated as 6.5, while the adjusted effect from ANCOVA is 4.6, as reported in Vickers et al.[28, Table 2] Figure 4 also shows many individuals, for whom the acupuncture

TABLE 1 Variable description for the headache data

Name	Description
`id`	Patient ID code
`diff`	Difference in headache severity score between one year follow-up and baseline, i.e., (`pk5` - `pk1`)
`group`	Randomized treatment assignment: 0 is control; 1 is acupuncture
`age`	Age
`sex`	sex: 0 male; 1 female
`migraine`	Migraine: 0 No and 1 Yes
`chronicity`	Chronicity
`pk1`	Severity score at baseline
`f1`	Headache frequency at baseline
`pf1`	Baseline SF36 (36-Item Short Form Health Survey) physical functioning
`rlp1`	Baseline SF36 role limitation physical
`rle1`	Baseline SF36 role limitation emotional
`ef1`	Baseline SF36 energy fatigue
`ewb1`	Baseline SF36 emotional well being
`sf1`	Baseline SF36 social functioning
`p1`	Baseline SF36 pain
`gen1`	Baseline SF36 general health
`hc1`	Baseline SF36 health change
`painmedspk1`	Medication Quantification Scale (MQS) at baseline
`prophmqs1`	MQS of prophylactic medication at baseline
`allmedsbaseline`	Total MQS at baseline

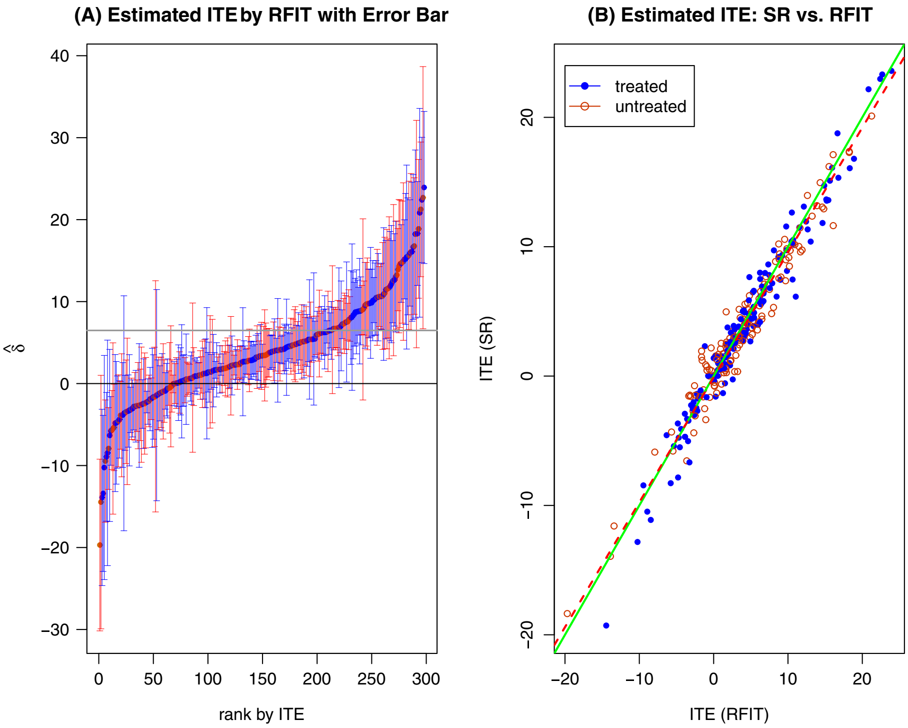

FIGURE 4 Random forests of interaction trees (RFIT) analysis of the headache data: A, the error bar plot of the estimated individualized treatment effect (ITE) ± standard error and B, estimated ITE by separate regression (SR) versus estimated ITE by RFIT. In panel A, individuals are ranked by estimated ITE. The gray horizontal line indicates the unadjusted average treatment effect 6.5, ie, the mean difference in headache severity score change from baseline to month 12. In panel B, the solid green line is the reference line $y = x$, while the dashed red line corresponds to the least-squares regression line [Colour figure can be viewed at wileyonlinelibrary.com]

treatment did not help much. Two individuals, the 44th (with patient ID 222) and the 224th (with patient ID 630), are noteworthy. Both are female patients aged 60 and 58, suffering migraine headaches and being assigned to the control group, but they surprisingly achieved a reduction of 36 and 29.75 in headache severity score, respectively. Their initial severity scores are relatively similar as well: 44.25 and 37. Their estimated ITEs turn out to be −14.81 and −9.09, indicating a detrimental effect from acupuncture. Although the performances of these 2 patients are quite unusual relative to the rest of the patients, they may indicate a small subgroup that is worth further investigation. Figure 4B plots the estimated ITE by SR versus the estimated ITE by RIFT. The LS fitted (red dashed) line almost overlaps with the reference (solid green) line $y=x$, indicating that the 2 methods provide similar ITE estimates in this example.

5 | DISCUSSION

We have implemented RFs of ITs to tackle the problem of estimating ITEs. To this end, we have introduced SSS splitting to speed up RFIT and possibly improve its performance. We have also derived an SE for the estimated ITE by applying the IJ method. Altogether, RFIT provides enlightening results for deploying personalized medicine by informing a new patient about the potential effect of the treatment on him or her.

According to our numerical experiments, RFIT outperforms the SR approach for estimating ITE. Separate regression estimates the potential outcomes separately and then takes difference. In RFIT, we group individuals so that those with similar treatment effects are put together and then estimate the treatment effect by taking differences within each group. Comparatively, RFIT focuses on predictive covariates and estimation of ITE directly, while SR has to deal with both prognostic and predictive covariates. Since SR is used as an intermediary step in other causal inference procedures, our method might contribute to their improvement as well.

To conclude, we identify several avenues for future research. First of all, our discussion has been restricted to data from randomized experiments. Assessing treatment effects with data from observational data can be very different, entailing adjustment for potential confounders.[31-33] Secondly, the SE formula provides some assessment for precision in estimating ITE; however, issues such as consistency of RFIT, asymptotic normality of estimated ITE (see comments in Efron[14]), and multiplicity have not been thoroughly addressed as of yet. Thirdly, the current version of RFIT is not free of variable selection bias,[7] and how to address this problem with the SSS approach awaits further investigation. Fourthly, several other features in RFs including variable importance ranking, partial dependence plots, and the proximity matrix[19] have yet to be explored for RFIT. Like RFs, RFIT is essentially a black-box tool for predicting ITE, although the IJ-based SE supplies additional reliability measure. The last direction of future research is closely related to how to extract meaningful interpretations of RFIT. Specifically, the variable importance measure can sort out important effect modifiers of the treatment; the partial dependence plot can depict how a covariate modifies the treatment effect under the intertwined influences of other covariates; the proximity matrix can identify a neighborhood of a future patient in terms of how similarly they react to the treatment. To make these additional features of RF better suitable for ITE assessment, major modifications are needed, which warrants future research.

ACKNOWLEDGMENTS

X.S. was partially supported by NIMHD grant 2G12MD007592 from NIH; L.L. was partially supported by AHRQ grant HS 020263; R.L. was supported in part by NSF grant 1633130. The authors wish to thank the editor (Dr Joel Greenhouse), the associate editor, and 2 anonymous reviewers, whose insightful and constructive comments played an important role in improving the paper.

ORCID

Xiaogang Su http://orcid.org/0000-0002-9642-9412
Lei Liu http://orcid.org/0000-0003-1844-338X

REFERENCES

1. Lipkovich I, Dmitrienko A, D'Agostino RB. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Stat Med*. 2017;36(1):136-196.
2. Breiman L, Friedman J, Olshen R, Stone C. *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group; 1984.
3. Su X, Tsai CL, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. *J Mach Learn Res*. 2009;10:141-158.
4. Foster JC, Taylor JMC, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med*. 2011;30(24):2867-2880.
5. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search (SIDES): a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat Med*. 2011;30(21):2601-2621.
6. Dusseldorp E, van Mechelen I. Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions. *Stat Med*. 2014;33(2):219-237.
7. Loh WY, He X, Man M. A regression tree approach to identifying subgroups with differential treatment effects. *Stat Med*. 2015;34(11):1818-1833.
8. Murphy SA. Optimal dynamic treatment regimes (with discussion). *J R Stat Soc Ser B*. 2003;65(2):331-366.
9. Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E. Estimating optimal treatment regimes from a classification perspective. *STAT*. 2012;1(1):103-114.
10. Laber EB, Zhao Y. Tree-based methods for individualized treatment regimes. *Biometrika*. 2015;102(3):501-514.
11. Ballman KV. Biomarker: predictive or prognostic? *J Clin Oncol*. 2015;33(33):3968-3971.
12. van der Laan M, Polley E, Hubbard A. Super learner. *Stat Appl Genet Mol Biol*. 2007;6(1): Article 25.
13. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
14. Efron B. Estimation and accuracy after model selection (with discussion). *J Am Stat Assoc*. 2014;109(507):991-1007.
15. Neyman J. On the application of probability theory to agricultural experiments. *Essay on Principles 1923, Section 9. Stat Sci*. 1990;5(4):465-472. Translated by Dabrowska DM and Speed TP.
16. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688-701.
17. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *J Am Stat Assoc*. 2005;100(469):322-331.
18. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81(396):945-960.
19. Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002;2(3):18-22.
20. Abadie A. Semiparametric difference-in-differences estimators. *Rev Econ Stud*. 2005;72(1):1-19.

21. Brent R. *Algorithms for Minimization without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall; 1973.

22. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2017. https://www.R-project.org/

23. LeBlanc M, Crowley J. Survival trees by goodness of split. *J Am Stat Assoc*. 1993;88(422):457-467.

24. Wager S, Hastie T, Efron B. Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *J Mach Learn Res*. 2014;15:1625-1651.

25. Breiman L. Using adaptive bagging to debias regressions. Technical Report #547, Department of Statistics University of California at Berkeley; 1999.

26. Gilli M, Maringer D, Schumann E. *Numerical Methods and Optimization in Finance*. Kidlington, Oxford, UK: Elsevier; 2011.

27. Friedman JH. Multivariate adaptive regression splines. *Ann Stat*. 1991;19(1):1-67.

28. Vickers AJ, Rees RW, Zollman CE, et al. Acupuncture for chronic headache in primary care: large, pragmatic, randomised trial. *Br Med J (Prim Care)*. 2004;328:744.

29. Vickers AJ. Whose data set is it anyway? Sharing raw data from randomized trials. *Trials*. 2006;7:15.

30. Stekhoven DJ, Buehlmann P. missForest – nonparametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28:112-118.

31. Su X, Kang J, Fan J, Levine R, Yan X. Facilitating score and causal inference trees for large observational data. *J Mach Learn Res*. 2012;13:2955-2994.

32. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc*. 2017. published online.

33. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.

34. Efron B. *The Jackknife, the Bootstrap and Other Resampling Plans*. *CBMS-NSF Regional Conference Series in Applied Mathematics 38*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM); 1982.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Su X, Peña AT, Liu L, Levine RA. Random forests of interaction trees for estimating individualized treatment effects in randomized trials. *Statistics in Medicine*. 2018;1–14. https://doi.org/10.1002/sim.7660