SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2018, Vol. 27(1) 142–157

DOI: 10.1177/0962280215623981
journals.sagepub.com/home/smm
SAGE

Identification of predicted individual treatment effects in randomized clinical trials

Andrea Lamont,[1] Michael D Lyons,[2] Thomas Jaki,[3] Elizabeth Stuart,[4] Daniel J Feaster,[5] Kukatharmini Tharmaratnam,[6] Daniel Oberski,[7] Hemant Ishwaran,[5] Dawn K Wilson[1] and M Lee Van Horn[8]

Abstract

In most medical research, treatment effectiveness is assessed using the average treatment effect or some version of subgroup analysis. The practice of individualized or precision medicine, however, requires new approaches that predict how an individual will respond to treatment, rather than relying on aggregate measures of effect. In this study, we present a conceptual framework for estimating individual treatment effects, referred to as predicted individual treatment effects. We first apply the predicted individual treatment effect approach to a randomized controlled trial designed to improve behavioral and physical symptoms. Despite trivial average effects of the intervention, we show substantial heterogeneity in predicted individual treatment response using the predicted individual treatment effect approach. The predicted individual treatment effects can be used to predict individuals for whom the intervention may be most effective (or harmful). Next, we conduct a Monte Carlo simulation study to evaluate the accuracy of predicted individual treatment effects. We compare the performance of two methods used to obtain predictions: multiple imputation and non-parametric random decision trees. Results showed that, on average, both predictive methods produced accurate estimates at the individual level; however, the random decision trees tended to underestimate the predicted individual treatment effect for people at the extreme and showed more variability in predictions across repetitions compared to the imputation approach. Limitations and future directions are discussed.

Keywords

Predicted individual treatment effects, heterogeneity in treatment effects, individualized medicine, multiple imputation, random decision trees, random forests, individual predictions

1 Introduction

Understanding and predicting variability in treatment response is an important step for the advancement of individualized approaches to medicine. Yet, the effectiveness of an intervention assessed in a randomized clinical trial is typically measured by the average treatment effect (ATE) or a type of subgroup analysis (e.g. statistical interactions). The ATE (or conditional ATE for subgroup analysis) forms the basis of treatment recommendations for the individual without considering individual characteristics (such as genetic risk, environmental risk exposure or disease expression) that may alter a particular individual's response. Even in the case of cutting-edge, personalized

[1]Department of Psychology, Barnwell College, University of South Carolina, Columbia, USA
[2]Department of Psychology, University of Houston, Houston, USA
[3]Department of Mathematics and Statistics, Lancaster University, Lancaster, UK
[4]Department of Mental Health, Department of Biostatistics, and Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins, Baltimore, MD, USA
[5]Department of Public Health Sciences, Division of Biostatistics, University of Miami, Miami, FL, USA
[6]Lancaster University Fylde College, Lancaster University, Lancaster, UK
[7]Department of Methodology and Statistics, Tilburg University, Tilburg, the Netherlands
[8]Department of Individual, Family and Community Education, University of New Mexico, Albuquerque, NM, USA

Corresponding author:
Andrea Lamont, Barnwell College, University of South Carolina, 1512 Pendleton Street Columbia, SC 29208, USA.
Email: alamont082@gmail.com

protocols, treatment decisions are based on subgroups defined by a few variables (e.g. disease expression, biomarkers, genetic risk),[1–4] which may mask large effect variability. The ideal health-care scenario would be one in which treatment recommendations are based on the individual patient's most likely treatment response, given their biological and environmental uniqueness. While the reliance on aggregate measures is partially justified by long-established concerns over the dangers of multiplicity (false positives) in subgroup analyses,[5–11] the avoidance of false positives has come at the cost of understanding individual heterogeneity in treatment response. We propose that a principled statistical approach that allows for prediction of an individual patient's response to treatment is needed to advance the effectiveness of individualized medicine (Note: We define individualized medicine in this study broadly as the tailoring of interventions to the individual patient. Our definition overlaps with aspects of the fields of precision medicine, personalized medicine, individualized medicine, patient-centered/patient-oriented care and other related fields). Individual-level predictions would provide prognostic information for an individual patient, and allow the clinician and patient to select the treatment option(s) that would maximize benefit and minimize harm.

This paper proposes a framework for estimating individual treatment effects. Based on the principles of causal inference, we define a predicted individual treatment effect (PITE) and compare two different methods for deriving predictions. The PITE approach builds upon existing methods for identifying heterogeneity in treatment effects (HTE) and has direct implications for health-care practice. The structure of this paper proceeds as follows. Section 2 describes the theoretical foundations and methodological literature from which the PITE approach is derived. Section 3 outlines the PITE approach and the predictive models compared in this paper. In Section 4, we demonstrate the utility of the PITE approach using an applied intervention aimed to reduce behavioral and physical symptoms related to depression. In Section 5, we present a Monte Carlo simulation study to validate the PITE approach using two methods for deriving predictions – multiple imputation and random decision trees (RDT). We compare relative performance of each estimator in terms of both accuracy and stability of the estimator. This paper concludes in Section 6 with implications and next steps.

2 Theoretical foundations

Our definition of individual causal effects is rooted in the potential outcomes framework.[12,13] A potential outcome is the theoretical response of each unit under each treatment arm, i.e. the response each unit would have exhibited if they had been assigned to a particular treatment condition. Let Y_0 denote the potential outcome under control and Y_1 the potential outcome under treatment. Assuming that these outcomes are independent of the assignment other patients receive (Stable Unit Treatment Value Assumption; SUTVA),[14,15] individual-level treatment effects are defined as the difference between the responses under the two potential outcomes: $Y_1 - Y_0$. The fundamental problem of causal inference, as described by Neyman,[12] Rubin,[13] Rubin,[15] and Holland,[16] is that both potential outcomes for an individual cannot typically be observed. A single unit is assigned to *only one* treatment condition or the control condition, rendering direct observations in the other condition(s) (the *counterfactual* condition) and, by extension, observed individual causal effects, impossible.

Instead, researchers often focus on ATEs, which under the SUTVA assumption will simply equal the difference in expectations

$$\text{ATE} = \text{E}[Y_1 - Y_0] = \text{E}(Y_1) - \text{E}(Y_0) = \text{E}(Y|\text{treatment}) - \text{E}(Y|\text{control}).$$

Replacing the expectations by observed sample means under treatment and control yields estimates of treatment effects at the group level. While the advantage of this approach is that treatment effects can be estimated in the aggregate, this comes at the cost of minimizing information about individual variability in effects. This is problematic when used for individual treatment decision-making because an individual patient likely differs from the average participant in a clinical trial (i.e. the theoretical participant whose individual effect matches the average) on many biologic, genetic and environmental characteristics that explain heterogeneity in treatment response. When the individual patient differs from the average participant, the ATE can be an (potentially highly) inaccurate estimate of the individual response.

2.1 HTE

There is growing recognition of the importance of individual heterogeneity in treatment response, which has led to a rapid growth of methodological development in the area.[1,17–32] Methods are designed to estimate HTE, while avoiding problems associated with classical subgroup analysis. Modern HTE methods place the expectation for

heterogeneous treatment response at the forefront of analysis and define a statistical model that captures this variability. Proposed approaches for estimating HTE are diverse and include: the use of instrumental variables to capture essential heterogeneity,[17,23] LASSO constraints in a Support Vector Machine,[21] sensitivity analysis,[33] the derivation of treatment bounds to solve identifiability issues,[34,35] regression discontinuity designs,[36] general growth mixture modelling,[37] boosting,[38,39] predictive biomarkers,[40,41] Bayesian additive regression trees (BART),[26] virtual twins[18] and a myriad of other tree-based/recursive partitioning methods[22,32,42–50] and interaction-based methods.[51–54] Generally, the aim of existing methods has been the detection of differential effects across subgroups or the estimation of population effects given known heterogeneity. Most HTE methods have not been validated for individual-level prediction (an exception includes Basu[17]). The focus remains at the level of the subgroup. We argue that the estimation of the individual treatment effect itself is a meaningful and important result, without the aggregation to subgroups. There are important clinical implications for detecting how an individual would respond, independent of the subgroup they belong. Thus, in this study, we build upon and extend existing methods to validate their use at the individual level.

Individual-level predictions are particularly important when estimating a patient's treatment effect in an applied setting. This type of prediction – i.e. predicting responses for out-of-sample individuals or individuals not involved in the original clinical trial – can help bridge the gap between treatment planning in a clinical setting (e.g. *What are the chances that this particular individual will have a positive, null, or iatrogenic response to treatment?*) and the results of clinical trials (e.g. *What is the ATE for a pre-specified subpopulation of interest?*). Individual-level predictions can support data-informed medical decision-making for an individual patient, given that patient's unique constellation of genetic, biological and environmental risk. It is a realistic scenario to imagine the case where a physician has access to medical technologies that input data on the patient (e.g. genetic risk data, environmental risk exposure) to obtain a precise estimate (PITE) of the patient's predicted treatment prognosis, rather than relying on the ATE of a phase III clinical trial for treatment decision-making.

3 Methodology: PITEs

Let the PITE be defined for individual i based on the predicted outcome (Y_i^*) given observed covariates u and treatment condition T as

$$PITE_i = E\big(Y_i^* | U = u_i, Tx = 1\big) - E\big(Y_i^* | U = u_i, Tx = 0\big)$$

which is the difference between the predicted value under treatment and predicted value under control for each individual. The major difference between the PITE approach and the ATE approach is that the PITE approach estimates a treatment effect for each individual, rather than a single effect based on means. There is no single summative estimate; rather, the estimand is the individual-level prediction.

3.1 Predictive models

A strength of the PITE framework is its generality. We suspect that there are multiple methods that can be used to obtain predictions in the PITE framework and that there will be no single predictive methods that is best across all scenarios.[55] In this paper, we contrast two distinct methods for deriving predictions: multiple imputation and RDT. These two methods were selected because: 1) they have been shown to handle a large number of covariates; and 2) the RDT approach has been designed to work with out-of-sample individuals. Also, they come from distinct statistical traditions, rely on their own set of assumptions, and have been applied in rather different ways. In this study, we specifically focus on how well different methods estimate PITEs with a continuous outcome variable.

Importantly, our purpose is not to present an exhaustive comparison of the methodological approaches or present a general mathematical proof of their predictive ability. A complete comparison of the predictive methods outside the context of the PITEs is beyond the scope of this study. Instead, we focus on a side-by-side comparison of the two predictive methods for estimating PITEs. A simple example of how the methods differ within the PITE framework and outside the PITE framework is the case of extreme predicted values. An individual may obtain a rather high prediction under treatment (potentially even at the extremes) when estimated using either predictive method. A comparison of the methods outside the PITE framework would focus on how well the estimator predicted this value at the extreme. Both predictive methods may perform well in this case, but it is not the focus of the PITE. The focus of the PITE is the difference between predicted values under each treatment arm.

Thus, if this same individual also has a high predicted value under control, the PITE estimate will be rather small, perhaps even near-zero. Detecting this effect requires a much more nuanced estimator at the individual level. While the two are clearly related, the PITE has much more practical utility than the predictions under treatment or control separately and warrants explicit empirical attention.

In the following subsections, we provide a brief overview of the predictive methods employed in this study.

3.1.1 Parametric multiple imputation

In a potential outcomes framework, each individual has a potential response under every treatment condition. Yet, there is only one realized response (i.e. the response under the counterfactual condition is never observed). Conceptualized in this way, the unobserved values associated with the counterfactual condition could be considered a missing data problem and handled with modern missing data techniques. Since missingness in the PITE approach is completely due to randomization, missingness is known to be completely at random.

Multiple imputation is a flexible method for handling a large class of missing data problems optimally used when data are assumed to be missing at random.[56,57] Multiple imputation was originally developed to produce parameter estimates and standard errors in the context of missingness[58] and has been expanded to examine subgroup effects.[59] In this paper, we test an extension of the multiple imputation model to obtain individual-level effects. We suggest that the missingness in the counterfactual condition could be addressed by imputing $m > 1$ plausible values based on a large set of observed baseline covariates. The PITE could then be defined as the average difference between values of Y_1 and Y_0 across imputations, for which data are now available for every individual.

We focus on a regression-based parametric model to derive imputations implemented through the *chained equations* algorithm (also referred to as *sequential regression* or *fully conditional specification*) in the imputation process.[60] *Chained equations* is a regression-based approach to imputing missing data that allow the user to specify the distribution of each variable conditional upon other variables in the dataset. Imputations are accomplished in four basic steps, iterated multiple times. First, a simple imputation method (e.g. mean imputation) is performed for every missing value in the dataset. These simple imputations are used as place holders to be improved upon in later steps. Second, for one variable at a time, place holders are set back to missing. This variable becomes the only variable in the model with missingness. Third, the variable with missingness becomes the dependent variable in a regression equation with other variables in the imputation model as predictors. Predictors include a vector of selected covariates and their interactions. The same assumptions that one would make when performing a regression model (e.g. linear, logistic, Poisson) outside of the context of imputation apply. Imputations are drawn from the posterior predictive distribution. Last, the missing values on the dependent variable are replaced by predictions (imputations) from the regression model. These imputed values replace the place holders in all future iterations of the algorithm. These four basic steps are repeated for every variable with missingness in the dataset. The cycling through each variable constitutes one iteration, which ends with all missing values replaced by imputed values. Imputed values are improved through multiple iterations of the procedure. The end result is a single dataset with no missing values. By the end of the iterative cycle, the parameters underlying the imputations (coefficients in the regression models) should have converged to stability, thereby producing similar imputed values across iterations and avoiding any dependence on the order of variable imputation.

3.1.2 RDT

RDT is a recursive partitioning method derived from the fields of machine learning and classification. RDTs fall under the broader class of models known as Classification and Regression Trees (CART) and have been commonly employed by others for finding HTE.[22,32,42–50] CART analysis operates through repeated, binary splits of the population (''the parent node'') into smaller subgroups (''the child nodes''), based on Boolean questions that optimize differences in the outcome. A recursive partitioning algorithm searches the data for the strongest predictors and splits the population based on empirically determined thresholds; for example, is $X \geq \theta_j$?, where X is the value of a predictor variable and θ_j represents an empirically determined threshold value. The splitting procedure continues until a stopping rule is reached (e.g. minimum number of people in the final node, number of splits, variance explained). The final node or ''terminal node'' reflects homogeneous subsets of similar cases, and, by extension, an estimate of an individual's predicted values.

A very simple version of a single decision tree appears in Figure 1. The fictitious population (''parent node'') comprised $N = 1000$ individuals for whom the analyst wanted to divide into subgroups defined by their expected response on hypothetical outcome variable Y. CART analyses work by exploring the data for the most salient predictors of outcome response. In this case, X_i was identified as the most salient predictor, which split the

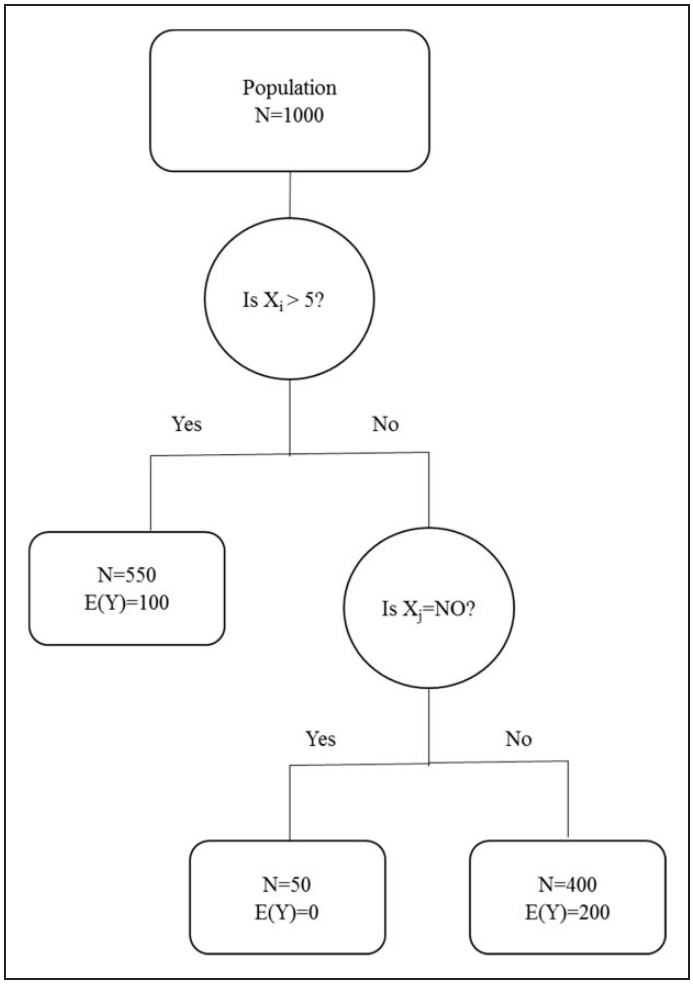

Figure 1. A single hypothetical decision tree.

population in to two groups based on the empirically defined threshold of 5. For individuals with $X_i < 5$, their expected response on the outcome was 100. For the remaining $N = 450$ individuals, more splits of the data were possible in order to create homogenous groups. The model searched the data for another salient predictor and selected $X_j = NO$. Individuals who responded ''NO'' on item X_i had an expected response of 0, whereas those who responded ''YES'' had an expected response of 200. No more splits were available based on a rule specified by the analyst a priori.

There are certain problems associated with implementing a single decision tree, such as that depicted in Figure 1. One issue with single decision trees is that when single trees are grown very large, trees are observed to overfit the data, resulting in low bias but high variance.[55] To circumvent this limitation, RDTs construct a series of decision trees, where each tree is ''grown'' on a bootstrapped sample from the original data. A ''forest'' of many decisions trees is grown and predictions are averaged across trees. Optimal splits are identified from a set of randomly selected variables at each split (or ''node''). An advantage of RDTs is that this is a non-parametric method that does not require the data to meet any assumptions regarding the distribution or specification of the model. As a result, RDTs can fit data with a large number of predictors, data that are non-normally distributed, or data with complex, higher-order interactions.

4 Applied example: Predicting treatment effects in behavioral medicine

The PITE approach for understanding variability in treatment effects has applicability to a broad range of behavioral and physical health outcomes. In this section, we demonstrate the utility of the approach using a program for the prevention of depression among new mothers. Data came from the Building stronger

families/family expectations randomized trial,[61] a federally funded intervention for unmarried, romantically involved adults across eight research sites; only data from Oklahoma (N = 1010) were used due to differences in implementation across sites. Data are publically available for research purposes subjected to application to the Inter-university Consortium for Political and Social Research.[62] At the 15-month impact assessment, there was an overall positive impact of the program such that women in the treatment group experienced significantly less depression than those in the control group. However, the effect size (measured as the standardized mean difference of the impact or treatment effect divided by the standard deviation of the outcome in the control group) of this impact was rather small (Cohen's d = −.22), posing a challenge to the overall clinical value of the program. This is a common scenario in many health interventions, where, despite significant gains, small overall effects suggest that the practical impact of the intervention is limited. Interpretation of this overall effect may misconstrue the true impact of the intervention making it unlikely for an applied practitioner to recommend the program to a patient.

In this demonstration, we tested the utility of the PITE approach for providing predictions for a new set of individuals (out-of-sample individuals). We used the PITE approach to extend the findings of the original trial to determine particular individuals for whom the intervention is most likely to show positive results, despite minimal impact on average. Specifically, we used trial data to estimate predictive models, then used these use these models to predict how a new individual would respond to treatment.

4.1 Methods

From June 2006 through March 2008, 1010 unmarried couples from Oklahoma were randomized into treatment (N = 503) and control (N = 507) conditions. In order to create the conditions of out-of-sample prediction, we randomly removed 250 individuals from the original sample and saved them for out-of-sample estimation. Predictive models were built on the remaining 760 individuals only. Outcome data from the 250 out-of-sample individuals were ignored to create a scenario similar to an applied setting where treatment recommendations are made before outcomes are known.

Seventy-five baseline covariates came from in-person surveys that assessed demographics, education and work, relationship status and satisfaction, parenting, financial support, social support and baseline levels of depression. For all items (except marriage status and number of children), ratings from both mother and father were included. Separate mother and father ratings were included due to inconsistent responses. If items required consistency in order to be valid (e.g. whether the couple was married, number of children), inconsistent responses were set to missing. Maternal depression at 15-month follow-up, the primary outcome variable, was measured using a 12-item version of the Center for epidemiologic studies depression scale (CES-D).[63] Factor scores, created in M*plus* software,[64] were used (standardized) for the observed outcome variable, maternal depression. Missingness on the baseline covariates was handled via single imputation using bootstrap methods, as implemented in the mi package[65] in R version 3.1.3.[66] The same imputed values were used for both treatment and control conditions. We relied on a comprehensive set of diagnostics to determine the quality of imputations (which showed the single imputation method matched the underlying distribution of the covariates well). We acknowledge potential limitations of missingness on baseline covariates. We buffered against potential threats by using different data to generate the predictive model and for predictions (out-of-sample estimation) and by using thorough diagnostics to identify potential problems; however, the best approach for integrating missingness into these models remains an empirical question.

4.1.1 Estimation of predictive models

Imputations were conducted using the package mi[65] in R software version 3.1.3.[66] One hundred imputations were created per repetition under a Bayesian linear regression. Convergence was assessed numerically using the $\hat{R}$ statistic, which measures the mixing of variability in the mean and standard deviation within and between different chains of the imputation.[65] The mean prediction across imputations was taken as the estimated predicted effect (PITE). RDTs were also grown in R using the randomForest®[67] package with all of the default settings, except tree depth. Tree depth was selected to minimize root mean square error (RMSE) in each treatment condition; specifically, RMSE was minimized under treatment when minimum node size equaled 12 and when minimum node size equaled 100 under control. For both the imputation and RDT methods, separate predictive models were estimated under treatment and control. This allowed for the estimation of predicted values under both treatment and control (which are needed to calculate PITE); and, by default, ensured that all one-way interactions between treatment and the covariates were modeled.

4.1.2 Calculation of predicted values

Once predictions were obtained, the PITE was calculated for the set of out-of-sample individuals. We assumed that the sample used to build the predictive model for both treatment and control (the training sample) was representative of the target population (the population we want to predict in). Out-of-sample predicted effects were calculated by multiplying the coefficients estimated by the predictive models to the baseline covariates of out-of-sample individuals. The PITE was taken as the difference of the fitted values under treatment and control. Observed Y values were not used to calculate the PITE in order to minimize overfitting to the data. In preliminary runs, there was significantly more variability in the out-of-sample predictions when observed data were used in the PITE calculation. We identified this inflated variance to be related to an overfitting of observed data (i.e. a predictive model that described random error in the data in addition to the underlying relationship) during model building. This led to an overly deflated correlation between the true treatment effect and PITE. This problem was avoided by using predicted values under both treatment conditions in the calculation of the PITE.

4.2 Results

For both the imputation method and RDT method, we estimated a predictive model on N = 760 in-sample individuals. These can be considered individuals from the original clinical trial. There were no issues of model convergence for either estimator, though the imputation approach took a few hours longer to estimate than the RDT approach, which was completed in minutes. We then calculated predictions for the retained N = 250 out-of-sample individuals, which we treated as out-of-sample individuals for whom we were testing whether the intervention would be effective.

Results are presented in Figure 2. As indicated, the PITE approach was able to detect variability in individual effects, using both the imputation and RDT methods. For certain individuals, the intervention would improve depressive symptoms; for others, participation in the program would not be recommended. We note that, in this demonstration, we illustrate predictions for N = 250 individuals so that heterogeneity in predictions can be seen. In applied practice, a PITE could be estimated for a single individual. The strength and direction of the prediction could then be used during treatment planning for that individual.

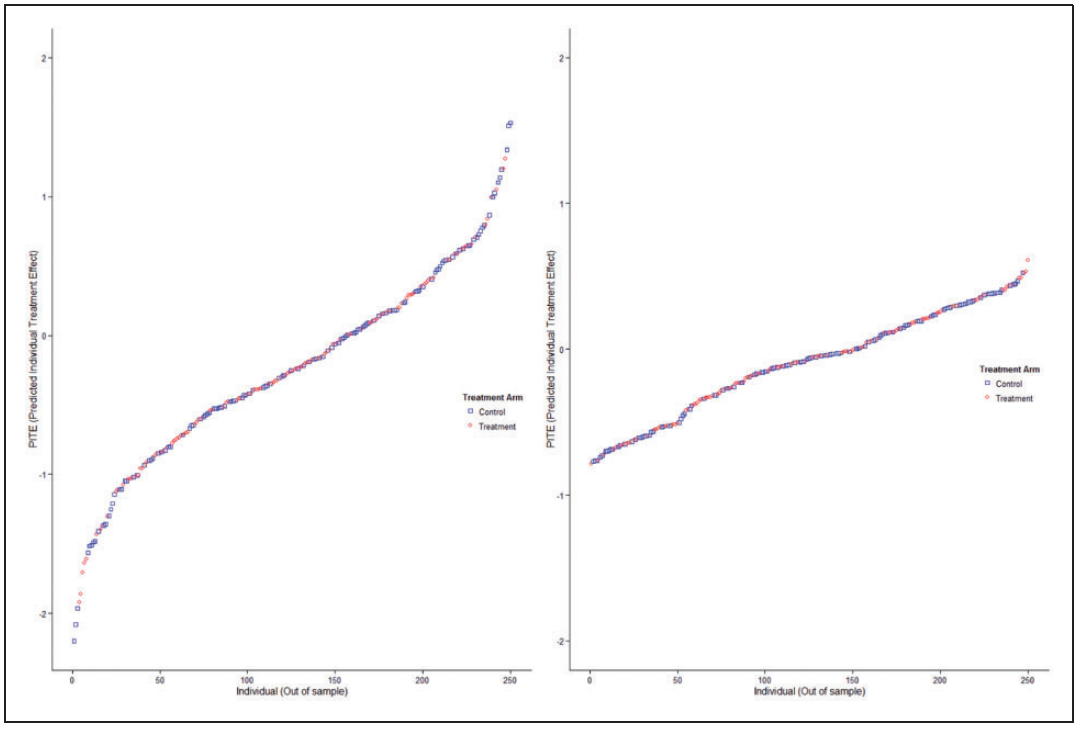

Figure 2. Predicted individual treatment effects for out-of-sample individuals, estimated using the imputation (left) and random decision trees (right) methods on the applied example.

Comparing across predictive methods, RDT approach tended to produce estimates closer to the mean, whereas the imputation approach provided a wider range of PITE values. The simulation study in the next section is intended to test which is providing more accurate and stable predictions.

5 Simulation study

One of the limitations of the applied example is that we do not know the true effects of the individuals in the sample, making the accuracy of the estimates unknown. In this section, we present the results of a Monte Carlo simulation study used to test the quality of these estimated individual effects.

5.1 Methods

Data were generated in R software version 3.1.3.[66] A total of N = 10,000 independent and identically distributed cases were generated. Fifty percent of the cases (N = 5000) were used for the derivation of predictive models, and the remaining N = 5000 cases were reserved for out-of-sample estimation. Cases were randomly assigned to balanced treatment and control groups.

Following the design of the BSF trial, data were generated with the same number of categorical covariates as the applied data. The true value under control was generated as $Y_c \sim N(0, 1)$. The true treatment effect for each individual was linearly related to a set of seven binary baseline covariates, generated using the following equation

$$TE = -1.3(X_1) - 1.2(X_2) - .6(X_3) + .3(X_4) + .5(X_5) + 1.1(X_6) + 1.2(X_7).$$

Binary covariates (generated from a random binomial distribution with the probability of endorsement equal to .5) were used to resemble the design of the data in the motivating example, which included only categorical predictors. No modifications would be necessary to extend to continuous predictors. Effect sizes were selected so that the mean effect was near zero (but not completely symmetric around zero) with individual effect ranging from small to large. Since the PITE approach is designed to detect HTE without pre-specifying the variable(s) responsible for the differential effects, we wanted to include a comprehensive set of potential confounders, most of which end up being nuisance variables. Thus, we additionally included 68 nuisance variables whose coefficients (X_8 through X_{75}) were set to zero. The true response under treatment (Y_t) was defined as $Y_{ti} = Y_{ci} + TE$ for that individual.

One hundred repetitions of the simulated data were generated. Baseline covariates (and by extension the true treatment effects) were set to be the same across repetitions. This established a scenario where the same individual was repeated, allowing for intuitive interpretations about the number of times an individual's predicted value reflects the true treatment effect. Y_c (and by extension Y_t) varied across repetitions. The procedures for estimating the predictive models and for calculating PITEs in this simulation study were identical to those used in the applied example (above).

We note that since this is our first study on the PITE approach, we specifically designed these conditions to be optimal in the sense of a correctly specified model with no major violations of model assumptions (e.g. all effect moderators observed and exchangeability of the in-sample and out-of-sample individuals) and a large sample size. While this may limit generalizability to other scenarios, we see this study as the first study of a larger program of research that will gradually test more complex scenarios that are more consistent with a range of applied examples. The primary purpose of this simulation is to test the feasibility of predicting individual-level response, particularly among out-of-sample individuals. This type of prediction is rather different from traditional, group-based statistical approaches and warranted tests under optimal conditions before pushing the boundaries of the approach under varying scenarios. Our primary focus in this paper is on point estimation of predicted effects. We acknowledge that variance calculations (e.g. credible intervals) will be critical before dissemination. We have begun developing credible intervals for the PITE approach and will continue this important area of work.

5.2 Results

We tested the performance of the PITE approach by comparing estimation quality at the individual level. Specifically, we were interested in two aspects of estimator quality: bias (accuracy in predictions) and variability (stability) of point estimates across multiple repetitions. Because the PITE is an individual-level estimate, all statistics were calculated within individuals across repetitions. We defined bias as how accurately the PITE recaptured true treatment effects,

which was calculated as the mean difference between true and predicted values across all repetitions. We examined the accuracy of the estimate for each individual, rather than a single summative measure for the whole sample. Variability was defined as the stability or reliability of the predicted values across repetitions. Variability provided information about the degree of similarity (or dissimilarity) of repeated predictions for an individual. Examination of both bias (accuracy) and variability (stability) provides a more comprehensive understanding of the quality of PITE estimates. Although the PITEs may be highly accurate (near the true value on average across repetitions), actual than examination of either bias or variability alone values may be highly variable across repetitions (unstable/unreliable), which would limit the usability of the method in applied realms. Because we were also interested in comparing the performance of imputations and RDTs as underlying predictive methods, we additionally compared the composite measure root mean squared error, which combines information about both bias and variability to understand estimator quality overall. Last, we examined the relationship between observed and predicted values within single repetitions as a descriptive measure for understanding model performance.

Neither multiple imputation nor RDTs experienced convergence problems or other problems with estimation. The imputation approach took significantly longer (required parallelization of R) and required extensive computational resources (e.g. RAM). Without parallelizing, the imputation approach took roughly two to three weeks to complete the full simulation (all repetitions).

5.2.1 Predictive bias

In this study, we use the term bias to refer to the predictive accuracy of the estimator or how well the predicted treatment effects for each individual recapture that individual's true treatment effect across all repetitions: $\frac{\sum_1^r [\hat{\theta}_i - \theta_i]_i}{R}$ where $\hat{\theta}_i$ represents the predicted value for individual i, θ_i is the value of the true treatment effect for individual i, and R is the total number of repetitions, $R = (1, \ldots, r)$. Measures of bias are not scaled in this evaluation, since we used a side-by-side evaluation of methods with the same conditions compared across methods. Overall, across individuals, both the imputation and RDT approach appear to be accurate estimators of the true treatment effect (*imputation:* mean bias $= -.0023$; *RDT:* mean bias $= .0028$). Yet, estimates of bias varied across individuals, particularly for the RDT approach. Figure 3 shows the distribution of bias across individuals using the imputation (red) and RDT (blue) methods. While the imputation method produced fairly unbiased estimates for all individuals, the RDT method showed substantial bias for some individuals. In fact, despite being

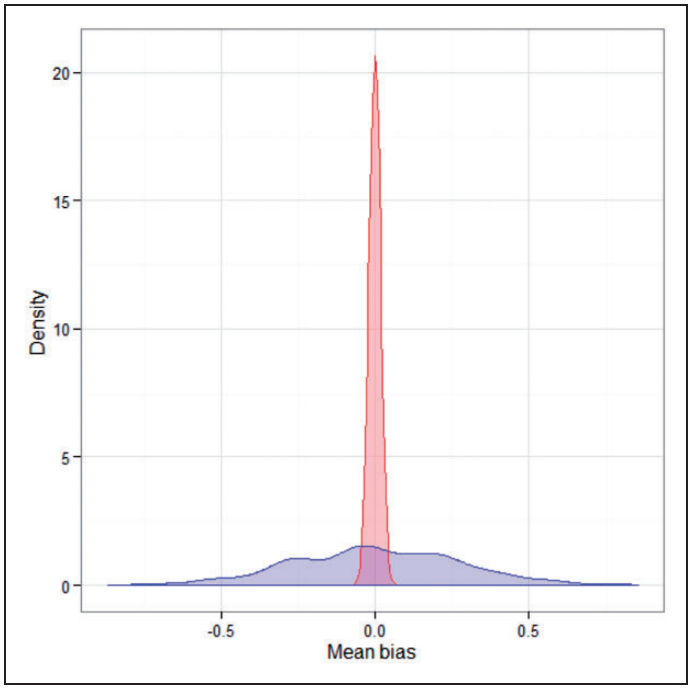

Figure 3. Distribution of individual-level bias using the imputation (red) and random decision trees (blue) estimators.

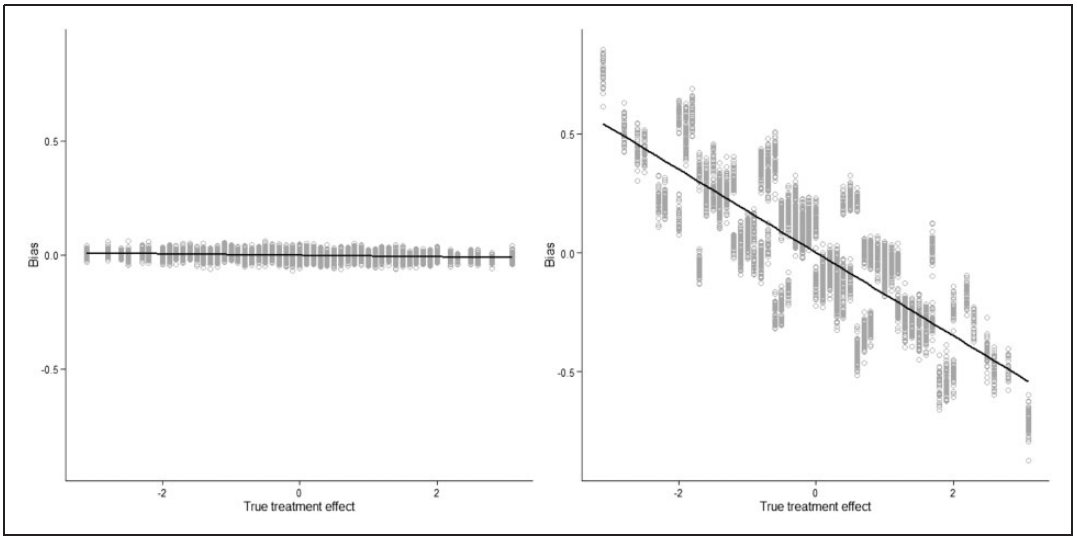

Figure 4. Bias as a function of true treatment effects for the imputation method (left column) and random decision trees method (right column).

unbiased at the mean, the range of estimated bias across individuals ranged from −0.8711 to 0.8536 using the RDT method (compared to −0.0688-0.0600 using the imputation method).

To further explore for whom the RDT method may be producing biased results, we examined bias as a function of the true treatment effect. As seen in Figure 4, there is minimal relation between bias and a true treatment effect for the imputation method but a fairly strong relation between an individual's true treatment effect and bias for the RDT method. The RDT method performs well for individuals in the mid-range, but does not provide accurate estimates of treatment effects for individuals at the extremes.

We then investigated, within one randomly selected repetition, the relationship between the true Y under treatment and control and the predicted Y for the same condition. This was intended to diagnose potential pitfalls in the estimation. Results are presented in Figure 5. The top row shows the scatterplot for the imputation method, the bottom row uses the RDT method. Using both estimators, the PITE approach performed as expected. There was no relationship between predicted values and true values under control (which is consistent with the data generation) and a moderate relationship under treatment. Figure 6 shows the plots of true versus predicted treatment effect ($Y_t - Y_c$) using the imputation method (left) and RDT (right), with colors representing the treatment condition to which the individual was randomized. Consistent with previous results, the imputation method produced estimates that were highly related to true values without any apparent bias in the parameter space. The RDT method produced estimates with more error in the prediction, particularly at the tails of the distribution. Extreme true values tended to be downward biased.

Put together, these results show that both the imputation and RDT methods produce, on average, accurate estimates of the true treatment effect. But, the RDT showed bias in the magnitude of the treatment effect for individuals at the extremes in this simulation scenario.

5.2.2 *Variability of the estimator*

We were interested in assessing the variance of the estimator and comparing the variance across estimators. Variance was calculated as the average squared deviation from the mean PITE across repetitions

$$\frac{\sum_{i=1}^{r}(\textit{predicted treatment effect} - \overline{\textit{predicted treatment effect}})_i^2}{R-1},$$

where $R = 1, \ldots, r$ is the number of repetitions and i refers to the individual subject. Variance in this context refers to how stable or reliable the predicted treatment effects are across multiple repetitions for the same individual.

Results are shown in Figure 7. We note that, unlike bias (which is judged based on its distance from zero), we do not have a pre-existing criterion to aid in the interpretation of variability. This is, in part, a reason for comparing across predictive methods. We rely on the comparison of predictive methods to understand how much variability

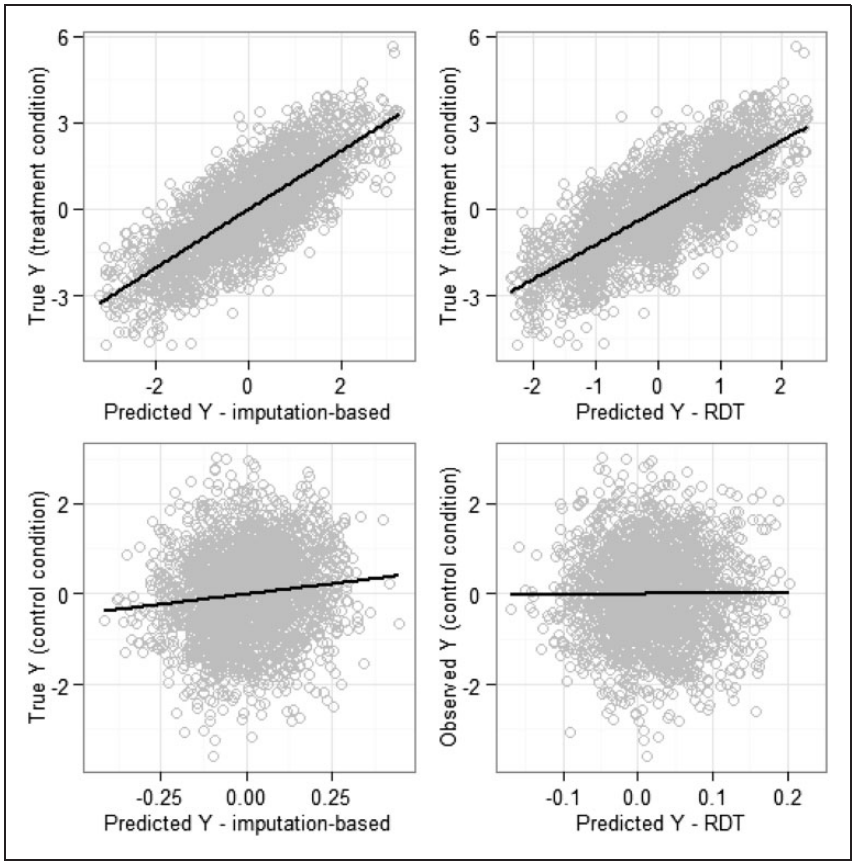

Figure 5. Scatterplot plots of the true and predicted value under treatment and control for the imputation (top row) and random decision trees (bottom row) methods.

RDT: random decision trees.

can be expected (lower the better) and to examine differences across predictive methods. The RDT estimator showed slightly more variability in predictions across simulations, indicating that the predicted values are less stable when this predictive method is used. Across individuals, the average variability for the imputation estimator was .0312 (range = .0175–.0486); average variance for the RDT estimator was .0406 (range = .0232–.0718). Moreover, for the RDT estimator there seemed to be certain individuals for whom variability was elevated in the RDT approach. This, however, did not appear to be a function of the true treatment effect (see Figure 7).

5.2.3 RMSE

RMSE was used as a composite measure for estimator comparison that takes both bias and variability into account. Results are shown in Figure 8. Given the previous results, it is unsurprising that the RMSE favors the imputation method under these simulation conditions (see Figure 8).

6 Discussion

Presently, there is a gap between the theory of individualized medicine and the statistical tools available to help implement individualized medical decision-making. Medical treatment recommendations are typically determined based on ATE, which predict response at the population or group level. Even when differential effects between subgroups are identified, these differential effects are defined by only a few broad-based variables that may or may not be meaningful for an individual patient in a clinical setting. In this paper, we presented a novel framework, *PITE,* which extends newly developed methods for estimating HTE[1,17–32] to estimate treatment response at the individual level. We use trial data to build predictive models under treatment and control, and then obtain model-based potential outcomes for each individual. The difference between the two potential outcomes is taken as the predicted treatment effect for that individual.

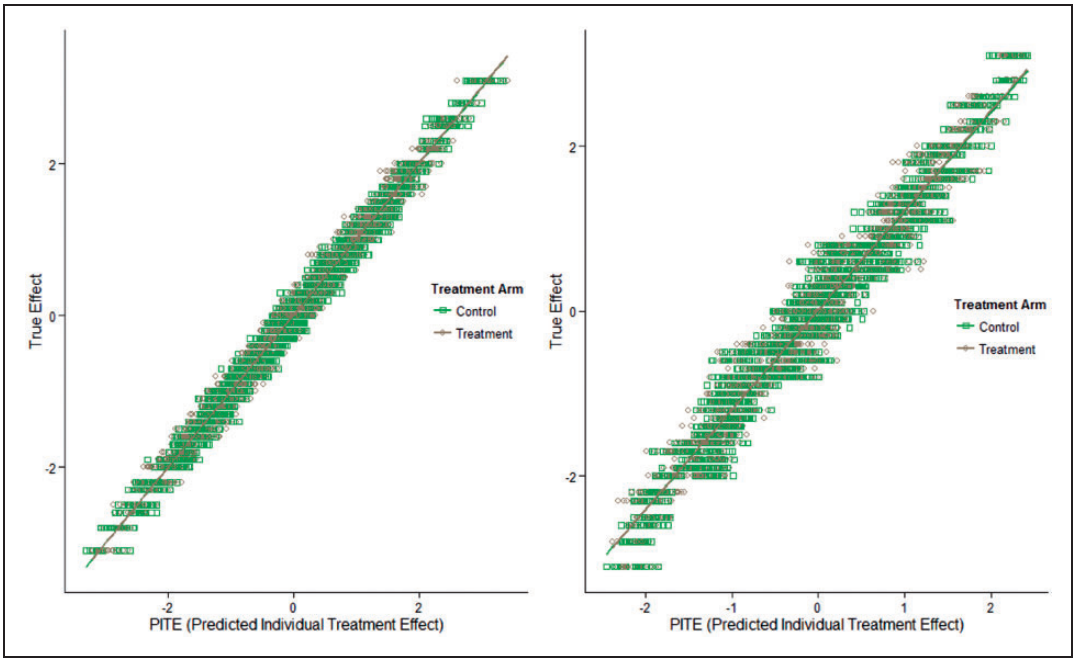

Figure 6. Scatterplot of the true versus predicted treatment effect for the imputation (left) and random decision trees (right) method.

PITE: predicted individual treatment effect.

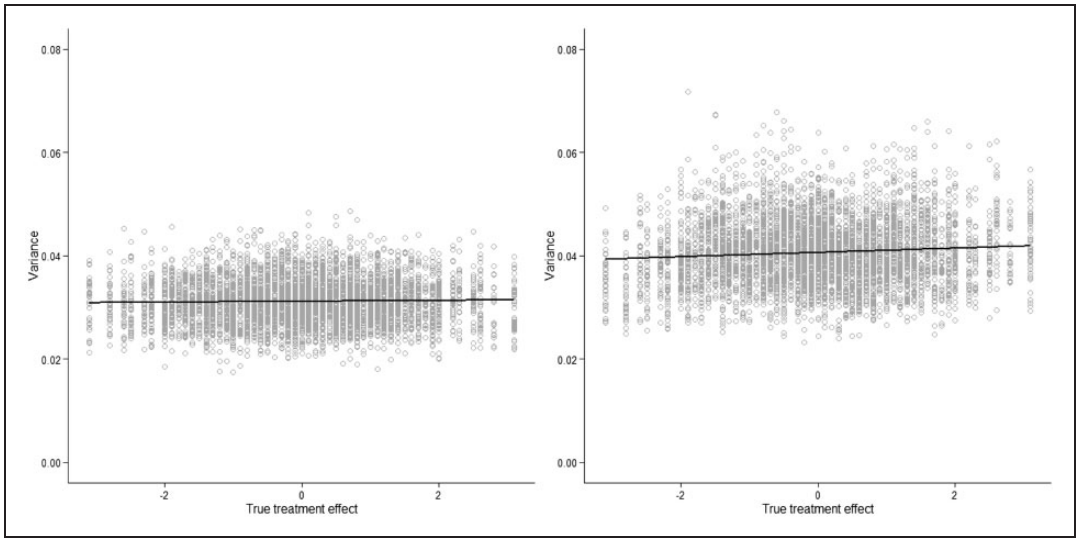

Figure 7. Variance of imputation (left) and random decision trees (right) estimators as a function of true treatment effect.

We began by first demonstrating the feasibility of the approach on applied data, whose original impact analysis – a group-level RCT – showed small average effects. Our re-analysis of the data using the PITE approach showed that the intervention did indeed have positive impacts for certain individuals (and iatrogenic effects for others). The PITE approach was used in conjuncture with clinical data to obtain a prediction for individuals who may be seeking treatment but are unsure whether the intervention would have an effect for them (out-of-sample individuals). Unlike the ATE, which provided critical information about the effectiveness overall, the PITE provided an estimate about how the individual would respond, given baseline risk. This information can be useful for informing future iterations of the program by targeting only those for whom the program would have a positive effect.

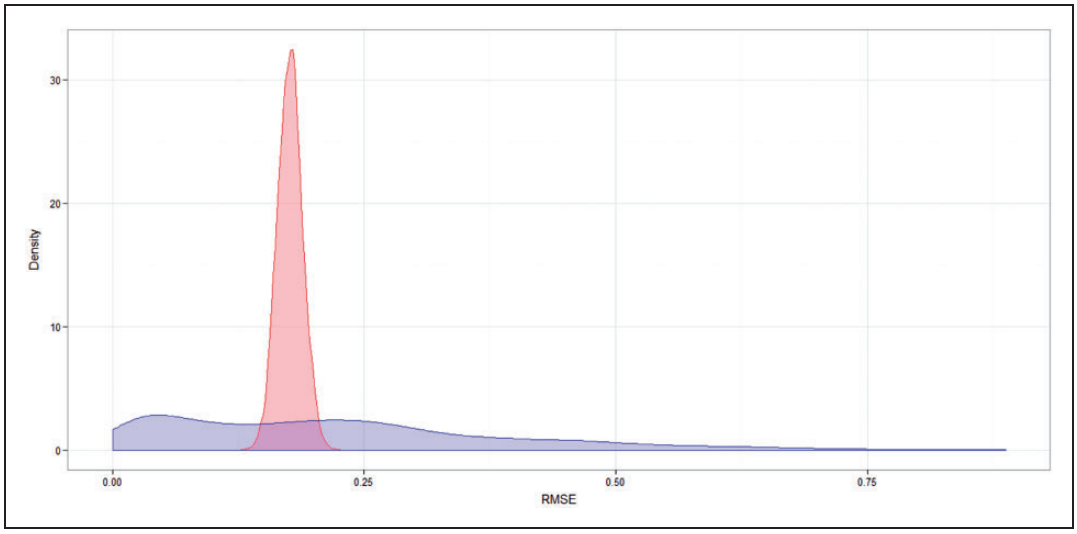

Figure 8. Distribution of RMSE across individuals for the imputation (red) and random decision trees (blue) estimators. RMSE: root mean square error.

The second aim of this study was to test the quality of the PITE estimates. We focus on two aspects of estimator quality: (1) the accuracy (closeness) of the PITE compared to the true treatment effects for each individual (bias); and (2) the stability of the PITE estimates over multiple repetitions of the data (variability). Bias was judged compared to the gold standard of no difference between estimate and true value. Variability, however, did not have an independent measure to aid in interpretation; thus, we relied on the comparison across predictive methods and descriptive information to interpret results.

Overall, our results are favorable for the feasibility of the PITE approach. Using two very different predictive methods, we were able to obtain fairly accurate individual-level predictions on average. At the individual level, imputations performed very well in terms of high accuracy and stability for all individuals. In contrast, the RDT approach showed some important limitations. Despite being having low bias on average, RDTs produced biased estimates for individuals with the strongest treatment response (extreme values of true treatment effect). Additionally, there was less stability in PITE estimates using the RDT approach than the imputation approach, at least in scenarios that match our data-generation model. Put together, these findings suggest that the RDT approach may not be a suitable estimator of PITE, despite having favorable properties for uncovering HTE in general.[55,68]

We expect that many established methods for estimating HTE can be used to derive predictions. We focused on two rather distinct methods in this study; and, despite the outperformance of the imputation method in our analysis, we emphasize that there is likely no single optimal predictive method.[55] The scenario we designed in this simulation was correctly specified for the imputation method (e.g. involved all the right covariates and interaction); therefore, it was not surprising that the imputations worked very well. A similar finding is reported by Hastieet al.,[69] where a linear model is shown to perform better than RDT in a scenario where the true model is linear. The correctly specified design was intentional, as the purpose was, in some sense, proof of concept that using methods designed for HTE can be used for individual predictions.

Future work is planned that will explore the limits of the methods under conditions of increased complexity. We anticipate that, for example, as higher order interactions are introduced to the data, the RDT method (along with other tree-based methods) will outperform the imputation approach. This expectation is based on the way that imputations and RDTs handle interactive effects. Whereas RDTs can easily accommodate interactions without any additional model specifications, imputations require the interactions to be specified in the imputation model. This creates a scenario in which the analyst must have a priori theory about which higher order interactions are driving heterogeneity in effects (which is arguably an unlikely situation), and a sufficient sample size to estimate a rather large and complex imputation equation. It is likely that inclusion of multiple higher order interaction to an already large imputation model will cause estimation problems and/or encounter problems of statistical power. In addition, another important area of future work for expanding the PITE framework is the handling of missingness of observed data. In this study, we used single imputation of

covariates and full information maximum likelihood (FIML) on outcomes in the applied study. The implications of this approach need to be more fully explored. A likely limitation of the PITE method is the case of differential attrition, particularly when imbalanced drop-out is informative. While missingness on the outcome is itself non-problematic, we suspect that informative differential attrition will lead to bias on the estimate of the response on treatment and, consequently, the PITE, unless the mechanism driving the differential attrition of modeled.

We acknowledge that the fundamental purpose of the potential outcomes framework is to come up with causal estimates at the population level. In this study, we presented an extension of the potential outcomes framework where we derived model-based estimates of potential outcomes under both treatment and control. Since the models tested in this study were correctly specified, additional work is needed to understand the assumptions required for these model-based individual estimates (rather than population estimates) as pure, causal effects. Related, this study presents a computer-assisted simulation as a conjecture of the PITE framework. We do not see this as a complete replacement for a formal mathematical proof; however, given our purposes of understanding the method and the conditions under which the method works, we see this approach as well fitting.

The PITE approach marks a first step to integrating a diverse methodological literature on HTE and provides a framework for increasing the clinical utility of established methods. The PITE approach directly focuses on the estimation of predicted effects for individuals who were not part of the original clinical trial. While extrapolation to external data is possible with many tree-based and regression-based models, this practice is often not empirically tested for accuracy and therefore rarely advised. In this study, we explicitly focus on out-of-sample predictions. This is an important aspect of this work because it – along with the focus on individual-level estimation – holds the potential to transform the ways in which treatment decisions in the practice of behavioral and health medicine are made. This type of prediction can greatly advance individualized medicine by providing interventionists and patients access to important individual-level data during treatment selection, prior to the initiation of a treatment protocol. The customization of treatment to the individual can potentially enhance the quality and cost-efficiency of services by allocating treatments to only those who are most likely to benefit.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by grant MR/L010658/1 from the Medical Research Council of the United Kingdom, awarded to Principal Investigator, Thomas Jaki, PhD.

References

1. Huang Y, Gilbert PB and Janes H. Assessing treatment-selection markers using a potential outcomes framework. *Biometrics* 2012; **68**: 687–696.
2. Li A and Meyre D. Jumping on the train of personalized medicine: a primer for non-geneticist clinicians: part 3. Clinical applications in the personalized medicine area. *Curr Psychiatry Rev* 2014; **10**: 118–132.
3. Goldhirsch A, Winer EP, Coates AS, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International expert consensus on the primary therapy of early breast cancer 2013. *Ann Oncol* 2013; **24**: 2206–2223.
4. Aquilante CL, Langaee TY, Lopez LM, et al. Influence of coagulation factor, vitamin K epoxide reductase complex subunit 1, and cytochrome P450 2C9 gene polymorphisms on warfarin dose requirements. *Clin Pharmacol Ther* 2006; **79**: 291–302.
5. Assmann SF, Pocock SJ and Enos LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000; **355**: 1064–1069.
6. Pocock SJ, Assmann SE, Enos LE, et al. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 2002; **21**: 2917–2930.
7. Brookes ST, Whitley E, Peters TJ, et al. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technol Assess* 2001; **5**: 1–56.
8. Cui L, James Hung HM, Wang SJ, et al. Issues related to subgroup analysis in clinical trials. *J Biopharm Stat* 2002; **12**: 347.
9. Fink G, McConnell M and Vollmer S. Testing for heterogeneous treatment effects in experimental data: false discovery risks and correction procedures. *J Dev Eff* 2014; **6**: 44–57.

10. Lagakos SW. The challenge of subgroup analyses – reporting without distorting. *N Engl J Med* 2006; **354**: 1667–1669.
11. Wang R, Lagakos SW and Ware JH. Statistics in medicine – reporting of subgroup analyses in clinical trials. *N Engl J Med* 2007; **357**: 2189–2194.
12. Neyman J. Sur les applications de la theorie des probabilites aux experiences agricoles: essai des principes (Masters Thesis); Justification of applications of the calculus of probabilities to the solutions of certain questions in agricultural experimentation. Excerpts English translation (Reprinted). *Stat Sci* 1923; **5**: 463–472.
13. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974; **66**: 688–701.
14. Rubin DB. Formal modes of statistical inference for causal effects. *J Stat Plan Inference* 1990; **25**: 279–292.
15. Rubin DB. Causal inference using potential outcomes. *J Am Stat Assoc* 2005; **100**: 322–331.
16. Holland PW. Statistics and causal inference. *J Am Stat Assoc* 1986; **81**: 945–960.
17. Basu A. Estimating person-centered treatment (PeT) effects using instrumental variables: an application to evaluating prostate cancer treatments. *J Appl Econ* 2014; **29**: 671–691.
18. Foster JC, Taylor JM and Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med* 2011; **30**: 2867–2880.
19. Doove LL, Dusseldorp E, Deun K, et al. A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions. *Adv Data Anal Classif* 2013; 1–23.
20. Freidlin B, McShane LM, Polley MY, et al. Randomized phase II trial designs with biomarkers. *J Clin Oncol* 2012; **30**: 3304–3309.
21. Imai K and Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann Appl Stat* 2013; **7**: 443–470.
22. Imai K and Strauss A. Estimation of heterogeneous treatment effects fromrandomized experiments, with application to the optimal planning of the Get-Out-the-Vote campaign. *Polit Anal* 2011; **19**: 1–19.
23. Heckman JJ, Urzua S and Vytlacil E. Understanding instrumental variables in models with essential heterogenetiy. *Rev Econ Stat* 2006; **88**: 389–432.
24. Zhang Z, Wang C, Nie L, et al. Assessing the heterogeneity of treatment effects via potential outcomes of individual patients. *J R Stati Soc C* 2013; **62**: 687–704.
25. Bitler MP, Gelbach JB and Hoynes HW. Can variation in subgroups' average treatment effects explain treatment effect heterogeneity? *Evidence from a social experiment*, National Bureau of Economic Research, NBER Working Paper No. 20142, May 2014. Available at: http://www.nber.org/papers/w20142
26. Green DP and Kern HL. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opin Quart* 2012; **76**: 491–511.
27. Shen C, Jeong J, Li X, et al. Treatment benefit and treatment harm rate to characterize heterogeneity in treatment effect. *Biometrics* 2013; **69**: 724–731.
28. Simon N and Simon R. Adaptive enrichment designs for clinical trials. *Biostatistics* 2013; **14**: 613–625.
29. Rosenbaum PR. Confidence intervals for uncommon but dramatic responses to treatment. *Biometrics* 2007; **63**: 1164–1171.
30. Poulson RS, Gadbury GL and Allison DB. Treatment heterogeneity and individual qualitative interaction. *Am Stat* 2012; **66**: 16–24.
31. Cai T, Tian L, Wong PH, et al. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* 2011; **12**: 270–282.
32. Ruberg SJ, Chen L and Wang Y. The mean does not mean as much anymore: finding sub-groups for tailored therapeutics. *Clin Trials* 2010; **7**: 574–583.
33. Gadbury G, Iyer H and Allison D. Evaluating subject-treatment interaction when comparing two treatments. *J Biopharm Stat* 2001; **11**: 313.
34. Gadbury GL and Iyer HK. Unit-treatment interaction and its practical consequences. *Biometrics* 2000; **56**: 882–885.
35. Gadbury GL, Iyer HK and Albert JM. Individual treatment effects in randomized trials with binary outcomes. *J Stat Plan Inference* 2004; **121**: 163.
36. Nomi T and Raudenbush SW. *Understanding treatment effects heterogeneities using multi-site regression discontinuity designs: example from a ''Double-Dose'' Algebra Study in Chicago*. Chicago: Society for Research on Educational Effectiveness, 2012.
37. Na C, Loughran TA and Paternoster R. On the importance of treatment effect heterogeneity in experimentally-evaluated criminal justice interventions. *J Quant Criminol* 2015; **31**: 289–310.
38. Schapire RE and Freund Y. *Boosting: foundations and algorithms*. Cambridge, MA: MIT Press, 2012.
39. LeBlanc M and Kooperberg C. Boosting predictions of treatment success. *Commentary* 2010; **107**: 13559–13560.
40. Lipkovich I and Dmitrienko A. Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. *J Biopharm Stat* 2014; **24**: 130–153.
41. Zhang Z, Qu Y, Zhang B, et al. Use of auxiliary covariates in estimating a biomarker-adjusted treatment effect model with clinical trial data. *Stat Meth Med Res* 2016; **25**: 2103–2119.
42. Su X and Johnson WO. Interaction trees: exploring the differential effects of intervention programme for breast cancer survivors. *J R Stat Soc C* 2011; **60**: 457–474.

43. Su X, Kang J, Fan J, et al. Facilitating score and causal inference trees for large observational studies. *J Mach Learn Res* 2012; **13**: 2955–2994.
44. Kang J, Su X, Hitsman B, Liu K, et al. Tree-structured analysis of treatment effects with large observational data. *J Appl Stat* 2012; **39**: 513–529.
45. Lipkovich I, Dmitrienko A, Denne J, et al. Subgroup identification based on differential effect search – a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat Med* 2011; **30**: 2601–2621.
46. Su X, Tsai C-L, Wang H, et al. Subgroup analysis via recursive partitioning. *J Mach Learn Res* 2009; **10**: 141–158.
47. Su X, Zhou T, Yan X, et al. Interaction trees with censored survival data. *Int J Biostat* 2008; **4**: 1–26.
48. Zeileis ATK. Model-based recursive partitioning. *J Comput Graph Stat* 2008; **17**: 492–514.
49. Dusseldorp E, Conversano C and Van Os BJ. Combining an additive and tree-based regression model simultaneously: STIMA. *J Comput Graph Stat* 2010; **19**: 514–530.
50. Ciampi A, Negassa A and Lou Z. Tree-structured prediction for censored survival data and the Cox model. *J Clin Epidemiol* 1995; **48**: 675–689.
51. Dai JY, Kooperberg C, Leblanc M, et al. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika* 2012; **99**: 929–944.
52. Dixon DO and Simon R. Bayesian subset analysis. *Biometrics* 1991; **47**: 871–881.
53. Gail M and Simon R. Testing for qualitative interactions between treatment effects and patient subsets (with appendix). *Biometrics* 1985; **41**: 361–372.
54. Simon R. Bayesian subset analysis: application to studying treatment-by-gender interactions. *Stat Med* 2002; **21**: 2909–2916.
55. Malley JD, Malley KG and Pajevic S. *Statistical learning for biomedical data*. New York: Cambridge University Press, 2011.
56. Schafer JL. *Analysis of incomplete multivariate data*. New York: Chapman & Hall/CRC, 1997.
57. Little RJA and Rubin DB. *Statistical analysis with missing data*, 2nd ed. New York: John Wiley, 2002.
58. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: J Wiley & Sons, 1987.
59. Dore DD, Swaminathan S, Gutman R, et al. Different analyses estimate different parameters of the effect of erythropoietin stimulating agents on survival in end stage renal disease: a comparison of payment policy analysis, instrumental variables, and multiple imputation of potential outcomes. *J Clin Epidemiol* 2013; **66**(8 Suppl): S42–S50.
60. Raghunathan TE, Lepkowski JM, Van Hoewyk J, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv Methodol* 2001; **27**: 85–95.
61. Hershey A, Devaney B, Wood RG, et al. *Building Strong Families (BSF) project data collection, 2005–2008*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2011.
62. Inter-university Consortium for Political and Social Research, www.icpsr.umich.edu/icpsrweb/ICPSR/studies/29781 (accessed 5 December 2015).
63. Radloff LS and Radloff LS. Center for Epidemiologic Studies Depression Scale. The CES-D Scale: a self-report depression scale for research in the general population. *Appl Psychol Meas* 1977; **1**: 385–401.
64. Muthén LK and Muthén BO. *Mplus user's guide*, 7th ed. Los Angeles: Muthén & Muthén, 1998–2012.
65. Su Y-S, Gelman A, Hill J and Masanao Y. Multiple imputation with diagnostics (mi) in R: opening windows into the black box. *J Stat Software* 2011; **45**: 1–31.
66. R Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2015.
67. Liaw A and Wiener M. Classification and regression by random forest. *R News* 2002; **2**: 18–22.
68. Breiman L. Random forests. *Mach Learn* 2001; **45**: 5–32.
69. Hastie T, Tibshirani R and Friedman J. *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed. New York: Springer, 2009, p.xxii.