# A Permutation Test for Assessing the Presence of Individual Differences in Treatment Effects

Chi Chang – Michigan State University
Thomas Jaki – Lancaster University and University of Cambridge
Muhammad Saad Sadiq – University of Miami
Alena Kuhlemeier – University of New Mexico
Daniel Feaster – University of Miami
Natalie Cole – University of New Mexico
Andrea Lamont – University of South Carolina
Daniel Oberski – Utrecht University
Yasin Desai – Lancaster University
The Pooled Resource Open-Access ALS Clinical Trials Consortium*
M. Lee Van Horn – University of New Mexico

Running Head: A Permutation Test for Assessing Heterogeneity in Treatment Effects

Corresponding Author:

M. Lee Van Horn, PhD. Tech 274, 1 University of New Mexico, Albuquerque, NM 87131

mlvh@unm.edu; (505) 277-4535

*Data used in the preparation of this article were obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database. As such, the following organizations and individuals within the PRO-ACT Consortium contributed to the design and implementation of the PRO-ACT Database and/or provided data, but did not participate in the analysis of the data or the writing of this report:

- Neurological Clinical Research Institute, MGH
- Northeast ALS Consortium
- Novartis
- Prize4Life
- Regeneron Pharmaceuticals, Inc.
- Sanofi
- Teva Pharmaceutical Industries, Ltd.

This paper was partially supported by grant # MR/L010658/1 awarded to Thomas Jaki by the United Kingdom Medical Research Council. For further information or comments please contact the senior author, M. Lee Van Horn at mlvh@unm.edu

## Abstract

An important goal of personalized medicine is to identify heterogeneity in treatments effects and then use that heterogeneity to target the intervention to those most likely to benefit. Heterogeneity is assessed using the predicted individual treatment effects (PITE) framework, and a permutation test is proposed to establish if significant heterogeneity is present given the covariates and predictive model or algorithm used for PITEs. We first show evidence for heterogeneity in the effects of Riluzole acrossan illustrative example data set. We then use simulations with two different predictive methods (linear regression model and Random Forests) to show that the permutation test has adequate type-I error control. Next, we use the example dataset as the basis for simulations to demonstrate the ability of the permutation test to find heterogeneity in treatment effects for a PITE estimate as a function of both effect size and sample size. We find that the proposed test has good power for detecting heterogeneity in treatment effects when the heterogeneity was due primarily to a single predictor, or when it was spread across the predictors. Power was found to be greater for predictions from a linear model than from random forests. This non-parametric permutation test can be used to test for significant differences across individuals in PITEs obtained with a given set of covariates using any predictive method with no additional assumptions.

*Keywords:* Predicted individual treatment effects, heterogeneity in treatment effects, personalized medicine, permutation test, Random Forests

## 1. Introduction

The key premise of personalized medicine is the identification and targeting of individuals most likely to benefit from a given intervention, (1) with the goal of improving health care outcomes and decreasing costs.(1,2) Much recent research has focused on statistical approaches for identifying a small number of subgroups of individuals who differ in their response to interventions, (3 –13) while a smaller body of research has focused on predicting intervention responses at an individual level. (4,14–19) For situations in which treatment response is related to a set of covariates, which is not a small number of clearly defined subgroups, individual-level predictions are particularly appropriate. Even if most covariates were categorical, with high dimensional data and finite samples, individual-level predictions may contain more information about heterogeneity in treatment effects than is contained in subgroups. This study focuses on the use of predicted individual treatment effects (PITE) (20,21), which builds on the potential outcomes framework (22,23) and results in predictions of the intervention response for individual patients.

The PITE approach utilizes data from a randomized clinical trial with a potentially large number of baseline covariates to generate predictions from a model or algorithm, which are then used in estimating PITEs. The same model or algorithm can then be used to generate treatment effect estimates for new subjects not used in training. Given that predictive algorithms have been trained, the next question becomes whether the data reveal more variability in individual predictions than would be expected due to chance. In other words, 'Do individuals differ in the predicted effects of the intervention?' is a question that should be answered before a given set of PITE estimates are used to ensure that this personalized method is only used when there are individual differences. This paper proposes a permutation test to answer this question. An advantage of the proposed method is that it can be generally applied to any method for estimating predictions for the treated group and the control group. While methods exist for estimating the significance of heterogeneity in treatment effects using kernel regression and instrumental variable regression (24,25) and for estimating whether subgroups identified by treatment response improve on group means, (26) the proposed permutation test provides flexibility in choosing the estimator and can use machine learning to estimate potential outcomes while retaining frequentist properties. The next section describes the PITE approach in general terms before providing details of our proposed permutation test. In Section 3, we use the PITE framework and the proposed test to evaluate heterogeneity in the effects of interventions for ALS; Section 4 uses this test on simulated data to show the type I error rates of the PITE permutation test using two different predictive models with and without main effects of covariates. In Section 5, we use the ALS example as the basis for simulations that demonstrate the ability of the permutation test to find heterogeneity in treatment effects as a function of both effect size and sample size. Section 6 concludes with a discussion of results.

## 2. **Permutation test for PITE**

Potential outcomes (23,27,28) provide a powerful way for understanding causal effects. In the context of a two-arm randomized trial, before treatment assignment, each individual has a potential outcome under both treatment conditions, which is the outcome that they would obtain if assigned to treatment ($Y_i^t$) and the outcome they would obtain under control ($Y_i^c$). The causal effect of treatment for an individual is defined as:

$$Y_i^t - Y_i^c \qquad (1)$$

The "fundamental problem of causal inference" (29) is that this effect is never observed because once an individual is assigned to a condition, the outcome can only be realized for that condition. The PITE framework proposes that we can use a predictive function to capture some proportion of the potential outcomes[1]:

$$Y_i^t = f_t(x_i) + \varepsilon_{it}, \qquad (2)$$

$$Y_i^c = f_c(x_i) + \varepsilon_{ic},$$

where $x_i$ contains covariates for individual $i$, $\varepsilon_{it} \sim \mathrm{N}(0, \sigma_t^2)$ is a patient-level random effect if treated, and $\varepsilon_{ic} \sim \mathrm{N}(0, \sigma_c^2)$ is a patient-level random effect if control, and $f(.)$ indicates any predictive function. Following Lamont et al. (20), PITE is defined as the difference between the predicted outcome under treatment and the predicted outcome under control for each patient *i* given their observed covariates.

$$\text{PITE}_i = \hat{Y}_i^t - \hat{Y}_i^c = \hat{f}\_t(x_i) - \hat{f}_c(x_i) \qquad (3)$$

The PITE, therefore, is an estimate of the treatment effect for a particular individual given the covariates and predictive model or algorithm used. The PITE is equal to the potential outcomes definition of an individual causal effect only in the unlikely, and unknowable, scenario that the random effects for both conditions are equal to zero. It should be noted that even if the average treatment effect equals zero, it is still possible that there are some individuals who would be expected to do better given the treatment than control and others who would be expected to do better under control. Therefore, in this paper, we exclude the expected value of the PITE from the test, as this value is an estimate of the average treatment effect and not evidence of individual differences. Because the PITE is defined very generally as the difference between two predictions, it can be used with any predictive model that provides outcome prediction on a patient-level (e.g., the General Linear Model, random forests,(30) Bayesian additive regression trees,(31) neural networks (32)). In addition, PITE can be used for predicting treatment effects given information on covariates for patients who are not originally part of the clinical trial.

The presence of individual differences has implications for how a treatment would be implemented: if there are individual differences in the treatment effect, it suggests that it may be worthwhile to collect and use individual-level data to help guide treatment decisions. Therefore,

[1] We gratefully acknowledge an anonymous reviewer who suggested this notation.

we propose a permutation test to evaluate whether there are individual differences in PITEs. This paper demonstrates the use of a permutation test with two different predictive approaches, Random Forests(30) and linear regression.

Following equation 2, we begin by using data from those randomized to treatment to estimate $f_t(x)$ which allows estimation of predicted values of the outcome under treatment ($\hat{Y}_i^t$) from a set of covariates $x$. Note that while the random effects, $\varepsilon_{it}$, are observed for those in the treatment group, we do not use these because 1) the random effect under control ($\varepsilon_{ic}$) is not known, and 2) the random effects for new individuals are unknown. The function for estimating the predicted potential outcome under control can be estimated in the same way. In the sense that PITE attempts to estimate potential outcomes from other variables, it can be considered a latent variable model. PITE is one method for personalized medicine that uses latent variables to investigate heterogeneity in treatment response; others include methods to estimate the proportion of subjects who benefit from treatment. (33) The example includes 1) latent classes defined by heterogeneity in treatment response, (34) and 2) Bayesian analyses that assess differential treatment response as a function of continuous and categorical latent variables. (35)

Once $f_t(x)$ and $f_c(x)$ are estimated from the trial data, individual-level PITE estimates for patients in the original trial and those who did not take part in it, can be obtained using Equation 3. It should also be noted that the algorithm or model used for prediction will determine both the assumptions made and the efficiency of the predictions (e.g., linear models assume linearity in the parameters and that all multiway interactions are included in the covariates and tend to have increased efficiency when those assumptions are met). This paper uses both Random Forests and the linear model to obtain predictions. We expect that the best model will be situation dependent.

We propose a permutation test (23,36–39) be used to test for individual differences in PITEs. Our test focuses on the standard deviation (SD) of the PITEs, because this estimate quantifies individual differences in the predicted treatment effects. More specifically, we test the hypothesis:

$H_0$: $\text{PITE}_i = \text{PITE}_j$ for all pairs of individuals $(i, j) \in \{1, .., n\}$

$H_A$: $\text{PITE}_i \neq \text{PITE}_j$ for at least one pair of individuals $(i, j) \in \{1, .., n\}$

In other words, the null hypothesis is that there are no individual differences in the PITEs obtained. To test this hypothesis, the permutation test first approximates the sampling distribution of PITEs' SDs under the null hypothesis, i.e., when the set of covariates in the PITE prediction models have the same effect on the outcome across treatment groups (the covariates may be prognostic, but not predictive). This is done by permuting treatment assignment. The observed SD of the PITEs from the data is compared against the resulting distribution.

The following algorithm describes the procedure in detail.

1) Estimate PITE models and compute $\widehat{PITE}_i$, for all $n$ individuals in the dataset using a prediction method and set of covariates
2) Estimate the standard deviation of the estimated PITEs as

$$\hat{\sigma}_{PITE} = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{PITE}_\imath - \overline{\widehat{PITE}_\imath})^2$$

where $\overline{\widehat{PITE}_\imath} = \frac{1}{n}\sum_{i=1}^{n} \widehat{PITE}_\imath$.

3) Randomly permute the treatment assignment of all patients in the study.
4) Estimate the PITE model and compute $\widehat{PITE}_i^{\,p}$, using the permuted data and the same prediction method as used in step 1.
5) Estimate the standard deviation, $\hat{\sigma}_{PITE}^{P}$ of $\widehat{PITE}_i^{\,p}$, in the same manner as in step 2.
6) Repeat steps 3 through 5, P times.
7) Obtain the p-value, $p^P$, associated with the above hypothesis as $p^P = \frac{\sum_{p=1}^{P} I(\hat{\sigma}_{PITE}^{P} > \hat{\sigma}_{PITE})}{P}$, with I(.) being an indicator function equal to 1 if the condition in the parenthesis is satisfied, and 0 otherwise.
8) Reject the above hypothesis at level α if $p^P < \alpha$.

The proposed PITE permutation test is intended to be conducted once per dataset and does not inherently involve multiple comparisons or multiple testing, which requires strong assumptions for permutation tests. (40)

For each of 1,000 replications in this study, 1,000 permutations were used in our subsequent evaluations. Following binomial arguments, this yields a .007% uncertainty in estimated p-values of which the true value should be 5%.

## 3. Demonstration of the permutation test: An intervention for individuals with ALS

Amyotrophic Lateral Sclerosis (ALS, also known as Motor Neuron Disease) is a neurodegenerative disorder that affects motor neurons in the brain and spinal cord. We use the Pooled Resource Open-Access ALS Clinical Trials database (PRO-ACT), (41) which is publicly available at their website: http://nctu.partners.org/ProACT. In 2011, Prize4Life, in collaboration with the Northeast ALS Consortium formed the PRO-ACT Consortium, which makes data from 23randomized trials of the drug Riluzole available. We note that one other study has used this data to evaluate a personalized approach based on the identification of subgroups of responders (26,42), finding evidence for significant individual differences in response to Riluzole.

The advantage of this dataset is that it pools data from many randomized trials of Riluzole and thus has the sample size needed to test for heterogeneity in treatment effects. However, to maintain confidentiality, the data does not include a study identifier, and thus it is impossible to model study-specific effects. This is potentially problematic because it is possible that systematic differences in subject populations as well as outcomes between studies exist, which could themselves result in the identification of heterogeneity in treatment effects. More generally, achieving personalized medicine in practice is likely to be the end result of a program of research involving many steps of which this is a preliminary example.

After estimating PITEs using a linear model with the PRO-ACT data, we then use the permutation test to examine heterogeneity in individual treatment effects. PRO-ACT includes information from more than 8,500 patients with ALS. Each of them participated in a clinical trial and received either a placebo or treatment. Following Küffner et al., (41) we used the slope of the ALSFRS score from a repeated measures model for each patient as the primary outcome, and the 2,910 patients (1,766 in experimental treatments and 1,144 in control ones) who had complete data for 17 covariates, treatment condition, and the outcome.

To avoid overfitting the data, we advocate either choosing both the predictive method and the covariates for the PITEs a priori, or adjusting for the variable selection process. (43) Here, we demonstrate the permutation test based on a linear model that included 7 (out of 17) covariates found to have significant interactions with treatment. In its simplest form, with a linear model, PITE captures baseline by treatment interactions; thus, we used this as the criteria for variable selection. PITEs, however, are much more general than these interactions as they capture the joint effect of many predictors and, depending on the predictive method used, implicitly capture non-linear and higher-order interactions. The seven covariates used for obtaining PITE estimates were: respiratory rate, systolic blood pressure, age, gender, limb only (coded 1 if the onset location was only in the limb), use of Riluzole, and delayed medication (coded 1 if the time duration between the first time the patient was assessed during the trial and the time medication was first given was more than a year, zero otherwise).

### *Results*

Appendix B and C showed the descriptive statistics of the 7 covariates that were found to have significant interactions with treatment. Appendix A showed the linear regression model's coefficients and standard errors for both treatment and control conditions with all 17 covariates in the ALS dataset. The results for the control group can be considered to be the main effects, in the absence of treatment, from the linear model and the differences between the treatment and control groups are the linear interactions that contribute to predicted individual differences in treatment effects. Figure 1 includes the permutation distribution of SDs of PITEs based on the procedure outlined above together with the observed SD from the ALS dataset, which at .127 is on the upper tail of this distribution. The p-value for the permutation test was .005, providing evidence for individual-level treatment heterogeneity based on the linear model and the 7 covariates included.

[Insert Figure 1 Here]

*Figure 1.* Permutation distribution of PITEs' SDs and the observed PITE's SD in the ALS study.

## 4. Type I error rates for the permutation test

The promise of this permutation test is that one can use any conventional or machine learning function and still get correct frequency properties. In this section, we investigate if this promise is fulfilled. We begin our evaluation of the proposed test by examining the type I error rate of the permutation test under the null hypothesis, i.e., that the treatment effect is the same for all individuals. Simulations were conducted with the true PITE for each individual being equal to the average treatment effect, meaning that covariates had no impact on the PITE. When the type I error is .05, the p-value obtained from step 7, above, should be below .05 in only 5% of the simulations.

In this phase of our investigation, PITE was estimated with sample sizes of 100, 250, 500, 1,000, and 5,000 using both linear regression (via the *lm* function in R) and Random Forests (via the *randomForestSRC* package in R, (44) tuned to have a node depth of 10). To make the simulation more realistic, we included five prognostic covariates that had the same effects across the treatment and control conditions. These included three normally distributed covariates with means of 0 and variances of 1, and two binary variables, each with a 0.5 probability of endorsing 1 or 0. The covariate effects of 0.406, -0.239, 0.703, -0.090, and -0.299, respectively, were identical across the treatment and the control groups – implying that they did not predict differential responses to the treatment - and were included in all analyses. To show type I error rates when many additional variables were included in the predictive model, we ran analyses with varying numbers of continuous variables with standard normal distribution and binary variables with binomial distributions with a probability of success of .5. They were generated to be unrelated to the outcome, hereafter called 'nuisance variables.' Because sample size limits the number of nuisance variables that can be included in a linear regression model, we examined an increased number of nuisance variables with larger samples. Analyses also varied the true treatment effect to show that the PITE, as described above, does not capture the main effects of treatment. For the Random Forest model, we used 500 trees and 10 random split points to split a node.

*Results*

The results (see Table 1) established that across all conditions, the permutation test rejected the null hypothesis between 4.7% and 6.3 % of the time with the linear regression model, and 2.9% to 6.3% of the time with Random Forests. The estimated type I errors appear to be mostly within simulation error (±0.007) with no discernable pattern detectable in relation to the number of nuisance variables, main effect, or total sample size.

[Insert Table 1 Here]

## 5. Power of the permutation test

Next, we used the ALS results in section 3 as the basis for simulations examining some of the factors that influence the permutation test's statistical power. In the data generation model for the power simulations, the parameters from the predictive linear model using the ALS example (shown in Appendix A) were used as the starting point. To mimic the real-world scenario, we included only the seven covariates that were found to have significant interactions with treatment in the simulation. The first three of these, respiratory rate, systolic blood pressure, and age, were generated as normally distributed random variables with means and SDs equal to the corresponding values estimated from the ALS dataset (details provided in Appendix B). Similarly, the remaining four binary random variables, gender, limb only, use of Riluzole, and Delayed Medication, were generated as binomial with the same probabilities as observed (in Appendix C). The covariance matrix was generated by mimicking that of the ALS data. The outcome was generated with the effect size scaled to mimic the ALS example with sample sizes of 1,000 (to examine if the effects could have been found with a smaller sample) and 3,000 (as in the ALS example), with equal size for the treatment and placebo groups. To assess the impact of adding further covariates when fitting PITE, we evaluated statistical power with 0, 20, 50, or 100 nuisance variables, all of which were generated either from standard normal distributions or binomial distributions with a probability of success of .5. The nuisance variables were included when estimating PITEs despite not being related to the outcome.

A challenge in estimating power was that measures of effect size for PITE have not been previously defined. In this study, we used the average PITE estimate divided by the pooled SD of the outcome as follows:

$$PITE\ Effect\ Size = \frac{\frac{\Sigma|PITE_i|}{N}}{\sqrt{\frac{(N_T - 1) \times (\sigma_T^2) + (N_C - 1) \times (\sigma_C^2)}{N_T + N_C - 1}}}$$

This resulted in an estimated effect size of 0.19 for the PITEs in the ALS example, meaning that the average person was 0.19 SD from the average effect size. When estimating power, data were generated with effect sizes of either 0.19 or 0.38, with the latter included to examine the method's ability to identify a larger effect.

We also examined the permutation test's power to detect heterogeneity that is mostly due to a single variable as well as when heterogeneity was spread across multiple variables which each contribute a small amount. Power simulations were run for six conditions, which differed from one another in the relative contributions of the 7 predictors. Specifically, these 6 conditions were: 1) the total heterogeneous effect is evenly spread across all 7 covariates ("Spread"); 2) 90% of the total heterogeneous effect is due to the first continuous variable, and 10% to the other 6 covariates ("90/10 Cont."); 3) as 90/10 Cont., but with 75%/25% split between the first continuous variable and the other 6 covariates ("75/25 Cont."); 4) as above, but with a 50%/50% split ("50/50 Cont."); 5) as above, but with a 25%/75% split ("25/75 Cont."); and 6) 90% of the total heterogeneous effect is due to the first binary variable, and 10% to the other 6 covariates ("90/10 Bin."). Thus, the power of PITE prediction was examined in a total of 96 conditions, i.e., in 2 (sample sizes) × 4 (numbers of nuisance variables) × 2 (effect sizes) × 6 (heterogeneity effect distributions).

Once data was generated, both the linear model and Random Forests were run for each dataset and under each condition using the procedures described above. For the Random Forest model, the depth was restricted to 10. The percentage of times that the permutation test was significant for each condition was recorded as the power estimate.

*Results*

Power for each of the 96 conditions was estimated as the proportion of 1000 simulations for which the permutation test was significant. The results obtained with an effect size of 0.19 are presented in Table 2, and those obtained with an effect size of 0.38 in Table 3. As expected, power increased both when sample size increased and when effect size increased. Our results also indicated that increasing the number of nuisance variables decreased power substantially, highlighting the importance of selecting meaningful covariates. With the ALS observed effect size and sample size, the predictive (post-hoc) power obtained from the linear regression model was adequate (>.80) when there were 50 nuisance variables, but not when there were 100. When using Random Forests for predictions, on the other hand, power was adequate with 20 nuisance variables at the same sample size. However, with a sample size of 1,000, the linear regression model's power was poor even with 20 nuisance variables and would be inadequate with Random Forests using the same tuning parameters.

[Insert Table 2 Here]

[Insert Table 3 Here]

With an effect size twice as great as observed, the power of the permutation test using linear regression model was low only when the sample size was 1,000, and there were 50 or 100 nuisance variables, but Random Forests' power was marginal even with a sample size of 3,000 if there were 100 nuisance variables. We note that power estimates for Random Forests were expected to be lower than those for the linear regression model given that the data were simulated using the latter method and that no higher-order interactions or non-linear effects were included.

The other factor that we varied across the simulations was how the effect of the covariates on heterogeneity in treatment effects was spread out. The reason for this is to show a core advantage of PITE, which is its ability to detect many small effects that add up to something meaningful rather than just one large effect. When looking across the six different spreads of the PITE effect, it was striking that when there was adequate power for one of them, there was usually adequate power for all. The only two exceptions to this were 1) when the effect was carried primarily by one binary variable (i.e., there are only two different kinds of responses), in which case, power was higher than for the other conditions; and 2) for random forests the power is lower for the binary predictor. The key result here is that, in the ALS example, power was about the same regardless of whether the heterogeneity in treatment effects is attributed primarily to one of the 7 important variables or when it is spread out across all 7.

One apparently inconsistent finding in these results was that in some conditions, power was less than 5%, the type I error rate. The reason for this is that any random variable will cause variability in the PITE to a certain extent. In some cases, with many nuisance variables, the effects of the predictors we were simulating were smaller than effects due to chance, resulting in a lower probability of finding heterogeneity in the effects of the predictor than would have been expected due to chance. Importantly, this implies that adding more predictors will increase the noise in PITE estimates and result in larger heterogeneity being estimated.

Finally, in order to illustrate that PITEs do not capture heterogeneity due to variables not included in the predictive models, we reran several sets of simulations, dropping the predictor accounting for most of the individual differences. Looking at the linear model with a sample size of 1000 and effect size of .38, we found that when the continuous variable accounting for the most heterogeneity was dropped, power went from 1, .94, .18, and 0 across levels of nuisance variables to under .05 for all conditions. When the binary indicator accounting for most of the variance was dropped, power went from 1, 1, .94, and .71 across levels of nuisance variables to 1, .96, .83, and .45. When even 1 important variable is left out of the PITE predictive model, the ability of the permutation test to find heterogeneity in treatment effects is reduced.

## 6. Discussion

For PITEs to be useful for quantifying individual differences in the effects of an intervention, it is necessary to have a test that can show that there are differences larger than chance between individuals. The proposed permutation test is therefore very important. Under the 96 conditions we examined, the test was shown to have appropriate, nominal type I error rates, practical utility in an applied example, and adequate power given a moderately sized sample and 20 to 50 nuisance covariates. The effect size observed in our applied example was fairly small (the average individual was 0.19 SD from the average treatment effect). However, if the effect size had been doubled, then the permutation test would have had adequate power even with 100 nuisance covariates. The permutation test also demonstrated an ability to detect heterogeneity in treatment effects due primarily to a single predictor, or when it was spread across the 7 predictors that had an impact in the ALS example. It also worked reasonably well when either the linear regression model or Random Forests were used as the predictive method. These are important because in practice, it means the test can be used with a large number of covariates when heterogeneity is due to just a small set of covariates, and it can be used with any predictive method while making no assumptions beyond those of that method. A final set of simulations illustrated that, while PITEs are inspired by potential outcomes, neither PITEs nor the permutation test allow us to detect the true individual-level causal effect. Instead, these methods detect only those individual differences which arise from the variables included in obtaining the predictions, the null hypothesis for the permutation test is that there are no individual differences for a given predictive method and set of covariates.

The permutation test was also found to have an unexpected benefit in that the variance of the PITEs across permuted datasets provides an estimate of the variability in PITEs that, due to chance, can be attributed to the number and distribution of the covariates used in a given application and to the model or algorithm used to obtain predictions. Thus, this test can help assess the amount of noise in PITEs for a given number of covariates and predictive methods.

We noted that for Random Forests with nuisance variables, the choice of tuning parameters made a meaningful difference in the results. If there is heterogeneity in treatment effects or added nuisance variables, the Random Forest requires more tuning or corrections for bias, irrespective of sample size. For instance, allowing deep trees led to high levels of overfitting, with many nuisance variables being identified as important. In this case, we chose a maximum node depth of 10 to reduce overfitting. While how to appropriately tune random forests when estimating PITEs is beyond the scope of this study, it is a non-trivial issue which merits further research.

Consistent with other studies finding evidence for heterogeneity in the effects of Riluzole with the PRO-ACT dataset (26,42), PITEs used with the permutation test suggest the possibility of individual differences in the effects of treatment. While our results suggest significant heterogeneity in the ALS dataset, the effect sizes were modest. Since this dataset consists of information pooled across a large number of clinical trials without an identifier for the clinical trial, we would like to stress that suggesting the heterogenetiy in practice is not our original intention. the result of the example dataset showed that the observed effects can be due to heterogeneity in both covariates and outcomes across trials rather than individual differences in the effects of treatment. However, further step would be needed in practice and in the substantive area, if researchers are interested in identifying the source of the heterogeneity in the ALS dataset.   that We concur with others who have warned against seeing personalized medicine as a panacea, which will result in effective treatments for everyone. (34) .

Other limitations of this study are that we examined the proposed PITE permutation test using predictions from only the linear regression model and Random Forests (with one set of tuning parameters), and under a set of conditions that were designed to clarify our understanding of its power via an applied example. In principle, we see no reason why this test should not work well with any method chosen but cannot claim that the present paper has established this. We should also note that, in the ALS example, the permutation test required a relatively large sample to attain adequate power. While the observed effect size was small in this case, even with a larger effect, the test required an N of 3,000 if many covariates were included. Because the outcome of interest is individual predictions, we believe that PITEs will generally require substantial sample sizes, unless the effects are very large. Nevertheless, as a very flexible test for the presence of individual differences, this permutation test is an important tool for personalized medicine.

# References

1. Ashley EA. The precision medicine initiative: A new national effort. JAMA - J Am Med Assoc. 2015;313(21):2119–20.

2. Smith M, Saunders R, Stuckhardt L, Mcginnis JM. Best care at lower cost: the path to continuously learning health care in America [Internet]. Vol. 51, Choice Reviews Online. 2014. 51-3277-51–3277 p. Available from: http://choicereviews.org/review/10.5860/CHOICE.51-3277

3. Carter GC, Cantrell RA, Zarotsky V, Haynes VS, Phillips G, Alatorre CI, et al. Comprehensive review of factors implicated in the heterogeneity of response in depression. Depress Anxiety [Internet]. 2012;29(4):340–54. Available from: http://ezproxy.msu.edu/login?url=http://search.proquest.com/docview/1017620373?accountid=12598

4. Liu LC, Hedeker D, Segawa E, Flay BR. Evaluation of longitudinal intervention effects: An example of latent growth mixture models for ordinal drug-use outcomes. J Drug Issues [Internet]. 2010 Jan 1;40(1):27–43. Available from: http://jod.sagepub.com/cgi/content/abstract/40/1/27

5. Taddy M, Gardner M, Chen L, Draper D. A nonparametric Bayesian analysis of heterogeneous treatment effects in digital experimentation. J Bus Econ Stat. 2016;34(4):661–72.

6. Lipkovich I, Dmitrienko A, D'Agostino RB. Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. Stat Med. 2017;36(1):136–96.

7. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search — A recursive partitioning method for establishing response to treatment in patient subpopulations. Stat Med. 2011;30:2601–21.

8. Loh W, Man M. A regression tree approach to identifying subgroups with differential treatment effects. Stat Med. 2015;34:1818–33.

9. Mayer C, Lipkovich I, Dmitrienko A. Survey Results on Industry Practices and Challenges in Subgroup Analysis in Clinical Trials. Stat Biopharm Res. 2015;7(4):272–82.

10. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc. 2018;113(523):1228–42.

11. Grimmer J, Messing S, Westwood SJ. Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods. Polit Anal. 2017;25(4):413–34.

12. Dusseldorp E, Mechelen I Van. Qualitative interaction trees: A tool to identify qualitative treatment – subgroup interactions. Stat Med. 2014;33:219–37.

13. Berger JO, Wang X, Shen L. A Bayesian Approach to Subgroup Identification. J Biopharm Stat. 2014;24(1):110–29.

14. Montgomery KL, Vaughn MG, Thompson SJ, Howard MO. Heterogeneity in drug abuse among juvenile offenders: Is mixture regression more informative than standard regression? Int J Offender Ther Comp Criminol [Internet]. 2013 Nov 1;57(11):1326–46. Available from: http://ijo.sagepub.com/cgi/content/abstract/57/11/1326

15. Xu W, Hedeker D. A random-effects mixture model for classifying treatment response in longitudinal clinical trials. J Biopharm Stat. 2001;11(4):253–73.

16. Green DP, Kern HL. Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. Public Opin Q. 2012;76(3):491–511.

17. Zvoch K. Treatment fidelity in multisite evaluation: A multilevel longitudinal examination of provider adherence status and change. Am J Eval [Internet]. 2009;30(1):44–61. Available from: http://ezproxy.msu.edu/login?url=http://search.proquest.com/docview/621868470?accountid=12598

18. Shalit U, Johansson FD, Sontag D. Estimating individual treatment effect: generalization bounds and algorithms. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: Journal of Machine Learning Reseach; 2017. p. 3076–85.

19. Duan T, Rajpurkar P, Laird D, Ng AY, Basu S. Clinical Value of Predicting Individual Treatment Effects for Intensive Blood Pressure Therapy: AMachine Learning Experiment to Estimate Treatment Effects from Randomized Data. Circ Cardiovasc Qual Outcomes. 2019;12:1–10.

20. Lamont A, Lyons MD, Jaki T, Stuart E, Feaster DJ, Tharmaratnam K, et al. Identification of predicted individual treatment effects in randomized clinical trials. Stat Methods Med Res [Internet]. 2016;0962280215623981. Available from: http://smm.sagepub.com/cgi/doi/10.1177/0962280215623981

21. Ballarini NM, Rosenkranz GK, Jaki T, Konig F, Posch M. Subgroup identification in clinical trials via the predicted individual treatment effect. PLoS One. 2018;13(10):1–22.

22. Muthén BO, Brown HC. Estimating drug effects in the presence of placebo response : Causal inference using growth mixture modeling. Stat Med. 2009;28:3363–85.

23. Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. J Am Stat Assoc. 2005;100(469):322–31.

24. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Nonparametric tests for treatment effect heterogeneity. Rev Econ Stat. 2008;90(3):389–405.

25. Dunn G, Bentall R. Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). Stat Med. 2007;2006:4719–45.

26. Seibold H, Zeileis A, Hothorn T. Individual treatment effect prediction for amyotrophic lateral sclerosis patients. Stat Methods Med Res. 2018;27(10):3104–25.

27. Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol [Internet]. 1974;66(5):688–701. Available from: http://content.apa.org/journals/edu/66/5/688

28. Holland PW. Statistics and Causal Inference: Rejoinder. J Am Stat Assoc. 1986;81(396):968.

29. Rubin DB. Causal inference using potential outcomes. J Am Stat Assoc [Internet]. 2005 Mar [cited 2013 Oct 29];100(469):322–31. Available from: http://www.tandfonline.com/doi/abs/10.1198/016214504000001880

30. Breiman L. Random Forests. Mach Learn. 2001;45:5–32.

31. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. Ann Appl Stat. 2012;6(1):266–98.

32. Zhang GP. Neural Networks for Classification : A Survey. IEEE Trans Syst Man Cybern. 2000;30(4):451–62.

33. Yin Y, Liu L, Geng Z. Assessing the treatment effect heterogeneity with a latent variable. Stat Sin. 2018;28(1):115–35.

34. Senn S. Mastering variation: Variance components and personalised medicine. Stat Med. 2016;35(7):966–77.

35. Shahn Z, Madigan D. Latent class mixture models of treatment effect heterogeneity. Bayesian Anal. 2017;12(3):831–54.

36. Haviland A, Nagin DS, Rosenbaum PR, Tremblay RE. Combining group-based trajectory modeling and propensity score matching for causal inferences in nonexperimental longitudinal data. Dev Psychol. 2008;44(2):422–36.

37. Levine SZ, Rabinowitz J, Case M, Ascher-Svanum H. Treatment response trajectories and their antecedents in recent-onset psychosis: A 2-year prospective study. J Clin Psychopharmacol [Internet]. 2010;30(4):446–9. Available from: http://ezproxy.msu.edu/login?url=http://search.proquest.com/docview/754053927?accountid=12598

38. Prince MA, Maisto SA. The clinical course of alcohol use disorders: using joinpoint analysis to aid in interpretation of growth mixture models. Drug Alcohol Depend [Internet]. 2013 Dec 1 [cited 2014 Aug 2];133(2):433–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23880249

39. Rosenbaum PR. Conditional Permutation Tests and the Propensity Score in Observational Studies. J Am Stat Assoc [Internet]. 1984;79(387):565. Available from: http://www.jstor.org/stable/2288402?origin=crossref

40. Foster JC, Nan B, Shen L, Kaciroti N, Taylor JMG. Permutation Testing for Treatment–Covariate Interactions and Subgroup Identification. Stat Biosci [Internet]. 2016;8(1):77–98. Available from: http://dx.doi.org/10.1007/s12561-015-9125-9

41. Küffner R, Zach N, Norel R, Hawe J, Schoenfeld D, Wang L, et al. Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. Nat Biotechnol. 2015;33(1):51–7.

42. Korepanova N, Seibold H, Steffen V, Hothorn T. Survival forests under test: Impact of the

proportional hazards assumption on prognostic and predictive forests for amyotrophic lateral sclerosis survival. Stat Methods Med Res. 2020;29(5):1403–19.

43. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol [Internet]. 2016;69:245–7. Available from: http://dx.doi.org/10.1016/j.jclinepi.2015.04.005

44. Ishwaran H, Kogalur UB. Fast Unified Random Survival, Regression, and Classification (RF-SRC). 2019. p. R package version 2.9.1.

Table 1. Type 1 error rates for the PITE permutation test, the linear regression model and Random Forests.

| Sample Size | Number of Nuisance Continuous Covariates | Number of Nuisance Binary Covariates | Average Treatment Effects | Upper-sided Type I Error Rate- LM | Upper-sided Type I Error Rate- RF |
|---|---|---|---|---|---|
| 100 | 0 | 0 | 0 | 0.049 | 0.046 |
| 100 | 0 | 0 | 0.5 | 0.047 | 0.029 |
| 250 | 0 | 0 | 0 | 0.047 | 0.053 |
| 250 | 75 | 35 | 0 | 0.052 | 0.060 |
| 250 | 0 | 0 | 0.5 | 0.063 | 0.054 |
| 250 | 75 | 35 | 0.5 | 0.061 | 0.030 |
| 500 | 0 | 0 | 0 | 0.048 | 0.053 |
| 500 | 75 | 35 | 0 | 0.056 | 0.047 |
| 500 | 150 | 70 | 0 | 0.052 | 0.048 |
| 500 | 0 | 0 | 0.5 | 0.053 | 0.051 |
| 500 | 75 | 35 | 0.5 | 0.054 | 0.035 |
| 500 | 150 | 70 | 0.5 | 0.050 | 0.047 |
| 1000 | 0 | 0 | 0 | 0.043 | 0.051 |
| 1000 | 75 | 35 | 0 | 0.043 | 0.046 |
| 1000 | 150 | 70 | 0 | 0.053 | 0.039 |
| 1000 | 0 | 0 | 0.5 | 0.050 | 0.038 |
| 1000 | 70 | 35 | 0.5 | 0.046 | 0.042 |
| 1000 | 150 | 70 | 0.5 | 0.062 | 0.041 |

Table 2. Power to detect heterogeneity in treatment effects, based on the linear regression model and the Random Forest predictions from the ALS example with an effect size of 0.19

| Model Prediction | Sample Size | Number of Nuisance Variables | Effect Distribution | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Spread | 90/10 Cont. | 75/25 Cont. | 50/50 Cont. | 25/75 Cont. | 90/10 Bin. |
| Linear Regression | 3,000 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| | | 20 | 1 | 1 | 1 | 1 | 1 | 0.958 |
| | | 50 | 0.924 | 1 | 0.998 | 0.968 | 0.988 | 0.878 |
| | | 100 | 0.256 | 0.002 | 0.002 | 0.002 | 0.498 | 0.658 |
| | 1,000 | 0 | 0.908 | 1 | 1 | 1 | 0.996 | 0.648 |
| | | 20 | 0.294 | 0.062 | 0.058 | 0.098 | 0.496 | 0.392 |
| | | 50 | 0.008 | 0 | 0 | 0 | 0.006 | 0.182 |
| | | 100 | 0 | 0 | 0 | 0 | 0 | 0.112 |
| Random Forest | 3,000 | 0 | 0.96 | 1 | 1 | 0.97 | 1 | 0.99 |
| | | 20 | 0.94 | 0.91 | 0.91 | 0.90 | 0.97 | 0.91 |
| | | 50 | 0.46 | 0.28 | 0.32 | 0.29 | 0.47 | 0.26 |
| | | 100 | 0.01 | 0.00 | 0.00 | 0.02 | 0.01 | 0.00 |
| | 1,000 | 0 | 0.23 | 0.08 | 0.06 | 0.10 | 0.44 | 0.42 |
| | | 20 | 0.16 | 0.03 | 0.06 | 0.09 | 0.19 | 0.12 |
| | | 50 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | 100 | 0 | 0 | 0 | 0 | 0 | 0 |

| Ignoring One Heterogeneous Variable | Effect Size | Number of Nuisance Variables | Effect Distribution | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Spread | 90/10 Cont. | 75/25 Cont. | 50/50 Cont. | 25/75 Cont. | 90/10 Bin. |
| Ignoring the Last Binary Variable | 0.38 | 0 | 1 | 1 | 1 | 1 | 1 | 0.998 |
| | | 20 | 1 | 1 | 1 | 1 | 1 | 0.962 |
| | | 50 | 0.488 | 0.046 | 0.044 | 0.082 | 0.322 | 0.828 |
| | | 100 | 0 | 0 | 0 | 0 | 0 | 0.452 |
| | 0.19 | 0 | 0.996 | 1 | 1 | 1 | 0.998 | 0.972 |
| | | 20 | 0.548 | 0.128 | 0.152 | 0.242 | 0.694 | 0.804 |
| | | 50 | 0.012 | 0 | 0 | 0 | 0.016 | 0.544 |
| | | 100 | 0 | 0 | 0 | 0 | 0 | 0.276 |
| Ignoring the First Continuous Variable | 0.38 | 0 | 0.96 | 0.044 | 0.048 | 0.106 | 0.476 | 0.9 |
| | | 20 | 0.554 | 0.046 | 0.052 | 0.066 | 0.156 | 0.666 |
| | | 50 | 0.142 | 0.048 | 0.046 | 0.058 | 0.05 | 0.388 |
| | | 100 | 0.008 | 0.044 | 0.046 | 0.04 | 0.01 | 0.196 |
| | 0.19 | 0 | 0.504 | 0.036 | 0.038 | 0.048 | 0.12 | 0.65 |
| | | 20 | 0.144 | 0.04 | 0.046 | 0.05 | 0.042 | 0.408 |
| | | 50 | 0.042 | 0.046 | 0.046 | 0.044 | 0.016 | 0.222 |
| | | 100 | 0.006 | 0.038 | 0.038 | 0.036 | 0.006 | 0.124 |

Table 3. Power to detect heterogeneity in treatment effects, based on the linear regression model and the Random Forests predictions from the ALS example with an effect size of 0.38

| Model Prediction | Sample Size | Number of Nuisance Variables | Effect Distribution | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Spread | 90/10 Cont. | 75/25 Cont. | 50/50 Cont. | 25/75 Cont. | 90/10 Bin. |
| Linear Regression | 3,000 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| | | 20 | 1 | 1 | 1 | 1 | 1 | 1 |
| | | 50 | 1 | 1 | 1 | 1 | 1 | 1 |
| | | 100 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 1,000 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| | | 20 | 1 | 1 | 1 | 1 | 1 | 0.996 |
| | | 50 | 0.682 | 0.106 | 0.108 | 0.158 | 0.582 | 0.944 |
| | | 100 | 0 | 0 | 0 | 0 | 0 | 0.706 |
| Random Forest | 3,000 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| | | 20 | 1 | 1 | 1 | 1 | 1 | 1 |
| | | 50 | 1 | 1 | 1 | 1 | 1 | 1 |
| | | 100 | 0.81 | 0.60 | 0.55 | 0.52 | 0.66 | 0.74 |
| | 1,000 | 0 | 1 | 1 | 1 | 1 | 1 | 0.99 |
| | | 20 | 0.97 | 0.94 | 0.98 | 0.98 | 0.99 | 0.99 |
| | | 50 | 0.29 | 0.18 | 0.11 | 0.16 | 0.24 | 0.24 |
| | | 100 | 0 | 0 | 0 | 0 | 0 | 0 |

APPENDIX

A. Parameter estimates from the linear regression model using the ALS example with 17 selected covariates.

| Covariates | Treatment Group | | Control Group | |
|---|---|---|---|---|
| | Coefficients | SE | Coefficients | SE |
| (Intercept) | -3.56945 | 1.16889 | -2.88826 | 1.45050 |
| Delayed Medication | 0.01969 | 0.07101 | 0.17810 | 0.11253 |
| Respiratory Rate | -0.00099 | 0.00422 | 0.01067 | 0.00501 |
| Temperature | 0.10752 | 0.02962 | 0.09247 | 0.03714 |
| Weight(kg) | 0.00226 | 0.00102 | 0.00166 | 0.00129 |
| Height(cm) | -0.00555 | 0.00208 | -0.00442 | 0.00271 |
| Diastolic Blood Pressure | -0.00311 | 0.00160 | -0.00204 | 0.00211 |
| Systolic Blood Pressure | 0.00110 | 0.00105 | -0.00113 | 0.00133 |
| Pulse | -0.00365 | 0.00117 | -0.00431 | 0.00150 |
| Gender | 0.00712 | 0.03857 | -0.03439 | 0.04959 |
| Age | 0.00130 | 0.00131 | -0.00327 | 0.00162 |
| White | 0.03524 | 0.06360 | -0.01493 | 0.06223 |
| severity | -0.04591 | 0.02726 | -0.06893 | 0.03656 |
| Diagnosis Delta | -0.00036 | 0.00009 | -0.00022 | 0.00011 |
| Limb Only | -0.08336 | 0.11539 | 0.08838 | 0.04050 |
| Bulbar Only | -0.33667 | 0.11856 | -0.08348 | 0.04985 |
| Start Delta | -0.00019 | 0.00011 | -0.00037 | 0.00018 |
| Use Riluzole | -0.07150 | 0.03861 | -0.22549 | 0.05578 |

B. Means and SDs of the continuous covariates in the ALS example

| | Mean | SD | Median | Min. | Max. |
|---|---|---|---|---|---|
| Systolic Blood Pressure | 131.88 | 16.63 | 130 | 85 | 206 |
| Age | 54.70 | 11.35 | 55 | 18 | 80 |
| Respiratory Rate | 17.19 | 3.27 | 16 | 6 | 42 |

C. Distribution of binary covariates in the ALS example

| | Category | N | Percentage |
|---|---|---|---|
| Delayed Medication (longer than 1 year) | Yes | 93 | 3.20% |
| | No | 2817 | 96.80% |
| Limb Only | Yes | 1952 | 67.08% |

| | | | |
|---|---|---|---|
| | No | 958 | 32.92% |
| Gender | Male | 1848 | 63.51% |
| | Female | 1062 | 36.49% |
| Use Riluzole | Yes | 1112 | 38.21% |
| | No | 1798 | 61.79% |

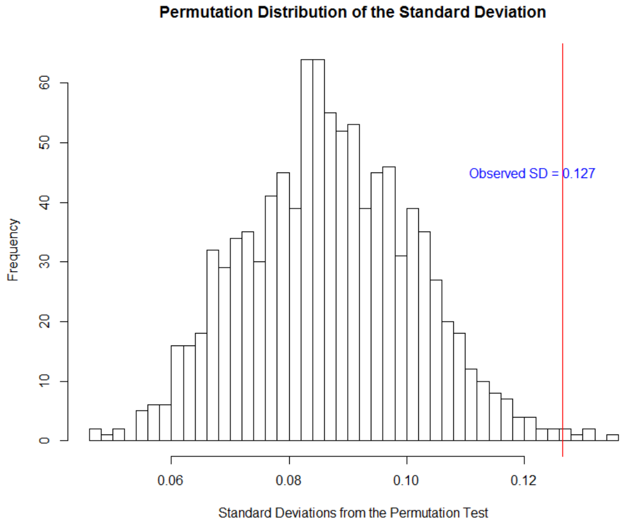

*Figure 1*. Permutation distribution of PITEs' SDs and the observed PITE's SD in the ALS study.