Research on Machine Learning for Biomedical Research

ZHAO, Kai

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
School of Biomedical Sciences

Supervised by

Prof. So Hon-cheong

The Chinese University of Hong Kong
July 2020

Thesis Assessment Committee

Professor Cheng Sze Lok Alfred (Chair)

Professor So Hon-cheong (Thesis Supervisor)

Professor Chen Yangchao (Committee Member)

Professor Sham Pak Chung (External Examiner)

Abstract of thesis entitled:

Research on Machine Learning for Biomedical Research

Submitted by ZHAO, Kai

for the degree of Doctor of Philosophy

at The Chinese University of Hong Kong in July 2020

This is the abstract in no more than 350 words.

Acknowledgement

I would like to thank my supervisor Prof. So Hon-cheong for offering me the opportunity to work with him. He teaches me how to conduct research works and gave me generous help in my daily life. He is humorous and highly empathic. Truth to be told, it is a wonderful journey to study and work with him, and I am lucky to have the experience.

I also would like to thank my wife. It's nearly ten years since the first meet, and she has become the most important person in my life. Thanks for supporting my decision to pursue a higher degree and bearing all burden from family to free me from distractions. This is not easy to her. You are a brave girl and wonderful mother for the kids. My any achievement is impossible without you.

Thanks my father and mother for never saying NO to the decision of further my education and for taking care of my kids in the past years. I know the hardness for you to bear the burden. I highly appreciate the support.

Thanks my kids for tolerating my absence of parental responsibility for the past year. You let me be your father and share tremendous joy with me. I promise my love to you will alway be the same.

Thanks my lab mates for having some awesome years with you. You are a part of my daily life in those years. The times we spent together will be a precious memory.

Thanks everybody who helped me for their kindness!

Dedicated to those who risked their life to fight against COVID-19.

Contents

Abstract i

Acknowledgement ii

Symbols and Acronyms viii

1 Introduction 1

2 Background Study 2

3 Drug Repurposing 3

4 Drug Target Discovery 4

5 Evaluating ITE of Genetic RFs on Survival 5

5.1 Motivation 5
5.2 Background 6
5.3 Overview of Related Work 8
5.3.1 Background Methods 8
5.3.2 Causal Forests 11
5.4 ITE Framework 14
5.4.1 Novel Tests for the presence of heterogeneity . 16
5.4.2 Modeling Survival Outcomes 19
5.5 Experiment Results 20
5.5.1 Simulations Studies 20
5.5.2 Applications on Real Data 26

5.6 Conclusion 26

6 Conclusion **27**

A Proof of Propositions **28**

B Publication List **29**

Bibliography **30**

List of Figures

List of Tables

5.1 Specifications for simulation scenarios 23
5.2 Comparison of power/type I error rate of different tests for the presence of heterogeneity 25

Symbols and Acronyms

In general, we denote a scalar by an italic lower case letter, a vector by a roman lower case bold letter, and a matrix by a roman upper case bold letter respectively, e.g., $a \in \mathbb{R}$, $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{M} \in \mathbb{R}^{p \times q}$, with any exceptions to be mentioned in the context case by case.

An identity matrix is written as $\mathbf{I}$. Specifically, an $n \times n$ identity matrix is written as $\mathbf{I}_n$. A zero matrix or vector is written as $\mathbf{0}$. Specifically, an $m \times n$ zero matrix is written as $\mathbf{0}_{m \times n}$.

Specialized symbols and major acronyms are defined as follows:

$p(\cdot)$	the probability density function (PDF)
$\Pr(\cdot)$	the probability value
$\mathbb{E}(\cdot)$	the expectation
$\mathbf{\Sigma}$	a covariance matrix
$\mathbf{N}(\mu, \mathbf{\Sigma})$	a normal distribution with mean μ and covariance $\mathbf{\Sigma}$
ε	a noise vector
$\mathbf{e}(\cdot)$	an error/residual function
$\mathbf{H}$	Hessian matrix
$\mathrm{tr}(\cdot)$	trace of a matrix
$\det(\cdot)$	determinant of a matrix

DNN	deep neuron network
GBM	gradient boosting machine
SVM	support vector machine
CF	causal forests
RF	risk factors
ML	machine learning
EN	elastic net
ITE	individual treatment effects
TE	treatment effects
tx	treatment
CM	cardiometabolic
GWAS	genome-wide association studies
CNV	copy number variation
SML	supervised machine learning
MCMC	Markov chain Monte Carlo
GRF	general causal forests
CF	causal forests
CV	cross validation
MSE	mean squared error
CI	confidence interval

Chapter 1

Introduction

□ End of chapter.

Chapter 2

Background Study

□ End of chapter.

Chapter 3

Drug Repurposing

□ End of chapter.

Chapter 4

Drug Target Discovery

□ End of chapter.

Chapter 5

Evaluating ITE of Genetic RFs on Survival

5.1 Motivation

Traditional biomedical or clinical studies in the area of estimating treatment effect mainly focus on the average effect of risk factors (RFs) or treatment (tx) in population level. However, in the clinical environment we can easily find that the same risk factor may affect patients differently. Thus, patients may pay more attention to how a risk factor will affect them in an individual level rather than in a population level, given their clinical backgrounds and genetic characteristics. The main aim of this study is to resolve this concern by estimating the ITEs for each patient, with consideration of their unique genetic and clinical information. In this study we treat the two terms ”risk factor” and ”treatment” conceptually equivalent, since a risk factor can be considered as a ”treatment” with adverse effects. The approach for estimating the ITEs for each patient allows us to offer tailored health management to individual patients. This enables us to deliver more cost-effective

prevention or treatment strategies to benefit them the most. This idea is also in the line with ”personalized medicine”, which has been advocated in recent years.

In spite of an increasing number of studies in this area, current studies in ITE are rather limited. Some critical limitations include a lack of well-established validation methods for treatment effect estimations and the contribution of key features ITE estimation, and failure in incorporating censored data. Even though genetic factors may determine heterogenous response to tx/RFs, especially to cancer treatments [12], current studies on ITE have not included genomic features. Here we proposed several methodologies to address the above limitations and applied the ITE framework to genomic data. In our approach genomic features were considered as risk factors or covariates that contribute to the heterogeneity of treatment effect.

5.2 Background

It has been well-known that different individuals response differently to the same risk factor (RF) or treatment (tx). For example, even though obesity is a risk factor for cardiometabolic (CM) diseases, there still are obese subjects who don't develop related complications [22]. The type and severity of such CM complications can also show heterogeneity among subjects [22]. Another evidence is that not all people suffering stressful life events are affected by depression, even if stressful events are risk factors for depression [34]. This fact can also be applied to other RFs or treatments. The heterogenous effect can be contributed by

different genetic and/or environmental factors of subjects, and these factors affect them differently. Here we would like to investigate the different treatment effect contributed by variants or mutations instead of clinical factors, since studies have shown that same variant/mutation can have varying outcomes on different subjects [1, 24, 32].

There are dramatic advances in the omics technology and a shape rising availability of biomedical data. However, current studies in cancer still mainly focus on one clinical/genetic RF at a time, without the consideration of presence of complex interactions among the subject's genetic and/or clinical factors.

One of most crucial concerns to patients is how a RF or treatment will affect them given their genetic and clinical information. However, current researches on this issue largely focus on the average treatment effect of RFs in population rather than individualized treatment effects.

Here we built a computational framework to unravel the individualized effects of RFs/treatment so that we can estimate the treatment effect for each individual with the incorporation of his/her genetic and/or clinical background. We also developed methods to discover genetic and/or clinical features contributing the most to the estimation. We employed our approaches to cancer data to estimate treatment effects of genetic changes (e.g. changes of expression level of risk genes, mutations, CNVs etc.) and other RFs on each individual's survival.

5.3 Overview of Related Work

5.3.1 Background Methods

Here we define notations for the clarification of following presentation. Let $\mathbf{X}^{n\times m}$ denotes the covariate variable, $\mathbf{Y}^n$ denotes outcome variable, and $\mathbf{W}^n$ denotes treatment variable. Given an observation i we denote its covariates as $\mathbf{x_i}^m$, the risk factor/tx status as w_i and outcome y_i. Here we restate that since a risk factor may also be considered as a "treatment" with adverse effects, methods for ITE estimation can also be employed to RFs. Assume that the outcome $\mathbf{Y}$ satisfy

$$\mathbf{Y} = f(\mathbf{X}, \mathbf{W}) + \varepsilon, \tag{5.1}$$

Where ε follows $N(\mathbf{0}, \mathbf{\Sigma})$. $\mathbf{\Sigma}$ is the covariance matrix.

Assumption assume $\mathbf{X}, \mathbf{W}, \mathbf{Y}$ fulfill unconfoundedness assumption (randomization conditional on the covariates),

$$\big[\mathbf{Y}_i(1), \mathbf{Y}_i(0)\big] \perp\!\!\!\perp \mathbf{W}_i \,|\, \mathbf{X}_i. \tag{5.2}$$

Under the unconfoundedness assumption 5.2 the key is to estimate the expected difference, the estimation of ITE, for each individual in response between treatment and control. The ITE for subject i is formulated as

$$\tau(\mathbf{x_i}) = \mu_1(\mathbf{x_i}) - \mu_0(\mathbf{x_i}), \tag{5.3}$$

with $\mu_1(x_i)$ and $\mu_0(x_i)$ defined as

$$\begin{aligned}\mu_1(x_i) &= \mathbb{E}(\mathbf{Y} = y_i|\mathbf{X} = \mathbf{x_i}, \mathbf{W} = w_i) = f(x_i, 1)\\ \mu_0(x_i) &= \mathbb{E}(\mathbf{Y} = y_i|\mathbf{X} = \mathbf{x_i}, \mathbf{W} = 1 - w_i) = f(x_i, 0)\end{aligned} \tag{5.4}$$

respectively, where $w_i = 1$ without the loss of generality. For a given subject i, y_i and w_i are scalars, and $\mathbf{x_i}$ is a vector of length m. Traditional machine learning method cannot handle this situation since they cannot capture the difference of ITE when the outcomes for RF/tx were absent or present.

A traditional solution to measure ITE is to estimate the difference of averaged outcome between treatments and controls in pre-specified subgroups [14] or subgroup defined by learning algorithms [29, 28, 3, 13]. Su et. al. employed interaction tree to iteratively searching subgroups based on treatment effect [29, 28]. Similarly, causal trees proposed by Athey and Imbens estimate the treatment effect at the leaves of the tree [3]. However, main drawbacks of the approach are that there is no ground truth for subgroup definition, and that the impediment of iteratively searching for subgroups present obvious treatment effect and reporting only the results for subgroups with extreme treatment effects to highlight heterogeneity may be highly spurious [2, 9]. In the high dimensional setting, it's still very challenging to divide subjects into appropriate subgroups [25]. It's the same case for genomic data.

Alternatively, a feasible approach is to use any supervised machine learning (SML) methods to fit $\mu_1(\mathbf{x})$ and $\mu_0(\mathbf{x})$ simultaneously / separately and estimate the difference by putting them together. Specifically, one may fit a single model $\mu(\mathbf{x}, w)$ or separate models for the

treated and control groups, and compute the different between $\mu(\mathbf{x}, w)$ and $\mu(\mathbf{x}, 1-w)$. Studies [21, 10] utilize different counterfactual random forests algorithms to estimates treatment effects by fitting separate random forests models to treatment and control groups. Several literature, including Green and Kern [15], Hill [17], and Hill and Su [16], has employed bayesian forest-based machine learning methods to estimate heterogeneous treatment effects. These studies utilize Bayesian additive regression tree (BART) method [8], and can obtain reliable intervals for treatment effects by MCMC sampling. For lasso-like methods for causal inference [18, 30], it's difficult to capture interactions, which may be naturally present in genomic data, in high-dimensional setting, in split of its simplicity and good ability in feature selection. A limitation of these studies is lack of formal statistical inference results [33]. Some other methods, like Meta learners and deep learning based, for ITE estimation could be found in [20, 19].

Here we are interested in methods with following characteristics: automatically select important features, well capture high interactions present, and have good asymptotic properties. Thus, methods such as causal forests [33] or GRF [4] are much more preferred. Causal forests [33] have been proposed with an objective to maximize the heterogeneity of $\tau(\mathbf{x})$. Recently, Athey et al. proposed an extension of causal forests, GRF [4], inspired by the R-learner proposed in [23]. Both of the two methods [33, 4] inherits the excellent capability of random forest in capturing complex interactions.

However, there are still substantial research gaps, including relative lack of methods for result validation, evaluation of key features

contributing to ITE estimation and handling censored data. Here we proposed methods to address these key issues and pioneer new applications to genomic data, which is the first of its kind.

5.3.2 Causal Forests

In this section, we will explain causal forests (CF) technically, a basis of our ITE framework. CF [33] originate from random forests [6], which are related to kernel or nearest neighborhood methods. However, random forests differ in that they determine weights received by nearby observations in a data-driven way, and this characteristics is critical in high dimensional environment or the present of high order interactions among covariates [33]. This is the same case for CF.

Here we begin with causal trees (CT) [3] since CF are made of a number of CT. In this part we follow a similar notations as appeared in [3]. A tree can be considered as a partitioning of the feature space $\mathbb{X}$, denoting as Π. A partition Π with a number of elements $\#(\Pi)$ can be written as

$$\Pi = \{l_1, l_2, \ldots, l_{\#(\Pi)}\},$$

and a union of all elements in partition is the whole feature space $\mathbb{X}$.

Let $\mathbb{P}$ denote the space of partitions, and $\mathbb{S}$ be the space of samples from a population of observations. We seek for a algorithm $\pi : \mathbb{S} \to \mathbb{P}$ that splits sample space $\mathbb{S}$ into partition Π.

Given a partition Π and sample S, the estimated conditional mean

for observation $\mathbf{x}$ is

$$\hat{\mu}(\mathbf{x};\Pi) \equiv \frac{1}{\#(i \in \mathrm{S} : \mathbf{x}_i \in l(\mathbf{x};\Pi))} \sum_{i \in \mathrm{S}:\mathbf{x}_i \in l(\mathbf{x};\Pi)} y_i,$$

which is an unbiased estimator for $\mu(\mathbf{x};\Pi)$. Here $l(\mathbf{x};\Pi)$ is the leaf to which $\mathbf{x}$ belongs.

A adjusted MSE criteria including $\mathbb{E}[\mathbf{Y}_i^2]$, a term that does not depend on the estimator, is defined as

$$\mathrm{MSE}_\mu(\mathrm{S}^{\mathrm{te}}, \mathrm{S}^{\mathrm{est}}, \Pi) \equiv \frac{1}{\#(\mathrm{S}^{\mathrm{te}})} \sum_{i \in \mathrm{S}^{\mathrm{te}}} \left\{ (y_i - \hat{\mu}(\mathbf{x}_i; \mathrm{S}^{\mathrm{est}}, \Pi))^2 - y_i^2 \right\}. \quad (5.5)$$

The expectation of the modified MSE is

$$\mathrm{EMSE}_\mu(\Pi) \equiv \mathbb{E}_{\mathrm{S}^{\mathrm{te}}, \mathrm{S}^{\mathrm{est}}} \left[\mathrm{MSE}_\mu(\mathrm{S}^{\mathrm{te}}, \mathrm{S}^{\mathrm{est}}, \Pi) \right].$$

The objective of CT is to maximize the criteria

$$Q^{\mathrm{H}}(\pi) \equiv -\mathbb{E}_{\mathrm{S}^{\mathrm{te}}, \mathrm{S}^{\mathrm{est}}, \mathrm{S}^{\mathrm{tr}}} \left[\mathrm{MSE}_\mu(\mathrm{S}^{\mathrm{te}}, \mathrm{S}^{\mathrm{est}}, \pi(\mathrm{S}^{\mathrm{tr}})) \right]. \quad (5.6)$$

This criteria shows better convergence properties of confidence intervals, compared with conventional practice that $\mathrm{S}^{\mathrm{est}}$ and S^{tr} are the same sample for both tree construction and estimation [33].

With the above setup for treatment effect estimation, a similar definition to 5.5, is defined as

$$\mathrm{MSE}_\tau(\mathrm{S}^{\mathrm{te}}, \mathrm{S}^{\mathrm{est}}, \Pi) \equiv \frac{1}{\#(\mathrm{S}^{\mathrm{te}})} \sum_{i \in \mathrm{S}^{\mathrm{te}}} \left\{ (\tau_i - \hat{\tau}(\mathbf{x}_i; \mathrm{S}^{\mathrm{est}}, \Pi))^2 - \tau_i^2 \right\}. \quad (5.7)$$

In reality, τ_i cannot be directly observed. The estimated counterparts

are defined as

$$\hat{\tau}(\mathbf{x}; \mathrm{S}, \Pi) \equiv \hat{\mu}(1, \mathbf{x}, \mathrm{S}, \Pi) - \hat{\mu}(1, \mathbf{x}; \mathrm{S}, \Pi), \tag{5.8}$$

where

$$\hat{\mu}(w, \mathbf{x}, \mathrm{S}, \Pi) \equiv \frac{1}{\#(i \in \mathrm{S_w} : \mathbf{x}_i \in l(\mathbf{x}; \Pi))} \sum_{i \in \mathrm{S_w} : \mathbf{x}_i \in l(\mathbf{x}; \Pi)} y_i^{\mathrm{obs}}.$$

With the fact that $\hat{\tau}$ is constant within each leaf and the fact that

$$\mathbb{E}_{\mathrm{S^{te}}}[\tau_i | i \in \mathrm{S^{te}} : i \in l(\mathbf{x}, \Pi)] = \mathbb{E}_{\mathrm{S^{te}}}[\hat{\tau}(\mathbf{x}; \mathrm{S^{te}}, \Pi)],$$

A crucial estimator for the infeasible in-sample goodness-of-fit criterion is derived as

$$-\mathrm{MSE}_\tau(\mathrm{S^{tr}}, \mathrm{S^{tr}}, \Pi) = \frac{1}{N^{\mathrm{tr}}} \sum_{i \in \mathrm{S^{tr}}} \hat{\tau}^2(\mathbf{x}_i; \mathrm{S^{tr}}, \Pi). \tag{5.9}$$

Then this leads to an estimator for the criterion relying only on $\mathrm{S^{tr}}$ and N^{est}

$$\begin{aligned} -\hat{\mathrm{EMSE}}_\tau(\mathrm{S^{tr}}, N^{\mathrm{est}}, \Pi) \equiv & \frac{1}{N^{\mathrm{tr}}} \sum_{i \in \mathrm{S^{tr}}} \hat{\tau}^2(\mathbf{x}_i; \mathrm{S^{tr}}, \Pi) - \\ & \left(\frac{1}{N^{\mathrm{tr}}} + \frac{1}{N^{\mathrm{est}}}\right) \sum_{l \in \Pi} \left(\frac{\mathrm{S}^2_{\mathrm{S^{tr}_{tx}}}(l)}{p} + \frac{\mathrm{S}^2_{\mathrm{S^{tr}_{ctl}}}(l)}{1-p} \right) \end{aligned} \tag{5.10}$$

where the last term in the second line of 5.10 is pooled within-leaf variance. Definitely the splits of the tree are chosen to maximize the variance of $\tau(\mathbf{x}_i)$. For more detailed derivations, refer to original publication [3].

Briefly, the objective for splitting for the adaptive version of CT, de-

noted CT-A, uses $-\text{MSE}_\tau(\text{S}^{\text{tr}}, \text{S}^{\text{tr}}, \Pi)$. The same objective function also is applicable to CV version of CT-A but evaluated at the samples $\text{S}^{\text{tr,cv}}$ and $\text{S}^{\text{tr,cv}}$. The splitting objective function for the honest CTS, CT-H, is $-\text{MSE}_\tau(\text{S}^{\text{tr}}, N^{\text{est}}, \Pi)$. The same objective function also can be applied to a CV version of CT-H, but evaluated at the cross-validation sample $\text{S}^{\text{tr,cv}}$ with known $N^{\text{est,cv}}$.

Procedure for causal forests with honesty and subsampling is proposed as follows: In above table, CT stands for causal tree algorithm

Algorithm 1 Causal forests with honesty and subsampling

Require: a samples $\text{S}^{n\times m}$ and pre-specified parameters, including the number of trees B, *mtry* and the sub-sampling rate s

1: **for all** i such that $0 \leq i \leq \text{B}$ **do**
2: samples for model construction $\text{S}_\text{i}^{\text{tr,est}} \leftarrow$ SUBSAMPLE (S, s), with remaining as test set $\text{S}_\text{i}^{\text{te}}$
3: training samples $\text{S}_\text{i}^{\text{tr}}, \text{S}_\text{i}^{\text{est}} \leftarrow$ SUBSAMPLE ($\text{S}_\text{i}^{\text{tr,est}}$)
4: causal tree $\text{T}_\text{i} \leftarrow \text{CT}(\text{S}_\text{i}^{\text{tr}}, \text{S}_\text{i}^{\text{est}})$
5: make out-of-bag prediction $\hat{\tau}(\mathbf{x}_j^i)$ for $\mathbf{x}_j^i \in \text{S}_\text{i}^{\text{te}}$
6: **end for**
7: **return** CTs $\text{T}_1, \text{T}_2, \ldots, \text{T}_\text{i}, \ldots, \text{T}_\text{B}$ and $\hat{\tau}(\mathbf{x}_j)$ by averaging all available out-of-bag predictions $\tau_j^{(.)}$ for $\mathbf{x}_j \in S$.

defined above in the section.

5.4 ITE Framework

In order to assess the ITE of RFs on disease outcome and discover key features that contribute to estimation of the heterogeneity we proposed an analytic framework to estimate the ITE of RFs/tx, with genetic features as primary RFs and/or covariates. The term 'individualized treatment effect' (ITE) will be used regardless of an RF or treatment being

considered, since we have declared that the two are conceptually equivalent entities in this study.

The estimated ITE may be formulated as 5.4 under a counterfactual outcomes framework [26]. In reality the true value of $\tau(\mathbf{x})$ cannot be directly observed, as we only have one of the two potential outcomes. In observational studies, the tx assignment may be associated with potential outcomes due to confounding variables. If the unconfoundedness assumption 5.2 satisfies then the causal ITE can still be captured. Then, in most observational studies, the study is still of significance in despite of the presence of residual confounding.

In that case we may still gain insights into TE heterogeneity at an association level, and covariates responsible for heterogeneity may still deserve further studies, and in practice a continuous variable $\mathbf{W}$ is also allowed [4]. When $\mathbf{W}$ is continuous, an average partial effect is estimated $Cov[\mathbf{Y}, \mathbf{W} \mid \mathbf{X} = \mathbf{x}]/Var[\mathbf{W} \mid \mathbf{X} = \mathbf{x}]$, which may be considered as the increase in Y given a unit increase in W, conditional on the covariates.

Our experiment studies rely on the framework of GRF, as it is a state-of-the-art approach which directly optimizes an objective function for ITE estimation, rather than fit conditional means for treatment and control observations. However, most of the methodologies and extensions presented in this study is data-driven such that it can be widely applied to any other ITE estimation models.

5.4.1 Novel Tests for the presence of heterogeneity

Unlike a SML model, an ITE model is not straightforward since the actual TE is not directly observed. There is no either empirically or theoretically well-established methods in ITE validation. We notice that the function `test_calibration` provided in the R package `grf` [31], and the idea behind it is borrowed from [7]. Thus, we would compare our methods with it in simulations in future section 5.5.

We would propose several novel statistical tests for ITE model evaluation:

Split-half correlation with multiple splits This method borrows the idea of cross validation (CV) in SML. Briefly, we proposed to split the dataset into 2 halves. Specifically, an ITE model is first fitted on the 1st half, then applied to the 2nd half to predict $\tau(\mathbf{x})$. The process can be repeated by reversing the training and testing sets. Then for each half, we have out-of-bag predictions $\tau_{\text{oob}}(\mathbf{x})$ and predictions $\tau_{\text{pred}}(\mathbf{x})$ by applying models fitted on the other half to it. We can assess the model fitting by examining the correlation between the two $\tau(\mathbf{x})$s.

The rational behind this statistical test for the presence of heterogeneity is that we expect the ITE to vary randomly around a constant in the absence of heterogenous treatment effects that can be explained by covariates, and hence the correlation between the two estimated τ values should be close to 0; otherwise, the stability and generality of ITE models can be examined by checking the replicability on an independent dataset using split-half correlation.

In order to reduce random variations that may be introduced by a single split, we performed split-half correlation test with multiple splits

on the data in practice and combined the results together. Standard Simes test [27] may be an options for the combination, since it is robust to positive dependency of p-values. Its alternative hypothesis (H1) is that at least one of the hypotheses is non-null (at least one out of n splits yields a positive significant correlation), so it may be relatively loose for our problem. To increase stringency, we introduced a partial conjunction test (partial Simes test) based on the work from Benjamini et al. [5], whose H1 assumes that at least r out of the n splits yield a positive significant correlation.

The threshold r defines the stringency/level of consistency for a finding that deserves further study. In experiment we set r to be 10% of n, which can give adequate type I error control, and employed 3 correlation measures (Pearson, Spearman and Kendall) to evaluate the split-half correlation.

A new permutation framework We proposed a novel permutation statistical test to assess the presence of heterogeneity. That is, it can be used to assess whether the predicted ITE are significantly better than predictions assuming a constant TE (which is the norm in most studies).

The objective of CF is to maximize $\widehat{\mathrm{Var}}(\tau)$, as stated in section 5.3.2. When there is no heterogeneity that can be explained by covariates, $\widehat{\mathrm{Var}}(\tau)$ should be low and close to 0. Equivalently, in this situation $\widehat{\mathrm{Var}}(\tau)$ is roughly equal to $\widehat{\mathrm{Var}}(\tau)$ yielded by model fitted on arbitrary covariates. Thus, a permutation approach we proposed is based on this rational to test the significance of $\widehat{\mathrm{Var}}(\tau)$ observed.

To model the null hypothesis we shuffled the covariates for each

permutation, such that there is no heterogeneity can be explained by covariates, and then computed the $\widehat{\text{Var}}_{\text{observed}}$ for the permuted data.

If we repeat this process N times, then a probability for the null hypothesis of $\widehat{\text{Var}}(\tau)$ statistical test is defined as

$$\Pr(\text{null}_{\text{var}}) = \frac{\#(\widehat{\text{Var}}_{\text{perm}}(\hat{\tau}) \geq \widehat{\text{Var}}_{\text{observed}}(\hat{\tau})}{N}, \tag{5.11}$$

A related but clinically relevant question is: whether the model that allows ITE outperforms that predicting a constant ITE? That is, whether patients benefit more with the introduction of individualized treatments than a conventional treatments with a consideration of averaged treatment effects only. In ordinary regression problem, the goodness-of-fit of model is assessed by the mean squared error (MSE) between the expected outcome and predictions, so it's preferred to compute the mean squared error (MSE) between the true and estimated τ for model assessment. If ITE model has a lower MSE than a constant model, which assumes that the ITE is the same for every subject, then ITE model outperforms the constant one. In reality, the true $\tau(x)$ cannot be directly observed, but Nie et al. [23] proposed that the MSE between the true and estimated $\tau(\mathbf{x})$, or $\tau_{\text{risk}}(\mathbf{x})$ can be defined as

$$\begin{aligned} \hat{\tau}_{\text{risk}} &= \sum_i \left((y_i - \hat{y}(\mathbf{x}_i) - \big(w_i - \hat{w}(\mathbf{x}_i)\big)\hat{\tau}(\mathbf{x}_i) \right)^2 \\ &= \sum_i \left(\tilde{y}_i - \tilde{w}_i \hat{\tau}(\mathbf{x}_i) \right)^2. \end{aligned} \tag{5.12}$$

Here $\hat{y}(\mathbf{x}_i)$ is an estimate of $\mathbb{E}(y_i|\mathbf{X} = \mathbf{x}_i)$ and $\hat{w}(\mathbf{x}_i)$ is an estimate of $\Pr(w_i = 1 \mid \mathbf{X} = \mathbf{x}_i)$ by any SML models. For simplicity, $\tilde{y}_i$ stands for

$y_i - \hat{y}(\mathbf{x}_i)$, and $\tilde{w}_i(\mathbf{x}_i)$ for $w_i - \hat{w}(\mathbf{x}_i)$. The two terms can be regarded as residualized outcome and treatment respectively. These quantities are out-of-bag estimations. We can then compute the τ_{risk} assuming an unrestricted $\hat{\tau}(\mathbf{x})$ estimated from an ITE model and a constant effect based on the **average treatment effect (ATE)** $\bar{\tau}(\mathbf{x})$. We proposed the following definition to assess the improvement in τ_{risk} due to the incorporation of ITE versus a constant treatment effect

$$\hat{\tau}_{\text{improve}} = \sum_i (\tilde{y} - \tilde{w}\tau(\hat{x}_i))^2 - \sum_i (\tilde{y} - \tilde{w}\tau(\bar{x}_i))^2. \tag{5.13}$$

To model the null distribution of $\hat{\tau}_{\text{improve}}$, we propose a permutation approach in which permuted $\hat{\tau}_{\text{improve}}$s are obtained by shuffling the covariates for a predefined number of times. The null hypothesis of above test is that there is no statistical difference between the observed $\hat{\tau}_{\text{improve}}$ and permuted $\hat{\tau}_{\text{improve}}$s. Note that the test does not require any distributional assumptions.

An adaptive permutation strategy with early stopping was adopted to reduce the computing time. Permutations will be stopped earlier if the result is unlikely to be significant in future runs. Specifically, we will calculate a 99% CI for the permutation p-value after each $k(k \ll N)$ runs. Here N is the total number permutations. If the lower CI > 0.05, the permutation will be terminated early.

5.4.2 Modeling Survival Outcomes

Time-to-event data are common in biomedical research, and standard ITE estimation methods may not work on survival data due to censor-

ing or lost to follow-up. Usually some subjects have not experienced the event at the end of follow-up, that is, their records of survival time are unavailable (right censored), so the actual survival time for them is unknown. We proposed a flexible approach that can incorporate survival data into GRF and any other ITE models, an approach based on weighted 'mean imputation'.

Given a subject, denote its actual survival time as T_i and its censor time as c_i. In reality we observe y_i which is $\min(T_i, c_i)$ due to censoring. Let $t_{(1)} < t_{(2)} < \cdots < t_{(j)}$ be the censored survival times in ascending order, and $\hat{K}$ be the Kaplan-Meier (KP) estimator function of survival. Given $T_i > c_i$ for subject i, its log of censored survival times can be estimated by

$$\log(y_i^*) = \sum_{t(j) > T_i^c} \log t(j) \frac{\Delta\hat{T}(t(j))}{\hat{T}(T_i^c)}, \tag{5.14}$$

where $\Delta\hat{T}(t(j))$ refers to to the jump size of $\hat{K}$ at $t_{(j)}$ [11]. Under the assumption of the log-normal distribution of survival time, the imputed survival times can be included in ITE estimation.

5.5 Experiment Results

5.5.1 Simulations Studies

In design of simulations to evaluate the performance of our ITE framework and to compare the power and type I error rate of our proposed statistical tests with that of `test_calibration` provided in the R package `grf`, we adopted similar strategies in the generation of synthetic data as introduced in [25]. Here we assume the log survival time

is normally distributed without the loss of generality. We considered the following six elements in simulations design:

1. *Sample size* The number of observations in the data is n, and p is the number of covariates.

2. *Distributions of covariates* Across all simulation scenarios, we drawn samples from standard normal distribution for features with odd column number, and for features with even column number we sampled from a Bernoulli distribution with $p = 1/2$. For simplicity, we denotes the distribution for the policy as D_x.

3. *Key functions* we denote propensity function for observations receiving a treatment $\pi(\cdot)$, average treatment effect $\mu(\cdot)$, and treatment effect $\tau(\cdot)$. The conditional mean effect for treatments and controls can be designed to be $\mu_1(\cdot) = \mu(\cdot) + \tau(\cdot)/2$ and $\mu_0(\cdot) = \mu(\cdot) - \tau(\cdot)/2$ respectively.

4. *Generation of survival time* Under the log-normal distribution of survival time, the survival time of observation can be generated by taking the natural exponentiation of the mean effect using $\exp(\cdot)$.

5. *Censor* We considered a censor rate of roughly 20% for all our scenarios. Given the uncensored simulated survival time, we found the cutoff r for the 80% quantile, and then generated censor time $T(\cdot)$ using the exponential distribution with rate parameter $1/r$. If the simulated $\log Y_i > T(\mathbf{x}_i)$, then $T(\mathbf{x}_i)$ should be used as outcome; otherwise $\log Y_i$ was used.

6. *Noise levels* The noise level $\sigma^2_{\log(Y)}$ was introduced in the generation of log survival time $\log Y_i$, which sampled from a normal distribution with mean $\mu(\cdot)+(w-1/2)\tau(\cdot)$ and variance $\sigma^2_{\log(Y)}$, where $w \sim \text{Bernoulli}(\pi(\cdot))$.

Given the above predefined components, our data generation for observations i is modeled as

$$\begin{aligned}
\mathbf{x}_i &\sim D_x, \\
W_i &\sim \text{Bernoulli}\big(\pi(\mathbf{x}_i)\big), \\
\log Y_i &\sim \text{Normal}\big(\mu(\mathbf{x}_i)+(W_i-1/2)\tau(\mathbf{x}_i), \sigma^2_{\log(Y)}\big), \\
T_i &\sim \text{Exp}(1/r), \\
c_i &= \begin{cases} 1, & \text{if } \log Y_i \le T_i \\ 0, & \text{otherwise} \end{cases}, \\
Y_i &= \exp(\min(\log Y_i, T_i)),
\end{aligned}$$

where r is a cutoff corresponding to a specific quantile of $\log Y_i$, and $\pi(\mathbf{x}_i)$ is defined as

$$\pi(\mathbf{x}_i) = \frac{\exp(\mu(\mathbf{x}_i)-\tau(\mathbf{x})/2)}{1+\exp(\mu(\mathbf{x}_i)-\tau(\mathbf{x})/2)} \tag{5.15}$$

for observational studies; for randomized studies $\pi(\mathbf{x}_i) = 1/2$ for all $\mathbf{x}_i$, which is the same as defined in [25]. In practice, the value for 0.8 quantile was used. c_i is an indicator for censor. If there is an event ($\log Y_i \le T_i$) for subject i, then c_i is 1; otherwise 0.

We used functions with minor changes from [25] for propensity probability of receiving a treatment $\pi(\cdot)$, average treatment effect $\mu(\cdot)$,

and treatment effect $\tau(\cdot)$. Within the simulation experiments, both randomized and observational studies are included, and 8 different functions of mean and treatment effects are made here to represent both univariate and multivariate, both additive and interactive, and both linear and piecewise constant relationships. They are defined as follows:

$$
\begin{aligned}
&f_1(x) = 0, f_2(x) = 5\mathbb{I}(x_1 > 1) - 5 * pnorm(-1), f_3(x) = 5x_1,\\
&f_4(x) = x_2x_4x_6 + 2x_2x_4(1-x_6) + 3x_2(1-x_4)x_6\\
&\qquad + 4x_2(1-x_4)(1-x_6) + 5(1-x_2)x_4x_6 + 6(1-x_2)x_4(1-x_6)\\
&\qquad + 7(1-x_2)(1-x_4)x_6 + 8(1-x_2)(1-x_4)(1-x_6) - 4.5,\\
&f_5(x) = x_1 + x_3 + x_5 + x_7 + x_8 + x_9,\\
&f_6(x) = 4\mathbb{I}(x_1 > 1)\mathbb{I}(x_3 > 0) + 4\mathbb{I}(x_5 > 1)\mathbb{I}(x_7 > 0) + 2x_8x_9 - 4 * pnorm(-1),\\
&f_7(x) = \frac{1}{\sqrt{2}}(x_1^3 + x_2 + x_3^3 + x_4 + x_5^3 + x_6 + x_7^3 + x_8 + x_9^3 - 7)\\
&f_8(x) = \frac{1}{2}(f_4(x) + f_5(x)),
\end{aligned}
$$

where $pnorm(x)$ is a function that calculates the cumulative distribution probability $F(x) = \Pr(X \leq x)$. Here X is with standard normal distribution.

Table 5.1: Specifications for simulation scenarios

	Scenarios							
	1,9	2,10	3,11	4,12	5,13	6,14	7,15	8,16
n	300	300	200	600	400	300	450	700
p	400	400	300	300	200	200	100	100
$\mu(x)$	$f_8(x)$	$f_5(x)$	$f_4(x)$	$f_7(x)$	$f_3(x)$	$f_1(x)$	$f_2(x)$	$f_6(x)$
$\tau(x)$	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$	$f_6(x)$	$f_3(x)$	$f_8(x)$
$\sigma^2_{\log Y}$	1	1/4	1	1/4	1	1	4	4

The 8 functions listed are centered and scaled to have mean close to 0 and roughly the same variance. Table 5.1 gives the specifications for simulation scenarios, including sample size n, number of features p, functions for mean and treatment effect $\mu(\cdot)$ and $\tau(\cdot)$, and variance of noise $\sigma^2_{\log Y}$. Specifications for sample size have been adjusted to accommodate our simulated survival data, compared with those in study [25]. Scenarios of odd number are randomized experiments, with $\pi(\mathbf{x}_i) = 1/2$ for all $\mathbf{x}_i$, but scenarios of even number are observational studies, in which $\pi(\mathbf{x}_i)$ is defined by equation 5.15 for each subject $\mathbf{x}_i$. Note there is no heterogeneity in treatment effects in scenario 1 and 9.

To explore the reliability of our developed statistical tests and compare them with the `test_calibration` (TC) from GRF, we repeated application of our ITE framework with different random seed to the above simulation scenarios 500 times and examined the fitting of our approach with the statistical tests and test_calibration for each run. We calculated the proportion of repeats with p-values ≤ 0.05 for every statistical test. Because of no heterogeneity in scenarios 1 and 9, the proportion for statistical tests in scenarios 1 and 9 are type I error rate. The proportion for other simulation scenarios is the power of statistical methods. Simulation results are shown in table 5.2.

Statistical tests with $\widehat{\text{Var}}_\tau$ and τ_{improve} and TC maintain good validity. They all have very strong power in capturing the presence of heterogeneity, and notably our methods with $\widehat{\text{Var}}_\tau$ and τ_{improve} completely dominate TC provided in package GRF cross all simulation scenarios except 1 and 9. In scenarios 1 and 9 they all show relatively low type I error rate, with roughly 0.05 for statistical tests with $\widehat{\text{Var}}_\tau$ and τ_{improve}

Table 5.2: Comparison of power/type I error rate of different tests for the presence of heterogeneity

Scenarios	SHC-P	SHC-K	SHC-S	TC	$\widehat{\text{Var}}_\tau$	τ_{improve}
1*	0.002	0.002	0.002	0.002	0.058	0.054
2	0.266	0.268	0.26	0.896	0.976	0.98
3	1	1	1	1	1	1
4	0.656	0.652	0.652	0.782	0.924	0.924
5	0.548	0.548	0.546	0.752	0.926	0.934
6	0.888	0.88	0.882	0.944	0.992	0.992
7	0.862	0.844	0.85	0.886	0.956	0.974
8	0.752	0.732	0.73	0.42	0.592	0.664
9*	0.004	0.004	0.004	0.002	0.054	0.068
10	0.162	0.164	0.164	0.652	0.856	0.88
11	0.672	0.666	0.672	1	1	1
12	0.904	0.908	0.908	0.92	0.968	0.986
13	0.636	0.628	0.632	0.742	0.924	0.934
14	0.6	0.6	0.598	0.74	0.938	0.95
15	0.784	0.774	0.774	0.604	0.85	0.87
16	0.99	0.986	0.988	0.956	0.978	0.988

1. SHC-P stands for split-half correlation with pearson method, SHC-K for split-half correlation with kendall method, and SHC-S for split-half correlation with kendall method method for correlation testing.
2. TC represents `test_calibration` from the R package `GRF`.
3. $\widehat{\text{Var}}_\tau$ and τ_{improve} stand for two statistical tests for the presence of heterogeneity using them, proposed in section 5.4.1.
4. Scenario 1 and 9 are masked with star, since they are two scenarios without heterogeneity.

and 0.002 for TC. Type I error rates for our methods are also within an acceptable range, and TC has lower type I error rate than our methods. However, our methods with $\widehat{\text{Var}}_\tau$ and τ_{improve} are more favored in detecting the presence of heterogeneity, even though they shows slight inflated type I error rate.

Split-half correlation (SHC) approaches with three different correlation evaluation methods show good type I error rates in scenario 1 and 9. However, their powers varies greatly cross all other scenarios. By investigating their performance in scenario 6, 7, 12, and 16, we found that they also can maintain good performance if there is strong heterogeneity present. Surprisingly, they apparently outperform the other three methods in Scenario 8. In short, we can still consider SHC approaches as good supplements to the other three methods.

5.5.2 Applications on Real Data

5.6 Conclusion

Bibliography data is put in database.bib.

□ **End of chapter.**

Chapter 6

Conclusion

Appendix A

Proof of Propositions

□ **End of chapter.**

Appendix B

Publication List

□ End of chapter.

Bibliography

[1] C. Aggarwal, C. W. Davis, R. Mick, J. C. Thompson, S. Ahmed, S. Jeffries, S. Bagley, P. Gabriel, T. L. Evans, J. M. Bauml, et al. Influence of tp53 mutation on survival in patients with advanced egfr-mutant non–small-cell lung cancer. *JCO precision oncology*, 2018.

[2] S. F. Assmann, S. J. Pocock, L. E. Enos, and L. E. Kasten. Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet*, 355(9209):1064–1069, 2000.

[3] S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[4] S. Athey, J. Tibshirani, S. Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

[5] Y. Benjamini and R. Heller. Screening for partial conjunction hypotheses. *Biometrics*, 64(4):1215–1222, 2008.

[6] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[7] V. Chernozhukov, M. Demirer, E. Duflo, and I. Fernandez-Val. Generic machine learning inference on heterogenous treatment effects in randomized experiments. Technical report, National Bureau of Economic Research, 2018.

[8] H. A. Chipman, E. I. George, R. E. McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[9] D. I. Cook, V. J. Gebski, and A. C. Keech. Subgroup analysis in clinical trials. *Medical Journal of Australia*, 180(6):289, 2004.

[10] A. Dasgupta, S. Szymczak, J. H. Moore, J. E. Bailey-Wilson, and J. D. Malley. Risk estimation using probability machines. *BioData mining*, 7(1):2, 2014.

[11] S. Datta, J. Le-Rademacher, and S. Datta. Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and lasso. *Biometrics*, 63(1):259–271, 2007.

[12] R. Fisher, L. Pusztai, and C. Swanton. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3):479–485, 2013.

[13] J. C. Foster, J. M. Taylor, and S. J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

[14] M. Gail and R. Simon. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, pages 361–372, 1985.

[15] D. P. Green and H. L. Kern. Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly*, 76(3):491–511, 2012.

[16] J. Hill and Y.-S. Su. Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, pages 1386–1420, 2013.

[17] J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[18] K. Imai, M. Ratkovic, et al. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

[19] F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.

[20] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

[21] M. Lu, S. Sadiq, D. J. Feaster, and H. Ishwaran. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27(1):209–219, 2018.

[22] I. J. Neeland, P. Poirier, and J.-P. Després. Cardiovascular and metabolic heterogeneity of obesity: clinical challenges and implications for management. *Circulation*, 137(13):1391–1406, 2018.

[23] X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

[24] W. K. O'Neal and M. R. Knowles. Cystic fibrosis disease modifiers: complex genetics defines the phenotypic diversity in a monogenic disease. *Annual review of genomics and human genetics*, 19:201–222, 2018.

[25] S. Powers, J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie, and R. Tibshirani. Some methods for heterogeneous treatment effect

estimation in high-dimensions. *arXiv preprint arXiv:1707.00102*, 2017.

[26] D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[27] R. J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[28] X. Su, K. Meneses, P. McNees, and W. O. Johnson. Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(3):457–474, 2011.

[29] X. Su, C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158, 2009.

[30] L. Tian, A. A. Alizadeh, A. J. Gentles, and R. Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

[31] J. Tibshirani, S. Athey, S. Wager, R. Friedberg, L. Miner, M. Wright, M. J. Tibshirani, L. Rcpp, R. I. DiceKriging, and G. SystemRequirements. Package 'grf'. 2018.

[32] P. A. VanderLaan, D. Rangachari, S. M. Mockus, V. Spotlow, H. V. Reddi, J. Malcolm, M. S. Huberman, L. J. Joseph, S. S. Kobayashi, and D. B. Costa. Mutations in tp53, pik3ca, pten and other genes in egfr mutated lung cancers: Correlation with clinical outcomes. *Lung Cancer*, 106:17–21, 2017.

[33] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[34] L. Yang, Y. Zhao, Y. Wang, L. Liu, X. Zhang, B. Li, and R. Cui. The effects of psychological stress on depression. *Current neuropharmacology*, 13(4):494–504, 2015.