frontiers
in Big Data

Evaluating the Effectiveness of Personalized Medicine With Software

Adam Kapelner[1]*, Justin Bleich[2], Alina Levine[1], Zachary D. Cohen[3], Robert J. DeRubeis[3] and Richard Berk[2]

[1]Department of Mathematics, Queens College, CUNY, Queens, NY, United States, [2]Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA, United States, [3]Department of Psychology, University of Pennsylvania, Philadelphia, PA, United States

We present methodological advances in understanding the effectiveness of personalized medicine models and supply easy-to-use open-source software. Personalized medicine involves the systematic use of individual patient characteristics to determine which treatment option is most likely to result in a better average outcome for the patient. Why is personalized medicine not done more in practice? One of many reasons is because practitioners do not have any easy way to holistically evaluate whether their personalization procedure does better than the standard of care, termed *improvement*. Our software, "Personalized Treatment Evaluator" (the R package PTE), provides inference for improvement out-of-sample in many clinical scenarios. We also extend current methodology by allowing evaluation of improvement in the case where the endpoint is binary or survival. In the software, the practitioner inputs 1) data from a single-stage randomized trial with one continuous, incidence or survival endpoint and 2) an educated guess of a functional form of a model for the endpoint constructed from domain knowledge. The bootstrap is then employed on data unseen during model fitting to provide confidence intervals for the improvement for the average future patient (assuming future patients are similar to the patients in the trial). One may also test against a null scenario where the hypothesized personalization are not more useful than a standard of care. We demonstrate our method's promise on simulated data as well as on data from a randomized comparative trial investigating two treatments for depression.

Keywords: personalized medicine, inference, bootstrap, treatment regimes, randomized comparative trial, statistical software

1 INTRODUCTION

Personalized medicine, sometimes called "precision medicine" or "stratified medicine" (Smith, 2012), is a medical paradigm offering the possibility for improving the health of individuals by judiciously treating individuals based on his or her heterogeneous prognostic or genomic information (Zhao and Zeng, 2013). These approaches have been described under the umbrella of "P4 medicine," a systems approach that combines predictive, personalized, preventive and participatory features to improve patient outcomes (Weston and Hood, 2004; Hood and Friend, 2011). And the interest in such personalization is exploding.

Fundamentally, personalized medicine is a statistical problem. However, much recent statistical research has focused on how to best estimate *dynamic treatment regimes* or *adaptive interventions* (Collins et al., 2004; Chakraborty and Murphy, 2014). These are essentially strategies that vary

treatments administered over time as more is learned about how particular patients respond to one or more interventions. Elaborate models are often proposed that purport to estimate optimal dynamic treatment regimes from multi-stage experiments (Murphy, 2005b) as well as the more difficult situation of inference in observational studies.

Thus, the extant work, at least in the field of statistics, is highly theoretical. There is a dearth of software that can answer two fundamental questions practitioners will need answered before they can personalize future patients' treatments: 1) How much better is this personalization model expected to perform when compared to my previous "naive" strategy for allocating treatments? 2) How confident can I be in this estimate? Can I reject a null hypothesis that it will perform no better than the standard of care? Chakraborty and Moodie (2013), page 168 believe that "more targeted research is warranted" on these important questions; and the goal of our paper is to provide a framework and usable software that fills in part of this gap.

Personalized medicine is a broad paradigm encompassing many real-world situations. The setting we focus on herein is a common one: using previous randomized comparative/controlled trial (RCT) data to be able to make better decisions for future patients. We consider RCTs with two treatment options (two-arm), with one endpoint measure (also called the "outcome" or "response" which can be continuous, binary or survival) and where the researchers also collected a variety of patient characteristics to be used for personalization. The practitioner also has an idea of a model of the response (usually a simple first-order interaction model). Although this setting is simplistic, our software can then answer the two critical questions listed above.

Our advances are modest but important for practitioners. 1) Practitioners now have easy-to-use software that automates the testing of their personalization models. 2) We introduce a more intuitive metric that gauges how well the personalization is performing: "improvement" vs. a baseline strategy. 3) Our estimates and hypothesis tests of improvement are all cross-validated, making the estimates honest even when the data at hand was overfit and thereby generalizable to future patients. This external validity is only possible if future patients can be thought to come from the same population as the clinical trial, a generalization that is debated but beyond the scope of our work. 4) We have extended this well-established methodology to the setting of binary and survival endpoints, the most common endpoints in clinical trials.

The paper proceeds as follows. In **Section 2**, we review the modern personalized medicine literature and locate our method and its limitations within. **Section 3** describes our methods and its limitations in depth, by describing the conceptual framework emphasizing our methodological advances. We then carefully specify the data and model inputs, define the improvement metric, and illustrate a strategy for providing practitioners with estimates and inference. **Section 4** applies our methods to 1) a simple simulated dataset in which the response model is known, 2) a more complicated dataset characterized by an unknown response model and 3) a real data set from a published clinical trial investigating two treatments for major depressive disorder. **Section 5** demonstrates the software for all three types of endpoints: continuous, binary and survival. **Section 6** concludes and offers future directions of which there are many.

2 BACKGROUND

Consider an individual seeking one of two treatments, neither of which is known to be superior for all individuals. "What treatment, by whom, is most effective for this individual with that specific problem, and under which set of circumstances?" (Paul, 1967).[1] Sometimes practitioners will select a treatment based informally on personal experience. Other times, practitioners may choose the treatment that their clinic or peers recommend. If the practitioner happens to be current on the research literature and there happens to be a published RCT whose results have clear clinical implications, the study's superior treatment (on average) may be chosen.

Each of these approaches can sometimes lead to improved outcomes, but each also can be badly flawed. For example, in a variety of clinical settings, "craft lore" has been demonstrated to perform poorly, especially when compared to very simple statistical models (Dawes, 1979). It follows that each of these "business-as-usual" *treatment allocation procedures* can in principle be improved if there are patient characteristics available which are related to how well an intervention performs.

Personalized medicine via the use of patient characteristics is by no means a novel idea. As noted as early as 1865, "the response of the average patient to therapy is not necessarily the response of the patient being treated" (translated by Bernard, 1957). There is now a substantial literature addressing numerous aspects of personalized medicine. The field is quite fragmented: there is literature on treatment-covariate interactions, locating subgroups of patients as well as literature on personalized treatment effects estimation. However, a focus on inference is rare in the literature and available software for inference is negligible.

Byar (1985) provides an early review of work involving treatment-covariate interactions. Byar and Corle (1977) investigates tests for treatment-covariate interactions in survival models and discusses methods for treatment recommendations based on covariate patterns. Shuster and van Eys (1983) considers two treatments and proposes a linear model composed of a treatment effect, a prognostic factor, and their interaction. Using this model, the authors create confidence intervals to determine for which values of the prognostic factor one of two treatments is superior.

---

[1]Note that this problem is encountered in fields outside of just medicine. For example, finding the movie that will elicit the most enjoyment to the individual (Zhou et al., 2008) or assessing whether a certain unemployed individual can benefit from job training (LaLonde, 1986). Although the methods discussed herein can be applied more generally, we will employ examples and the vocabulary from the medical field.

Many researchers also became interested in discovering "qualitative interactions," which are interactions that create a subset of patients for which one treatment is superior and another subset for which the alternative treatment is superior. Gail and Simon (1985) develop a likelihood ratio test for qualitative interactions which was further extended by Pan and Wolfe (1997) and Silvapulle (2001). For more information and an alternative approach, see Foster (2013).

Much of the early work in detecting these interactions required a prior specification of subgroups. This can present significant difficulties in the presence of high dimensionality or complicated associations. More recent approaches such as Su et al. (2009) and Dusseldorp and Van Mechelen (2014) favor recursive partitioning trees that discover important nonlinearities and interactions. Dusseldorp et al. (2016) introduce an R package called QUINT that outputs binary trees that separate participants into subgroups. Shen and Cai (2016) propose a kernel machine score test to identify interactions and the test has more power than the classic Wald test when the predictor effects are non-linear and when there is a large number of predictors. Berger et al. (2014) discuss a method for creating prior subgroup probabilities and provides a Bayesian method for uncovering interactions and identifying subgroups. Berk et al. (2020) uses trees to both locate subgroups and to provides valid inference on the local heterogeneous treatment effects. They do so by exhaustively enumerating every tree and every node, running every possible test and then providing valid post-selection inference (Berk et al., 2013b).

In our method, we make use of RCT data. Thus, it is important to remember that "clinical trials are typically not powered to examine subgroup effects or interaction effects, which are closely related to personalization ... even if an optimal personalized medicine rule can provide substantial gains, it may be difficult to estimate this rule with few subjects" (Rubin and van der Laan, 2012). This is why a major bulk of the literature does not focus on finding covariate-treatment interactions or locating subgroups of individuals, but instead on the entire model itself (called the "regime") and that entire model is then used to sort individuals. Holistic statements can then be made on the basis of this entire sorting procedure. We turn to some of this literature now.

Zhang et al. (2012a) consider the context of treatment regime estimation in the presence of model misspecification when there is a single-point treatment decision. By applying a doubly robust augmented inverse probability weighted estimator that under the right circumstances can adjust for confounding and by considering a restricted set of policies, their approach can help protect against misspecification of either the propensity score model or the regression model for patient outcome. Brinkley et al. (2010) develop a regression-based framework of a dichotomous response for personalized treatment regimes within the rubric of "attributable risk." They propose developing optimal treatment regimes that minimize the probability of a poor outcome, and then consider the positive consequences, or "attributable benefit," of their regime. They also develop asymptotically valid inference for a parameter similar to improvement with business-as-usual as the random (see our **Section 3.3**), an idea we extend. Within the literature, their work is the closest conceptually to ours. Gunter

et al. (2011b) develop a stepwise approach to variable selection and in Gunter et al. (2011a) compare it to stepwise regression. Rather than using a traditional sum-of-squares metric, the authors' method compares the estimated mean response, or "value," of the optimal policy for the models considered, a concept we make use of in **Section 3**. Imai and Ratkovic (2013) use a modified Support Vector Machine with LASSO constraints to select the variables useful in an optimal regime when the response is binary. van der Laan and Luedtke (2015) uses a loss-based super-learning approach with cross-validation.

Also important within the area of treatment regime estimation, but not explored in this paper, is the estimation of dynamic treatment regimes (DTRs). DTRs constitute a set of decision rules, estimated from many experimental and longitudinal intervals. Each regime is intended to produce the highest mean response over that time interval. Naturally, the focus is on optimal DTRs—the decision rules which provide the highest mean response. Murphy (2003) and Robins (2004) develop two influential approaches based on regret functions and nested mean models respectively. Moodie et al. (2007) discuss the relationship between the two while Moodie and Richardson (2009) and Chakraborty et al. (2010) present approaches for mitigating biases (the latter also fixes biases in model parameter estimation stemming from their non-regularity in SMART trials). Robins et al. (2008) focus on using observational data and optimizing the time for administering the stages—the "when to start"—within the DTR. Orellana et al. (2010) develop a different approach for estimating optimal DTRs based on marginal structural mean models. Henderson et al. (2010) develop optimal DTR estimation using regret functions and also focus on diagnostics and model checking. Barrett et al. (2014) develop a doubly robust extension of this approach for use in observational data. Laber et al. (2014) demonstrate the application of set-valued DTRs that allow balancing of multiple possible outcomes, such as relieving symptoms or minimizing patient side effects. Their approach produces a subset of recommended treatments rather than a single treatment. Also, McKeague and Qian (2014) estimate treatment regimes from functional predictors in RCTs to incorporate biosignatures such as brain scans or mass spectrometry.

Many of the procedures developed for estimating DTRs have roots in reinforcement learning. Two widely used methods are Q-learning Murphy (2005a) and A-learning (see Schulte et al., 2014 for an overview of these concepts). One well-noted difficulty with Q-learning and A-learning are their susceptibility to model misspecification. Consequently, researchers have begun to focus on "robust" methods for DTR estimation. Zhang et al. (2013) extend the doubly robust augmented inverse probability weighted method (Zhang et al., 2012a) by considering multiple binary treatment stages.

Many of the methods mentioned above can be extended to censored survival data. Zhao et al. (2015) describe a computationally efficient method for estimating a treatment regimes that maximizes mean survival time by extending the weighted learning inverse probability method. This method is doubly robust; it is protected from model misspecification if either the censoring model or the survival model is correct.

Additionally, methods for DTR estimation can be extended. Goldberg and Kosorok (2012) extend Q-learning with inverse-probability-of-censoring weighting to find the optimal treatment plan for individual patients, and the method allows for flexibility in the number of treatment stages.

It has been tempting, when creating these treatment regime models, to directly employ them to predict the differential response of individuals among different treatments. This is called in the literature "heterogeneous treatment effects models" or "individualized treatment rules" and there is quite a lot of interest in it.

Surprisingly, methods designed for accurate estimation of an overall conditional mean of the response may not perform well when the goal is to estimate these individualized treatment rules. Qian and Murphy (2011) propose a two-step approach to estimating "individualized treatment rules" based on single-stage randomized trials using ℓ_1-penalized regression while Lu et al. (2013) use quadratic loss which facilitates variable selection. Rolling and Yang (2014) develop a new form of cross-validation which chooses between different heterogeneous treatment models.

One current area of research in heterogeneous effect estimation is the development of algorithms that can be used to create finer and more accurate partitions. Kallus (2017) presents three methods for the case of observational data: greedily partitioning data to find optimal trees, bootstrap aggregating to create a "personalization forest" a la Random Forests, and using the tree method coupled with mixed integer programming to find the optimal tree. Lamont et al. (2018) build on the prior methods of parametric multiple imputation and recursive partitioning to estimate heterogeneous treatment effects and compare the performance of both methods. This estimation can be extended to censored data. Henderson et al. (2020) discuss the implementation of Bayesian Additive Regression Trees for estimating heterogeneous effects which can be used for continuous, binary and censored data. Ma et al. (2019) proposes a Bayesian predictive method that integrates multiple sources of biomarkers.

One major drawback of many of the approaches in the literature reviewed is their significant difficulty evaluating estimator performance. Put another way, given the complexity of the estimation procedures, statistical inference is very challenging. Many of the approaches require that the proposed model be correct. There are numerous applications in the biomedical sciences for which this assumption is neither credible nor testable in practice. For example, Evans and Relling (2004) consider pharmacogenomics, and argue that as our understanding of the genetic influences on individual variation in drug response and side-effects improves, there will be increased opportunity to incorporate genetic moderators to enhance personalized treatment. But we will ever truly understand the endpoint model well enough to properly specify it? Further, other biomarkers (e.g. neuroimaging) of treatment response have begun to emerge, and the integration of these diverse moderators will require flexible approaches that are robust to model misspecification (McGrath et al., 2013). How will the models of today incorporate important relationships that can be anticipated but have yet to be identified? Further, many

proposed methods employ non-parametric models that use the data to decide which internal parameters to fit and then in turn estimates these internal parameters. Thus a form of model selection that introduces difficult inferential complications (see Berk et al., 2013b).

At the very least, therefore, there should be an alternative inferential framework for evaluating treatment regimes that do not require correct model specification (and thereby obviating the need for model checking and diagnostics) nor knowledge of unmeasured characteristics (see discussion in Henderson et al., 2010) accompanied by easy-to-use software. This is the modest goal herein.

3 METHODS

This work seeks to be didactic and thus carefully explains the extant methodology and framework (**Section 3.1**), data inputs (**Section 3.2**), estimation (**Section 3.4**) and a procedure for inference (**Section 3.5**). For those familiar with this literature, these sections can be skipped. Our methodological contributions then are to 1) employ out-of-sample validation to (**Section 3.4**) specifically to 2) *improvement*, the metric defined as the difference in how patients allocated via personalized model fare in their outcome and a patients allocated via business-as-usual model fare in their outcome (**Section 3.3**) and 3) to extend this validation methodology to binary and survival endpoints (**Sections 3.4.2 and 3.4.3**). **Table 1** serves as a guide to the main notation used in this work, organized by section.

3.1 Conceptual Framework
We imagine a set of random variables having a joint probability distribution that can be also be viewed as a population from which data could be randomly and independently realized. The population can also be imagined as all potential observations that could be realized from the joint probability distribution. Either conception is consistent with our setup.

A researcher chooses one of the random variables to be the response Y which could be continuous, binary or survival (with a corresponding variable that records its censoring, explained later). We assume without loss of generality that a higher-valued outcome is better for all individuals. Then, one or more of the other random variables are covariates $X \in \mathcal{X}$. At the moment, we do not distinguish between observed and unobserved covariates but we will later. There is then a conditional distribution $\mathbb{P}(Y|X)$ whose conditional expectation function $\mathbb{E}[Y|X]$ constitutes the population response surface. No functional forms are imposed on the conditional expectation function and thus it is allowed to be nonlinear with interactions among the covariates which is the case in artificial intelligence procedures (machine learning).

All potential observations are hypothetical study subjects. Each can be exposed to a random treatment denoted $A \in \{0, 1\}$ where zero codes for the first experimental condition also equivalently referred to as T_1 (which may be considered the "control" or "comparison" condition) and one codes for another experimental condition equivalently referred to as T_2. We make the standard assumption of no interference

TABLE 1 | A compendium of the main notation in our methodology by section.

Notation	Description
Framework (**Section 3.1**)	
Y	The random variable (r.v.) for the outcomes for the subjects
$X, \mathcal{X}$	The r. v. for the observed measurements for the subjects, its support
A	The r. v. for the treatment
T_1, T_2	The two treatments, shorthand for their codes, zero and one
d or $d[f]$	The decision function; this function maps observed measurements to treatment
V or $V[d]$	The value of the decision function, the average outcome over all patients if this decision is used to allocate treatment
d^*	The unknown optimal decision function i.e. the one with highest V
d_0	A naive, baseline, business-as-usual or null decision function
μ_{I_0}	The unknown improvement of d over d_0; it is the difference of values
The RCT data (**Section 3.2.1**)	
n	The number of subjects in the randomized comparative trial (RCT)
p	The number of measurements assessed on each subject
x_i	The vector of p measurements for the ith subject
$x_{i,j}$	The jth measurement for the ith subject
$\boldsymbol{X}$	The $n \times p$ matrix of all measurements for all subjects
$\boldsymbol{A}$	The vector of treatments for all the n subjects
$\boldsymbol{y}$	The vector of outcomes for all the n subjects
The response model (**Section 3.2.2**)	
f	The function that relates the p measurements and A to the response
U_i	The r. v. for the unknown covariates for subject i
$\xi(x_i, A_i, U_i)$	The function that computes misspecification in the response
ε_i	The r. v. for irreducible noise for the ith subject
β_j	The linear coefficient for the jth measurement when $A = 0$
γ_j	The additional linear coefficient for the jth measurement when $A = 1$
Out of sample estimation and validation (**Section 3.4**)	
$\boldsymbol{X}_{train}, \boldsymbol{y}_{train}$	The subset of the data used to create the fit of f
$\boldsymbol{X}_{test}, \boldsymbol{y}_{test}$	The subset of the data used to validate the fit of f
$\widehat{f}, \widehat{d}, \widehat{V}, \widehat{I}_0$	The finite-sample estimates of f, d, V, μ_{I_0}
$\overline{y}_{set}$	The arithmetic average of $\{y_i : i \in set\}$
$\widehat{\beta}_j, \widehat{\gamma}_j$	The finite-sample estimates of β_j, γ_j
Inference (**Section 3.5**)	
B	The number of bootstrap samples
$\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}}$	A sample of the rows of $\boldsymbol{X}, \boldsymbol{y}$ with replacement
$\tilde{I}_{0,b}$	The bth estimate of μ_{I_0} in the bootstrap
α	The size of the hypothesis test
Personalization of future subjects' treatments (**Section 3.6**)	
x_*	A future subject (not part of the RCT)

between study subjects, which means that the outcome for any given subject is unaffected by the interventions to which other subjects are randomly assigned (Cox, 1958) and outcomes under either condition can vary over subjects (Rosenbaum, 2002, Section 2.5.1). In short, we employ the conventional Neyman-Rubin approach (Rubin, 1974) but treat all the data as randomly realized (Berk et al., 2013a).

A standard estimation target in RCTs is the population average treatment effect (PATE), defined here as $\mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0]$, the difference between the population expectations. That is, the PATE is defined as the difference in mean outcome were all subjects exposed to T_2 or alternatively were all exposed to T_1. In a randomized controlled trial, the PATE is synonymous with the overall efficacy of the treatment of interest and it is almost invariably the goal of the trial (Zhao and Zeng, 2013).

For personalization, we want to make use of any association between Y and X. For the hypothetical study subjects, there is a conditional population response surface $\mathbb{E}[Y|X, A = 1]$ and another conditional population response surface $\mathbb{E}[Y|X, A = 0]$, a key objective being to exploit the difference

in these response surfaces for better treatment allocation. The typical approach is to create a deterministic *individualized treatment decision rule d* that takes an individual's covariates and maps them to a treatment. We seek $d : \mathcal{X} \rightarrow \{0, 1\}$ based on knowledge of the differing conditional population response surfaces. The rule is sometimes called an *allocation procedure* because it determines which treatment to allocate based on measurements made on the individual. To compare different allocation procedures, our metric is the expectation of the outcome Y using the allocation procedure d averaged over all subjects $\mathcal{X}$. Following the notation of Qian and Murphy (2011), we denote this expectation as the value of the decision rule

$$V[d] := \mathbb{E}^d_{X,A}[Y] \triangleq \int_X \left(\sum_{a \in \{0,1\}} \left(\int_\mathbb{R} y f_{Y|X,A}(y,x,a) dy \right) \mathbb{1}_{a=d(x)} \right) f_X(x) dx.$$

Although the integral expression appears complicated, when unpacked it is merely an expectation of the response averaged

over $\mathcal{X}$, the space of all patients characteristics. When averaging over $\mathcal{X}$, different treatments will be recommended based on the rule, i.e. $a = d(x)$, and that in turn will modify the density of the response, $f_{Y|X}$. Put another way, $V[d]$ is the mean patient outcome when personalizing each patient's treatment.

We have considered all covariates to be random variables because we envision *future* patients for whom an appropriate treatment is required. Ideally, their covariate values are realized from the same joint distribution as the covariate values for the study subjects, an assumption that is debated and discussed in the concluding section.

In addition, we do not intend to rely on estimates of the two population response surfaces. As a practical matter, we will make do with a population response surface approximation for each. No assumptions are made about the nature of these approximations and in particular, how well or poorly either population approximation corresponds to the true conditional response surfaces.

Recall that much of the recent literature has been focused on finding the optimal rule, $d^* \triangleq \mathrm{argmax}_d\{V[d]\}$. Although this is an admirable ideal (as in Qian and Murphy 2011), our goals here are more modest. We envision an imperfect rule d far from d^*, and we wish to gauge its performance relative to the performance of another rule d_0, where the "naught" denotes a business-as-usual allocation procedure, sometimes called "standard of care". Thus, we define the population value improvement μ_{I_0} as the value of d minus the value of d_0,

$$\mu_{I_0} \triangleq V[d] - V[d_0] = \mathbb{E}_{X,A}^d[Y] - \mathbb{E}_{X,A}^{d_0}[Y], \tag{1}$$

which is sometimes called "benefit" in the literature. Since our convention is that higher response values are better, we seek large, positive improvements that translate to better average performance (as measured by the response). Note that this is a natural measure when Y is continuous. When Y is incidence or survival, we redefine μ_{I_0} (see **Sections 3.4.2** and **3.4.3**).

The metric μ_{I_0} is not standard in the literature but we strongly believe it to be the natural metric for personalization following Kallus (2017). There are many other such metrics beyond value and improvement. For example, Ma et al. (2019) uses three 1) the expected number of subjects misassigned to their optimal treatment, 2) expected gain or loss in treatment utility and 3) the expected proportion where the model correctly predicted the response which is useful only in the binary response case which we address later.

3.2 Our Framework's Required Inputs

Our method depends on two inputs 1) access to RCT data and 2) either a prespecified parametric model $f(x, A; \theta)$ for the population approximation of the true response surfaces or an explicit d function. If we prespecified f, we then use the RCT data to estimate parameters of the model $\widehat{\theta}$, and the estimates are embedded in the estimated model, $\widehat{f}$. This model estimate permits us, in turn, to construct an estimated decision rule $\widehat{d}$ and an estimate of the improved outcomes future subjects will experience (explained later in **Section 3.4**). We assume that the model f is specified before looking at the data. "Data snooping" (running

our method, checking the p-value, changing the model f and running again) fosters overfitting and can introduce serious estimation bias, invalidating our confidence intervals and statistical tests (Berk et al., 2013b).

3.2.1 The RCT Data
Our procedure strictly requires RCT data to ensure there is a causal effect of the heterogeneous parameters. Much of the research discussed in the background (**Section 2**) applies in the case of observational data. We realize this limits the scope of our proposal. The RCT data must come from an experiment undertaken to estimate the PATE for treatments T_1 and T_2 for a diagnosis of a disease of interest. T_1 and T_2 are the same treatments one would offer to future subjects with the same diagnosis.

There are n subjects each with p covariates which are denoted for the ith subject as $\boldsymbol{x_i} \triangleq [x_{i1}, x_{i2}, \ldots, x_{ip}]$. Because these covariates will be used to construct a decision rule applied with future patients in clinical settings, the $\boldsymbol{x_i}$'s in the RCT data must be the same covariates measured for new subjects. Thus, such characteristics such as the site of treatment (in a multi-center trial) or the identification of the medical practitioner who treated each subject or hindsight-only variables are not included.

We assume the outcome measure of interest y_i is assessed once per subject. Aggregating all covariate vectors, binary allocations and responses rowwise, we denote the full RCT data as the column-bound matrix $[\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{y}]$. In practice, missing data can be imputed (in both the RCT data and the future data), but herein we assume complete data.

We will be drawing inference to a patient population beyond those who participated in the experiment. Formally, new subjects must be sampled from that same population as were the subjects in the RCT. In the absence of explicit probability sampling, the case would need to be made that the model can generalize. This requires subject-matter expertise and knowledge of how the study subjects were recruited.

### 3.2.2 The Model for the Response Based on Observed Measurements
The decision rule d is a function of x through f and is defined as

$$d[f(x)] \triangleq \mathrm{argmax}_{A \in \{0,1\}} f(x, A) = \mathbb{1}_{f(x,1) - f(x,0)}. \tag{2}$$

As in Berk et al. (2014), we assume the model f provided by the practitioner to be an approximation using the available information,

$$Y_i = \underbrace{f(X_i, A_i) + \xi(X_i, U_i, A_i)}_{\mathbb{E}[Y_i|X_i, U_i, A_i]} + \varepsilon_i, \tag{3}$$

where f differs from the true response expectation by a term dependent on U, the unobserved information. The last term ε_i is the irreducible noise around the true conditional expectations and is taken to be independent and identically distributed, mean-centered and uncorrelated with the covariates. Even in the absence of ε_i, f will always differ from the true conditional
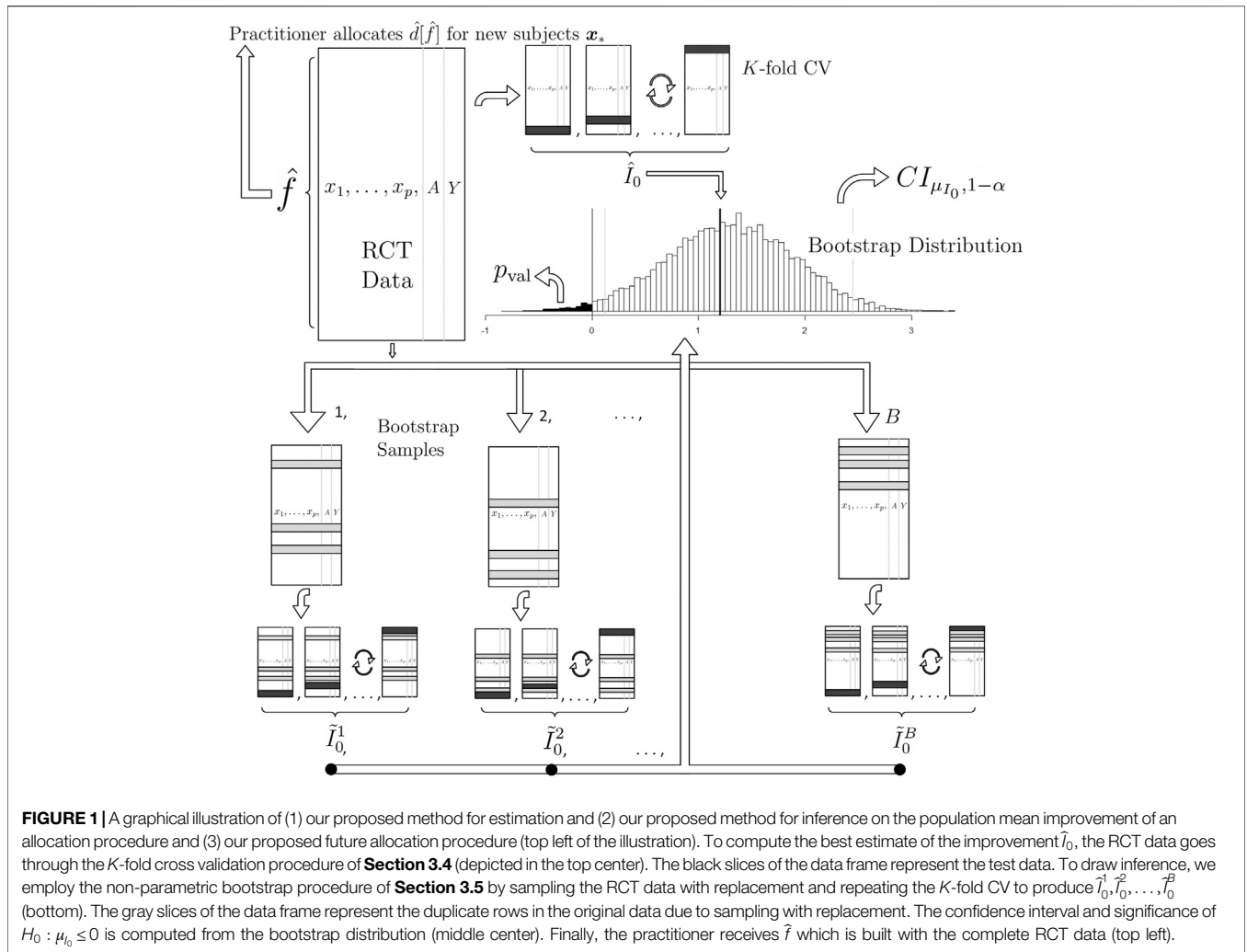
FIGURE 1 | A graphical illustration of (1) our proposed method for estimation and (2) our proposed method for inference on the population mean improvement of an allocation procedure and (3) our proposed future allocation procedure (top left of the illustration). To compute the best estimate of the improvement $\hat{I}_0$, the RCT data goes through the K-fold cross validation procedure of **Section 3.4** (depicted in the top center). The black slices of the data frame represent the test data. To draw inference, we employ the non-parametric bootstrap procedure of **Section 3.5** by sampling the RCT data with replacement and repeating the K-fold CV to produce $\tilde{I}_0^1, \tilde{I}_0^2, \ldots, \tilde{I}_0^B$ (bottom). The gray slices of the data frame represent the duplicate rows in the original data due to sampling with replacement. The confidence interval and significance of $H_0 : \mu_{I_0} \leq 0$ is computed from the bootstrap distribution (middle center). Finally, the practitioner receives $\hat{f}$ which is built with the complete RCT data (top left).

expectation function by $\xi_i (X_i, U_i, A_i)$, which represents model misspecification (Box and Draper, 1987, Chapter 13).

We wish only to determine whether an estimate of f is *useful* for improving treatment allocation for future patients (that are similar to the patients in the RCT) and do not expect to recover the optimal allocation rule d^* which requires the unseen U. Further, we do not concern ourselves with substantive interpretations associated with any of the p covariates, a goal of future research. Thus, our method is robust to model misspecification by construction.

What could f look like in practice? Assume a continuous response (binary and survival are discussed later) and consider the conventional linear regression model with first order interactions. Much of the literature we reviewed in **Section 2** favored this class of models. We specify a linear model containing a subset of the covariates used as main effects and a possibly differing subset of the covariates to be employed as first order interactions with the treatment indicator, $\{x_{1'}, \ldots, x_{p'}\} \subset \{x_1, \ldots, x_p\}$, selected using domain knowledge:

$$f(x_{i1}, A_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + A_i \left(\gamma_0 + \gamma_{1'} x_{1'} + \cdots + \gamma_{p'} x_{p'} \right). \quad (4)$$

These interactions induce heterogeneous effects between T_1 and T_2 for a subject x in a very interpretable way: $d[f(x)] = 1$ when $\gamma_0 + \gamma_{1'} x_{1'} + \cdots + \gamma_{p'} x_{p'} > 0$ and 0 otherwise. The γ's are the critical component of the model if there are systematic patient-specific differences between the interventions. Thereby, d varies over different points in $\mathcal{X}$ space. Note that rules derived from this type of conventional model also have the added bonus as being interpretable as a best linear approximation of the true relationship.

We stress that our models are *not* required to be of this form, but we introduce them here mostly for familiarity and pedagogical simplicity. There are times when this linear model will perform terribly even if $\{x_{1'}, \ldots, x_{p'}\}$ are the correct moderating variables. For a non-linear example, see Zhao and Zeng (2013), **Figure 1**, right. Although this model is the default implementation, the user can specify any model desired in the software. This will be discussed in **Section 5**.

We also stress that although the theory for estimating linear models' coefficients (as well as those for logistic regression and Weibull regression) is well-developed, we are not interested in inference for these coefficients in this work as our goal is only estimation and inference for overall usefulness of the personalization scheme, i.e. the unknown parameter μ_{I_0}. This will become clear in the next few sections.

3.3 Other Allocation Procedures

Although d_0 can be any allocation rule, for the purposes of the paper, we examine only two "business-as-usual" allocation procedures (others are discussed as extensions in **Section 6**). The first we call **random** denoting the allocation where the patient receives T_1 or T_2 with a fair coin flip, probability 50%. This serves as a baseline or "straw man" but nevertheless an important standard—the personalization model should be able to provide better patient outcomes than a completely random allocation.

The second business-as-usual procedure we call **best**. This procedure gives all patients the better of the two treatments as determined by the comparison of the sample average for all subjects who received T_1 denoted $\bar{y}_{T_1}$ and the sample average of all subjects who received T_2 denoted $\bar{y}_{T_2}$. This is used as the default in many frameworks for example Kang et al. (2014). We consider beating this procedure the gold standard in proof that the personalization truly works as practitioners most often employ the current best known treatment. However, some consider this comparison conservative (Brinkley et al., 2010, Section 7) and the next section will describe why it is statistically conservative as we lose sample size when demanding this comparison. Due to this conservativeness, barring conclusive evidence that either T_1 or T_2 is superior, the random procedure should be the standard of comparison. This case is not infrequent in RCTs which feature negative comparison results, the case in our clinical trial example of **Section 4.3**.

3.4 Estimating the Improvement Scores
3.4.1 For a Continuous Response
How well do unseen subjects with treatments allocated by d do on average compared to the same unseen subjects with treatments allocated by d_0? We start by computing the estimated improvement score, a sample statistic given by

$$\hat{I}_0 \triangleq \hat{V}[\hat{d}] - \hat{V}[\hat{d}_0], \tag{5}$$

where $\hat{d}$ is an estimate of the rule d derived from the population response surface approximation, $\hat{V}$ is an estimate of its corresponding value V and $\hat{I}_0$ is an estimate of the resulting population improvement μ_{I_0} (**Eq. 1**). The $\hat{d}_0$ notation indicates that sometimes the competitor d_0 may have to be estimated from the data as well. For example, the allocation procedure **best** must be calculated by using the sample average of the responses for both T_1 and T_2 in the data.

In order to properly estimate μ_{I_0}, we use cross-validation (Hastie et al., 2013, Chapter 7.10). We split the RCT data into two disjoint subsets: training data with n_{train} of the original n observations $[X_{\text{train}}, y_{\text{train}}]$ and testing data with the

TABLE 2 | The elements of y_{test} cross-tabulated by their administered treatment A_i and our model's estimate of the better treatment $\hat{d}(x_i)$.

	$\hat{d}(x_i) = 0$	$\hat{d}(x_i) = 1$
$A_i = 0$	P	Q
$A_i = 1$	R	S

remaining $n_{\text{test}} = n - n_{\text{train}}$ observations $[X_{\text{test}}, y_{\text{test}}]$. Then $\hat{f}_{\text{train}}$ can be fit using the training data to construct $\hat{d}$ via **Eq. 2**. Performance of $\hat{d}$ as calculated by **Equation 5**, is then evaluated on the test data. Hastie et al. (2013) explain that a single train-test split yields an estimate of the "performance" of the procedure on future individuals conditional on $[X_{\text{train}}, y_{\text{train}}]$, the "past". Thus, the $\hat{I}_0$ statistic defined in **Eq. 5** computed using $[X_{\text{test}}, y_{\text{test}}]$ can provide an honest assessment of improvement (i.e. immune to overfitting in $\hat{f}$) who are allocated using our proposed methodology compared to a baseline business-as-usual allocation strategy (Faraway, 2016). This can be thought of as employing a replicated trial, often required in drug development programs, which separates rule construction (in-sample) from rule validation (out-of-sample) as recommended by Rubin and van der Laan (2012). Note that this comes at a cost of more sample variability (as now our estimate will be based on the test subset with a sample size much smaller than n). Our framework and software is the first to provide user-friendly out-of-sample validation for the overall utility of personalized medicine models as a native feature.

Given the estimates $\hat{d}$ and $\hat{d}_0$, the question remains of how to explicitly compute $\hat{V}$ for subjects we have not yet seen in order to estimate $\hat{I}_0$. That is, we are trying to estimate the expectation of an allocation procedure over covariate space $\mathcal{X}$.

Recall that in the test data, our allocation prediction $\hat{d}(x_i)$ is the binary recommendation of T_1 or T_2 for each $x_{\text{test},i}$. If we recommended the treatment that the subject actually was allocated in the RCT, i.e. $\hat{d}(x+_i) = A_i$, we consider that subject to be "lucky". We define lucky in the sense that by the flip of the coin, the subject was randomly allocated to the treatment that our model-based allocation procedure estimates to be the better of the two treatments.

The average of the lucky subjects' responses should estimate the average of the response of new subjects who are allocated to their treatments based on our procedure d and this is the estimate of $\hat{V}[\hat{d}]$ we are seeking. Because the x's in the test data are assumed to be sampled randomly from population covariates, this sample average estimates the expectation over $\mathcal{X}$, i.e. $\mathbb{E}_{X,A}^d[Y]$ conditional on the training set. In order to make this concept more clear, it is convenient to consider **Table 2**, a 2×2 matrix which houses the sorted entries of the out-of-sample y_{test} based on the predictions, the $\hat{d}(x_i)$'s.

The diagonal entries of sets P and S contain the "lucky" subjects. The notation $\bar{y}_{\cdot}$ indicates the sample average among the elements of y_{test} specified in the subscript located in the cells of the table.

How do we compute $\hat{V}[\hat{d}_0]$, the business-as-usual procedure? For **random**, we simply average all of the y_{test} responses; for **best**, we average the y_{test} responses for the treatment group that has a larger sample average. Thus, the sample statistics of **Eq. 5** can be written as

$$\hat{I}_{\text{random}} \triangleq \bar{y}_{P \cup S} - \bar{y}_{\text{test}}, \tag{6}$$

$$\hat{I}_{\text{best}} \triangleq \bar{y}_{P \cup S} - \begin{cases} \bar{y}_{P \cup Q}, & \text{when} \quad \bar{y}_{P \cup Q} \geq \bar{y}_{R \cup S}, \\ \bar{y}_{R \cup S}, & \text{when} \quad \bar{y}_{P \cup Q} < \bar{y}_{R \cup S}. \end{cases} \tag{7}$$

Note that the plug-in estimate of value $\hat{V}[\hat{d}] = \bar{y}_{P \cup S}$ is traditional in the personalized medicine literature. For example, in Kallus (2017), Corollary 3 it is written as

$$\hat{V}[\hat{d}] := \frac{\sum_{i=1}^{n} Y_i \mathbb{1}_{\hat{d}(x_i)=A_i}}{\sum_{i=1}^{n} \mathbb{1}_{\hat{d}(x_i)=A_i}}. \tag{8}$$

There is one more conceptual point. Recall that the value estimates $\hat{V}[\cdot]$ are conditional on the training set. This means they do not estimate the unconditional $\mathbb{E}_{X,A}^{d}[Y]$. To address this, Hastie et al. (2013), Chapter 7 recommend that the same procedure be performed across many different mutually exclusive and collectively exhaustive splits of the full data. This procedure of building many models is called "K-fold cross-validation" (CV) and its purpose is to integrate out the effect of a single training set to result in the unconditional estimate of generalization. This is "an alternative approach ... [that] for simplicity ... [was not] consider [ed] ... further" in the previous investigation of Chakraborty et al. (2014), page 5.

In practice, how large should the training and test splits be? Depending on the size of the test set relative to the training set, CV can trade bias for variance when estimating an out-of-sample metric. Small training sets and large test sets give more biased estimates since the training set is built with less data than the n observations given. However, large test sets have lower variance estimates since they are composed of many examples. There is no consensus in the literature about the optimal training-test split size (Hastie et al., 2013, page 242) but 10-fold CV is a common choice employed in many statistical applications and provides for a relatively fast algorithm. In the limit, n models can be created by leaving each observation out, as done in DeRubeis et al. (2014). In our software, we default to 10-fold cross validation but allow for user customization.

This estimation procedure outlined above is graphically illustrated in the top of **Figure 1**. We now extend this methodology to binary and survival endpoints in the next two sections.

3.4.2 For a Binary Response

In the binary case, we let V be the expected probability of the positive outcome under d just like in Kang et al. (2014). We consider three improvement metrics 1) the probability difference, 2) the risk ratio and 3) the odds ratio:

(a) $\mu_{I_0} \triangleq V[d] - V[d_0],$

(b) $\mu_{I_0} \triangleq \dfrac{V[d]}{V[d_0]},$

(c) $\mu_{I_0} \triangleq \dfrac{V[d]/(1 - V[d])}{V[d_0]/(1 - V[d_0])}.$

and the estimate of all three (the probability difference, the risk ratio and the odds ratio) is found by placing hats on each term in the definitions above (all V's, d's and d_0's).

Following the example in the previous section we employ the analogous model, a logistic linear model with first order treatment interactions where the model f now denotes the probability of the positive outcome $y = 1$,

$$f(x_{i1}, A_i) = \text{logit}\big(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$
$$+ A_i(\gamma_0 + \gamma_{1'} x_{1'} + \ldots + \gamma_{p'} x_{p'})\big). \tag{9}$$

This model, fit via maximum likelihood numerically (Agresti, 2018), is the default in our software implementation. Here, higher probabilities of success imply higher logit values so that algebraically we have the same form of the decision rule estimate, $\hat{d}[\hat{f}(x)] = 1$ when $\hat{\gamma}_0 + \hat{\gamma}_{1'} x_{1'} + \cdots + \hat{\gamma}_{p'} x_{p'} > 0$.

If the risk ratio or odds ratio improvement metrics are desired, **Eqs. 6, 7** are modified accordingly but otherwise estimation is then carried out the same as in the previous section.

3.4.3 For a Survival Response

Survival responses differ in two substantive ways from continuous responses: 1) they are always non-negative and 2) some values are "censored" which means it appropriates the value of the last known measurement but it is certain that the true value is greater. The responses y are coupled with this censoring information c, a binary vector of length n where the convention is to let $c_i = 0$ to indicate that y_i is censored and thus set equal to its last known value.

To obtain $\hat{d}$, we require a survival model. For example purposes here we will assume the exponential regression model (the exponentiation enforces the positivity of the response values) with the usual first order treatment interactions,

$$f(x_{i1}, A_i) = \exp\big(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$
$$+ A_i(\gamma_0 + \gamma_{1'} x_{1'} + \cdots + \gamma_{p'} x_{p'})\big). \tag{10}$$

Under the exponential model, the convention is that the noise term ε is multiplicative instead of additive, $Y_i = f(x_{i1}, A_i)\varepsilon_i$. Note that at this step, a fully parametric model is needed; the non-parametric Kaplan-Meier or the semi-parametric Cox proportion hazard model are insufficient as we need a means of explicitly estimating $\mathbb{E}[Y \mid X, A]$ for all values of X and both values of A.

Moreso than for continuous and incidence endpoints, parameter estimation is dependent on the choice of error distribution. Following Hosmer and Lemeshow (1999), a flexible model is to let $\ln(\varepsilon_1), \ldots, \ln(\varepsilon_n) \overset{\text{iid}}{\sim} \text{Gumbel}(0, \sigma^2)$, implying the popular Weibull model for survival (and the default in our software). As was the case previously, the user is free to choose whatever model they wish. The β_j's, γ_j's and the nuisance scale parameter σ^2 are fit using maximum likelihood taking care to ensure the correct contributions of censored and uncensored values. Similar to the case of logistic regression, the

likelihood function does not have a closed form solution and must be approximated numerically.

Some algebra demonstrates that the estimated decision rule is the same as those above, i.e. $\widehat{d}[\widehat{f}(x)] = 1$ when $\widehat{\gamma}_0 + \widehat{\gamma}_{1'} x_{1'} + \cdots + \widehat{\gamma}_{p'} x_{p'} > 0$. In other words, the subject is given the treatment that yields the longest expected survival.

Subjects are then sorted in cells like **Table 2** but care is taken to keep the corresponding c_i values together with their paired y_i values following Yakovlev et al. (1994). At this point, we need to specify analogous computations to **Eqs. 6**, **7** that are sensitive to the fact that many y_i values are censored (The sample averages $\bar{y}$ obviously cannot be employed here because it ignores this censoring).

Of course we can reemploy a new Weibull model and define improvement as we did earlier as the difference in expectations (**Eq. 1**). However, there are no more covariates needed at this step as all subjects have been sorted based on $\widehat{d}(x)$. Thus, there is no reason to require a parametric model that may be arbitrarily wrong.

For our default implementation, we have chosen to employ the difference of the Kaplan-Meier median survival statistics here because we intuitively feel that a non-parametric estimate makes the most sense. Once again, the user is free to employ whatever they feel is most appropriate in their context. Given this default, please note that the improvement measure of **Eq. 1** is no longer defined as the difference in survival expectations, but now the difference in survival medians. This makes our framework slightly different in the case of survival endpoints.

3.5 Inference for the Population Improvement Parameter

Regardless of the type of endpoint, the $\widehat{I}_0$ estimates are drawn from an elaborate estimator whose sampling distribution is not available in closed form. We can employ the nonparametric bootstrap to obtain an asymptotic estimate of its sampling variability, which can be used to construct confidence intervals and testing procedures (Efron and Tibshirani, 1994).

In the context of our proposed methodology, the bootstrap procedure works as follows for the target of inference μ_{I_0}. We take a sample with replacement from the RCT data of size n denoted with tildes: $[\tilde{X}, \tilde{y}]$. Using the 10-fold CV procedure described at the end of **Section 3.4**, we create an estimate $\tilde{I}_0$. We repeat the resampling of the RCT data and the recomputation of $\tilde{I}_0$ B times where B is selected for resolution of the confidence interval and significance level of the test. In practice we found $B = 3000$ to be sufficient, so we leave this as the default in our software implementation. Because the n's of usual RCTs are small, and the bootstrap is embarrassingly parallelizable, this is not an undue computational burden.

In this application, the bootstrap approximates the sampling of many RCT datasets. Each $\tilde{I}$ that is computed corresponds to one out-of-sample improvement estimate for a particular RCT dataset drawn from the population of RCT datasets. We stress again that the frequentist confidence intervals and tests that we develop for the improvement measure do *not* constitute inference for a new individual's improvement, it is inference

for the average improvement for future subjects vs. **random** allocation, μ_{I_0}.

To create a $1 - \alpha$ level confidence interval, first sort the $\{\tilde{I}_{0,1}, \ldots, \tilde{I}_{0,B}\}$ by value, and then report the values corresponding to the empirical $\alpha/2$ and $1 - \alpha/2$ percentiles. This is called the "percentile method." "Although this direct equation of quantiles of the bootstrap sampling distribution with confidence limits may seem initially appealing, its "rationale is somewhat obscure" (Rice, 1994, page 272). There are other ways to generate asymptotically valid confidence intervals using bootstrap samples but there is debate about which has the best finite sample properties. We have also implemented the "basic method" (Davison and Hinkley, 1997, page 194) and the bias-corrected "BC_a method" of Efron (1987) that DiCiccio and Efron (1996) claim performs an order of magnitude better in accuracy than the percentile method. Implementing other confidence interval methods for the bootstrap may be useful future work.

If a higher response is better for the subject, we set $H_0 : \mu_{I_0} \le 0$ and $H_a : \mu_{I_0} > 0$. Thus, we wish to reject the null hypothesis that our allocation procedure is at most as useful as a naive business-as-usual procedure. To obtain an asymptotic p value based on the percentile method, we tally the number of bootstrap sample $\tilde{I}$ estimates below 0 and divide by B. This bootstrap procedure is graphically illustrated in the bottom half of **Figure 1** and the bootstrap confidence interval and p value computation is illustrated in the center. Note that for incidence outcomes where the improvement is defined as the risk ratio or odds ratio, we use $H_0 : \mu_{I_0} \le 1$ and $H_a : \mu_{I_0} > 1$ and tally the number of $\tilde{I}$ estimates below 1.

We would like to stress once again that we are not testing for qualitative interactions—the ability of a covariate to "flip" the optimal treatment for subjects. Tests for such interactions would be hypothesis tests on the γ parameters of **Eqs. 4**, **9**, **10**, which assume model structures that are not even required for our procedure. Qualitative interactions are controversial due to model dependence and entire tests have been developed to investigate their significance. In the beginning of **Section 2** we commented that most RCTs are not even powered to investigate these interactions. "Even if an optimal personalized medicine rule [based on such interactions] can provide substantial gains it may be difficult to estimate this rule with few subjects" (Rubin and van der Laan, 2012). The bootstrap test (and our approach at large) looks at the holistic picture of the personalization scheme without focus on individual covariate-treatment interaction effects to determine if the personalization scheme in totality is useful, conceptually akin to the omnibus F-test in an OLS regression.

3.5.1 Concerns With Using the Bootstrap for This Inference

There is some concern in the personalized medicine literature about the use of the bootstrap to provide inference. First, the estimator for V is a non-smooth functional of the data which may result in an inconsistent bootstrap estimator (Shao, 1994). The non-smoothness is due to the indicator function in **Eq. 8** being non-differentiable, similar to the example found in (Horowitz,

2001, Chapter 52, Section 4.3.1). However, "the value of a fixed [response model] (i.e., one that is not data-driven) does not suffer from these issues and has been addressed by numerous authors" (Chakraborty and Murphy, 2014). Since our $\widehat{V}$ is constructed out-of-sample, it is merely a difference of sample averages of the hold-out response values that are considered pre-sorted according to a fixed rule.[2] This setup does not come without a substantial cost. Estimation of the improvement score out-of-sample means the effective sample size of our estimate is small and our power commensurately suffers. One can also implement the double bootstrap (see e.g. the comparisons in Chakraborty et al., 2010) herein and that is forthcoming in our software (see **Section 5**).

There is an additional concern. Some bootstrap samples produce null sets for the "lucky subjects" (i.e. $P \cup S = \varnothing$ of **Table 2** or equivalently, all values of the indicator in **Eq. 8** are zero). These are safe to ignore as we are only interested in the distribution of estimates conditional on feasibility of estimation. Empirically, we have noticed that as long as $n > 20$, there are less than 1% of bootstrap samples that exhibit this behavior. Either way, we print out this percentage when using the PTE package and large percentages warn the user that inference is suspect.

3.6 Future Subjects

The implementation of this procedure for future patients is straightforward. Using the RCT data, estimate f to arrive at $\widehat{f}$. When a new individual, whose covariates are denoted $\boldsymbol{x}_\star$, enters a clinic, our estimated decision rule is calculated by predicting the response under both treatments, then allocating the treatment which corresponds to the better outcome, i.e. $\widehat{d}(\boldsymbol{x}_\star)$. This final step is graphically illustrated in the top left of **Figure 1**.

It is important to note that $\widehat{d}(\boldsymbol{x}_\star)$ is built with RCT data where treatment was allocated randomly and without regard to the subject covariates. In the example of the first order linear model with treatment interactions, the γ parameters have a causal interpretation—conditional causation based on the values of the moderating covariates. Thus $\widehat{d}(\boldsymbol{x}_\star)$ reflects a treatment allocation that causes the response to be higher (or lower). We reiterate that this would not be possible with observational data which would suffer from elaborate confounding relationships between the treatment and subject covariates (see discussion in **Sections 2** and **6.1**).

4 DATA EXAMPLES

We present two simulations in **Sections 4.1** and **4.2** that serve only as illustrations that our methodology both works as purported but degrades in the case of pertinent information that goes missing. We then demonstrate a real clinical setting in **Section 4.3**.

4.1 Simulation With Correct Regression Model

Consider a simulated RCT dataset with one covariate x where the true response function is known:

$$Y = \beta_0 + \beta_1 X + A(\gamma_0 + \gamma_1 X) + \varepsilon, \tag{11}$$

where ε is mean-centered. We employ $f(x, A)$ as the true response function, $\mathbb{E}[Y|X, A]$. Thus, $d = d^*$, the "optimal" rule in the sense that a practitioner can make optimal allocation decisions (modulo noise) using $d(x) = \mathbb{1}_{\gamma_0 + \gamma_1 x > 0}$. Consider d_0 to be the **random** allocation procedure (see **Section 3.3**). Note that within the improvement score definition (**Eq. 1**), the notation $\mathbb{E}_X^d[Y]$ is an expectation over the noise $\mathcal{E}$ and the joint distribution of X, A. After taking the expectation over noise, the improvement under the model of **Eq. 11** becomes

$$
\begin{aligned}
\mu_{I_0} &= \mathbb{E}_X\left[\beta_0 + \beta_1 X + \mathbb{1}_{\gamma_0 + \gamma_1 X > 0}(\gamma_0 + \gamma_1 X)\right] - \mathbb{E}_X\left[\beta_0 + \beta_1 X + 0.5(\gamma_0 + \gamma_1 X)\right] \\
&= \mathbb{E}_X\left[\left(\mathbb{1}_{\gamma_0 + \gamma_1 x > 0} - 0.5\right)(\gamma_0 + \gamma_1 X)\right] \\
&= \gamma_0\left(\mathbb{P}(\gamma_0 + \gamma_1 X > 0) - 0.5\right) + \gamma_1\left(\mathbb{E}_X\left[X\mathbb{1}_{\gamma_0 + \gamma_1 X > 0}\right] - 0.5\mathbb{E}_X[X]\right).
\end{aligned}
$$

We further assume $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and we arrive at

$$
\mu_{I_0} = (\gamma_0 + \gamma_1 \mu_X)\left(0.5 - \Phi\left(-\frac{\gamma_0}{\gamma_1}\right)\right) + \gamma_1 \frac{\sigma_X}{\sqrt{2\pi}} \exp
$$
$$
\left(-\frac{1}{2\sigma_X^2}\left(-\frac{\gamma_0}{\gamma_1} - \mu_X\right)^2\right).
$$

We simulate under a simple scenario to clearly highlight features of our methodology. If $\mu_X = 0, \sigma_X^2 = 1$ and $\gamma_0 = 0$, neither treatment T_1 or T_2 is on average better. However, if $x > 0$, then treatment T_2 is better in expectation by $\gamma_1 \times x$ and analogously if $x < 0$, T_1 is better by $-\gamma_1 \times x$. We then set $\gamma_1 = \sqrt{2\pi}$ to arrive at the round number $\mu_{I_0} = 1$. We set $\beta_0 = 1$ and $\beta_1 = -1$ and let $\mathcal{E}_i \overset{iid}{\sim} \mathcal{N}(0, 1)$. We let the treatment allocation vector A be a random block permutation of size n, balanced between T_1 and T_2. Since there is no PATE, the random and best d_0 procedures (see **Section 3.3**) are the same in value. We then vary $n \in \{100, 200, 500, 1000\}$ to assess convergence for both d_0 procedures and display the results in **Figure 2**.

Convergence to $\mu_{I_0} = 1$ is observed clearly for both procedures but convergence for d_0 best is slower than d_0 rand. This is due to the $\widehat{V}$ being computed with fewer samples: $\overline{y}_{\text{test}}$, which uses all of the available data, vs. $\overline{y}_{P \cup Q}$ or $\overline{y}_{R \cup S}$, which uses only half the available data on average (see **Eqs. 6, 7**). Also note that upon visual inspection, our bootstrap distributions seem to be normal. Our intuition is that non-normality in this distribution when using the software package warns the user that the inference is suspect.

In this section we assumed knowledge of f and thereby had access to an optimal rule. In the next section we explore convergence when we do not know f but pick an approximate model yielding a non-optimal rule that would still provide clinical utility.

---

[2]Note also that we do not have the additional non-smoothness created by Q-learning during the maximization step (Chakraborty et al., 2010, Section 2.4).
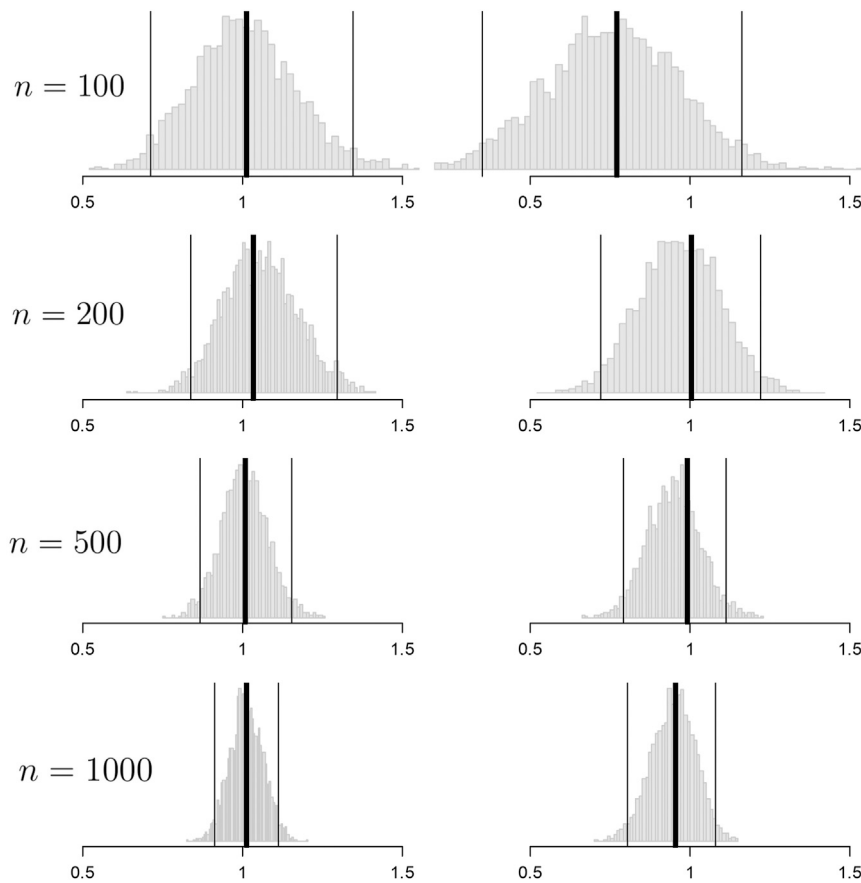
FIGURE 2 | Histograms of the bootstrap samples of the out-of-sample improvement measures for d_0 **random** (left column) and d_0 **best** (right column) for the response model of **Eq. 11** for different values of n. $\hat{I}_0$ is illustrated with a thick black line. The $\text{CI}_{\mu_{I_0}, 95\%}$ computed by the percentile method is illustrated by thin black lines.

4.2 Simulation With an Approximate Regression Model

Consider RCT data with a continuous endpoint where the true response model is

$$Y = \beta_0 + \beta_1 X + \beta_2 U + A\left(\gamma_0 + \gamma_1 X^3 + \gamma_2 U\right) + \varepsilon, \qquad (12)$$

where X denotes a covariate recorded in the RCT and U denotes a covariate that is not included in the RCT dataset. The optimal allocation rule d^* is one when $\gamma_0 + \gamma_1 X^3 + \gamma_2 U > 0$ and 0 otherwise. The practitioner, however, does not have access to the information contained in U, the unobserved covariate, and has no way to ascertain the exact relationship between X and the treatment. Consider the default model that is an approximation of the true population response surface,

$$f(X, A) = \beta_0 + \beta_1 X + A\left(\gamma_0 + \gamma_1 X\right), \qquad (13)$$

which is different from the true response model due to (a) the misspecification of X (linear instead of cubic) and (b) the absence of covariate U (see **Eq. 3**). This is a realistic scenario; even with infinite data, d^* cannot be located because of both ignorance of the true model form and unmeasured subject characteristics.

To simulate, we set the X's, U's and $\mathcal{E}$'s to be standard normal variables and then set $\beta_0 = 1, \beta_1 = -1, \beta_2 = 0.5, \gamma_0 = 0, \gamma_1 = 1$ and $\gamma_2 = -3$. The X_i's and the U_i's are deliberately made independent of one another so that the observed covariates cannot compensate for the unobserved covariates, making the comparison between the improvement under d^* and d more stark. To find the improvement when the true model's d^* is used to allocate, we simulate under **Eq. 12** and obtain $\mu_{I_0}^* \approx 1.65$ and analogously, to find the improvement under the approximation model's d, we simulate under **Eq. 13** and obtain $\mu_{I_0} \approx 0.79$. Further simulation shows that not observing U is responsible for 85% of this observed drop in improvement performance and employing the linear X in place of the non-linear X^3 is responsible for the remaining 15%. Since $\gamma_0 = 0$ and the seen and unseen covariates are mean-centered, there is no PATE and thus these simulated improvements apply to both the cases where d_0 is **random** and d_0 is **best**.

Figure 3 demonstrates results for $n = \{100, 200, 500, 1000\}$ analogous to **Figure 2**. We observe that the bootstrap confidence intervals contain μ_{I_0} but not $\mu_{I_0}^*$. This is expected; we are not allocating using an estimate of d^*, only an estimate of d.

Convergence toward $\mu_{I_0} = 0.79$ is observed clearly for both procedures and once again the convergence is slower for the best
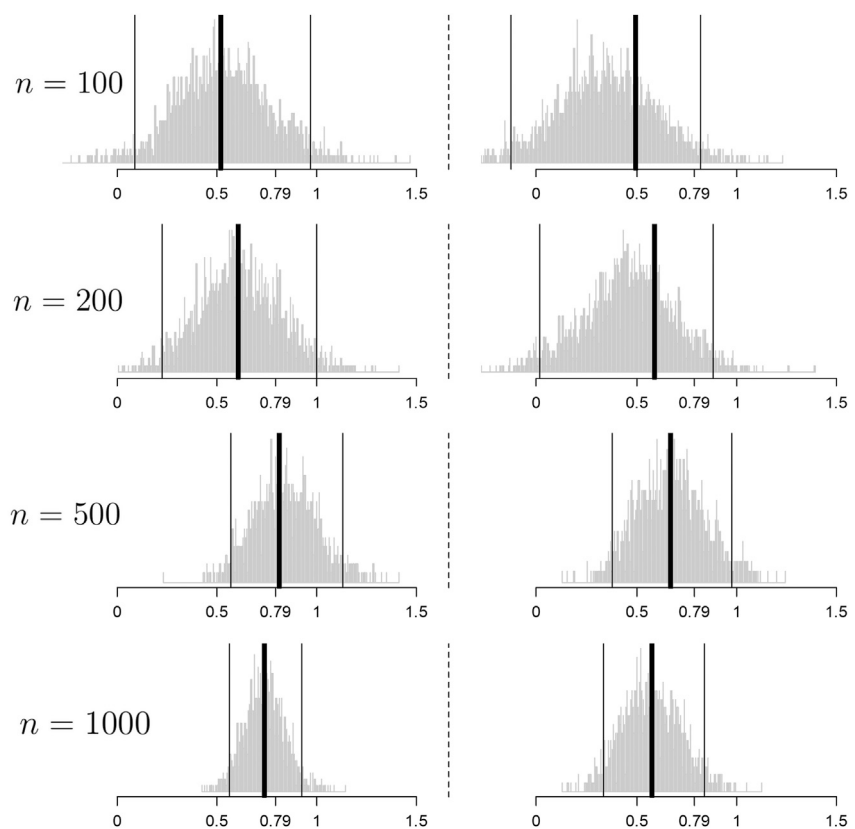
FIGURE 3 | Histograms of the bootstrap samples of the cross-validated improvement measures for d_0 **random** (left column) and d_0 **best** (right column) for the response model of **Eq. 12** for different values of n. $\hat{I}_0$ is illustrated with a thick black line. The $CI_{\mu_{I_0}, .95\%}$ computed via the percentile method is illustrated by thin black lines. The true population improvement $\mu_{I_0}^\star$ given the optimal rule $d^\star$ is illustrated with a dotted black line.

procedure for the same reasons outlined in **Section 4.1**. Note that the coverage illustrated here is far from $\mu_{I_0}^\star$, the improvement using the optimal allocation rule. Kallus (2017) presents a coefficient of personalization metric similar to R^2 where a value of 100% represents perfect personalization and 0% represents standard of care. Here, we would fall far short of the 100%.

The point of this section is to illustrate what happens in the real world: the response model is unknown and important measurements are missing and thus any personalized medicine model falls far short of optimal. However, the effort can still yield an improvement that can be clinically significant and useful in practice.

There are many cases where our procedure will not find signal yielding improvement in patient outcomes either because it does not exist or we are underpowered to detect it. For example 1) in cases where there are many variables that are important and a small sample size. Clinical trials are not usually powered to find even single interaction effects, let alone many. The small sample size diminishes power to find the effect, similar to any statistical test. 2) If the true heterogeneity in the functional response cannot be approximated by a linear function. For instance, parabolic or sine functions cannot be represented whatsoever by best fit lines.

In the next section, we use our procedure in RCT data from a real clinical trial where both these limitations apply. The strategy is to approximate the response function using a reasonable model f built from domain knowledge and the variables at hand and hope to find demonstrate a positive, clinically meaningful μ_{I_0} knowing full well it will be much smaller than $\mu_{I_0}^\star$.

4.3 Clinical Trial Demonstration

We consider an illustrative example in psychiatry, a field where personalization, called "*precision psychiatry*, promises to be even more transformative than in other fields of medicine" (Fernandes et al., 2017). However, evaluation of predictive models for precision psychiatry has been exceedingly rare, with a recent systematic review identifying 584 studies in which prediction models had been developed, with only 10.4% and 4.6% having conducted proper internal and external validation, respectively (Salazar de Pablo et al., 2020).

We will demonstrate our method (that provides this proper validation) on the randomized comparative trial data of DeRubeis et al. (2005). In this depression study, there were two treatments with very different purported mechanisms of action: cognitive behavioral therapy (T_1) and paroxetine, an antidepressant medication (T_2). After omitting patients who dropped out

TABLE 3 | Baseline characteristics of the subjects in the clinical trial example for the moderating variables employed in our personalization model. These statistics differ slightly from those found in the table of DeRubeis et al. (2005, page 412) as here they are tabulated for subjects only after dropout ($n = 154$).

Variable	Sample Average or Proportion
Age	40.3 ± 11.3
Chronicity	55.1%
Life stressors	6.6 ± 4.8
Personality disorder	48.1%
Unemployed	14.9%
Married	37.6%

there were $n = 154$ subjects with 28 baseline characteristics measured. Although this dataset did not have explicit patient-level identifying information, using these 28 characteristics could potentially identify some of the patients. Note that this study was funded and begun before clinical trials registration and thus it does not have a clinical trial registration number.

The primary outcome measure y is the continuous Hamilton Rating Scale for Depression (HRSD), a composite score of depression symptoms where lower means less depressed, assessed by a clinician after 16 weeks of treatment. A simple t test revealed that there was no statistically significant HRSD difference between the cognitive behavioral therapy and paroxetine arms, a well-supported finding in the depression literature. Despite the seeming lack of a population mean difference among the two treatments, practitioner intuition and a host of studies suggest that the covariates collected can be used to build a principled personalized model with a significant negative μ_{I_0}. The lack of a difference also suggests that the **random** d_0 is an appropriate baseline comparison. If there were to be a clinically and statistically significant average difference in treatment outcomes between T_1 and T_2, then the **best** d_0 would be appropriate. In this latter case, we have found in our experience (in analyses of other datasets) that proving a personalization improvement is elusive as it is difficult to beat **best**. Even though the **best** d_0 is inappropriate in our context, we still provide its results in this section for illustrative purposes.

We now must specify a model, f. For the purpose of illustration, we employ a linear model with first-order interactions with the treatment (as in **Eq. 4**). Which of the 28 variables should be included in the model? Clinical experience and theory should suggest both prognostic (main effect) and moderator (treatment interaction) variables (Cohen and DeRubeis, 2018). We should not use variables selected using methods performed on this RCT data such as the variables found in DeRubeis et al. (2014), **Table 3**. Such a procedure would constitute data snooping and it will invalidate the inference provided by our method. The degree of invalidation is not currently known and is much needed to be researched.

Of the characteristics measured in this RCT data, previous researchers have found significant treatment moderation in age
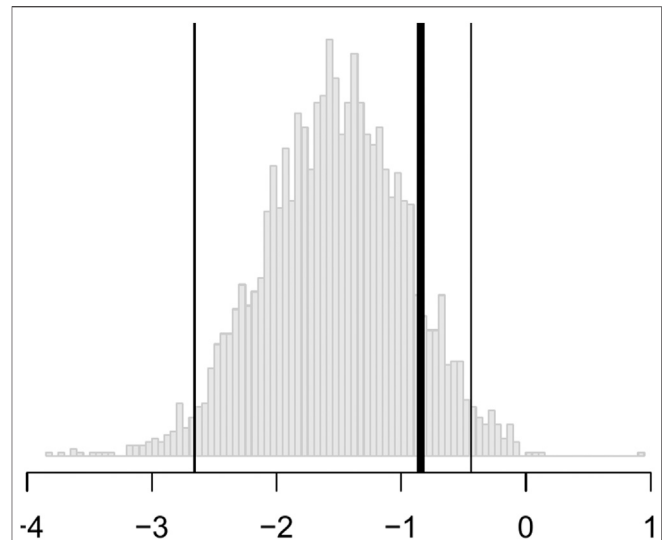
FIGURE 4 | Histograms of the bootstrap samples of $\bar{I}_{\mathrm{Rand}}$ i.e. for the random d_0 business-as-usual allocation procedure. The thick black line is the best estimate of $\widehat{I}_0$, the thin black lines are the confidence interval computed via the percentile method. More negative values are "better" as improvement is defined as lowering the HSRD composite score corresponding to a patient being less depressed.

and chronicity (Cuijpers et al., 2012), early life trauma (Nemeroff et al., 2003) (which we approximate using a life stressor metric), presence of personality disorder (Bagby et al., 2008), employment status and marital status (Fournier et al., 2009) but almost remarkably baseline severity of the depression does not moderate (Weitz et al., 2015) and baseline severity is frequently the most important covariate in response models (e.g. Kapelner and Krieger, 2020, **Figure 1B**). We include these $p = 6$ variables as moderators and as main effects[3] and statistics of their baseline characteristics are found in **Table 3**.

The output from 3,000 bootstrap samples are shown in **Figure 4**. From these results, we anticipate that a new subject allocated using our personalization model will be less depressed on average by 0.84 HRSD units with a 95% confidence interval of [0.441, 2.657] compared to that same subject being allocated randomly to cognitive behavioral therapy or paroxetine. We can easily reject the null hypothesis that personalized allocation over **random** is no better for the new subject (p value = 0.001).

In short, the results are statistically significant, but the estimated improvement may not be clinically significant. According to the criterion set out by the National Institute for Health and Care Excellence, three points on the HRSD is considered clinically important. Nevertheless, this personalization scheme can be implemented in practice with new patients for a modest improvement in patient outcome at little cost.

---

[3] Although this is standard linear modeling practice, it is not absolutely essential in our methodology, where our goal is neither inference for the variables nor prediction of the endpoint.

5 THE PTE PACKAGE

5.1 Estimation and Inference for Continuous Outcomes

The package comes with two example datasets. The first is the continuous data example. Below we load the library and the data whose required form is 1) a vector of length n for the responses, y and 2) a matrix X of dimension $n \times (p + 1)$ where one column is named "treatment" and the other p are appropriate names of the covariates.

```
R> library(PTE); library(dplyr)
R> data(continuous_example)
R> X = continuous_example$X
R> y = continuous_example$y
R> continuous_example$X %>% sample_n(5)
# A tibble: 5 x 6
  treatment     x1     x2     x3     x4         x5
      <dbl> <fctr> <fctr> <fctr> <fctr>      <dbl>
1         1     NO    OFF    YES MEDIUM  1.3009448
2         0    YES    OFF    YES MEDIUM -0.5483983
3         0     NO    OFF    YES    LOW  0.3762733
4         1     NO    OFF    YES MEDIUM -1.1648459
5         1    YES     ON    YES   HIGH -0.8566221
> round(head(continuous_example$y), 3)
[1] -0.746 -1.359  0.020  0.632 -0.823 -2.508
```

The endpoint y is continuous and the RCT data has a binary treatment vector appropriately named (this is required) and five covariates, four of which are factors and one is continuous. We can run the estimation for the improvement score detailed in **Section 3.4.1** and the inference of **Section 3.5** by running the following code:

```
R> options(mc.cores = 4)
R> pte_results = PTE_bootstrap_inference(X, y,
     B = 1000, run_bca_bootstrap = TRUE)
R> plot(pte_results)
```

Here, 1,000 bootstrap samples were run on four cores in parallel to minimize runtime. The f model defaults to a linear model where all variables included are interacted with the treatment and fit with least squares. Below are the results.

```
R> pte_results
  I_random observed_est = 0.077,  pctile p-val = 0.021,
    95% CI's: basic = [-0.237, 0.143], pctile = [0.011, 0.391],
      BCa = [-0.085, 0.145],
  I_best observed_est = 0.065,  pctile p-val = 0.089,
    95% CI's: basic = [-0.198, 0.2], pctile = [-0.071, 0.328],
      BCa = [-0.146, 0.198]
```

Note how the three bootstrap methods are different from another. The percentile method barely includes the actual observed statistic for the random comparison (see discussion in **Section 3.5**). The software also plots the $\tilde{I}$'s in a histogram (unshown).

To demonstrate the flexibility of the software, consider the case where the user wishes to use x_1, x_2, x_3, x_4 as main effects and x_5 as the sole treatment moderator. And further, the user wishes to estimate the model parameters using the ridge penalty instead of OLS. Note that this is an elaborate model that would be difficult to justify in practice and it is only shown here as an illustration of the customizability of

the software. Below is the code used to test this approach to personalization.

```
R> library(glmnet)
R> pte_results = PTE_bootstrap_inference(X, y, B = 1000,
  personalized_model_build_function = function(Xytrain){
    Xytrain_mm = model.matrix(~ . - y + x5 * treatment, Xytrain)
    cv.glmnet(Xytrain_mm, Xytrain[, ncol(Xytrain)], alpha = 0)
  },
  predict_function = function(mod, Xyleftout){
    Xyleftout$censored = NULL
    Xyleftout_mm = model.matrix(~ . + x5 * treatment, Xyleftout)
    predict(mod, Xyleftout_mm)
  })
```

Here, the user passes in a custom function that builds the ridge model to the argument `personalized_model_build_function`. The specification for ridge employed here uses the package `glmnet` (Friedman et al., 2010) that picks the optimal ridge penalty hyperparameter automatically. Unfortunately, there is added complexity: the `glmnet` package does not accept formula objects and thus model matrices are generated both upon model construction and during prediction. This is the reason why a custom function is also passed in via the argument `predict_function` which wraps the default `glmnet` predict function by passing in the model matrix.

5.2 Estimation and Inference for Binary Outcomes

In order to demonstrate our software for the incidence outcome, we use the previous data but threshold its response arbitrarily at its 75th percentile to create a mock binary response (for illustration purposes only).

```
R> y = ifelse(y > quantile(y, 0.75), 1, 0)
```

We then fit a linear logistic model using all variables as fixed effects and interaction effects with the treatment. As discussed in **Section 3.4.2**, there are three improvement metrics for incidence outcomes. The default is the odds ratio. The following code fits the model and performs the inference.

```
R> pte_results = PTE_bootstrap_inference(X, y, B = 1000,
    regression_type = "incidence", run_bca_bootstrap = TRUE)
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Note that the response type incidence has to be explicitly made known otherwise the default would be to assume the endpoint is continuous and perform regression. Below are the results.

```
R> pte_results
  I_random observed_est = 1.155,  pctile p-val = 0.104,
    95% CI's: basic = [0.317, 1.475], pctile = [0.836, 1.994],
      BCa = [0.497, 1.503],
  I_best observed_est = 1.04,  pctile p-val = 0.323,
    95% CI's: basic = [0.329, 1.431], pctile = [0.649, 1.751],
      BCa = [0.52, 1.528]
```

The p value is automatically calculated for $H_0 : \mu_{I_0} < 1$ (i.e. the odds of improvement is better in d_0 than d). Other tests can be specified by changing the `H_0_mu_equals` argument. Here, the test failed to reject H_0. Information is lost when a continuous metric

is coerced to be binary. If the user wished to define improvement via the risk ratio (or straight probability difference), an argument would be added to the above, `incidence_metric = "risk_ratio"` (or `"probability_difference"`).

5.3 Estimation and Inference for Survival Outcomes

Our package also comes with a mock RCT dataset with a survival outcome. In addition to the required input data y, X described in **Section 5.1**, we now also require a binary vector c also of length n where a value $c_i = 1$ denotes that the ith subject's y_i is a censored value. Below, we load the data.

```
R> data(survival_example)
R> X = survival_example$X
R> y = survival_example$y
R> censored = survival_example$censored
```

There are four covariates, one factor and three continuous. We can run the estimation for the improvement score detailed in **Section 3.4.3** and inference for the true improvement by running the following code.

```
R> pte_results = PTE_bootstrap_inference(X, y, censored = censored,
  B = 1000, run_bca_bootstrap = TRUE, regression_type = "survival")
```

The syntax is the same as the above two examples except here we pass in the binary c vector separately and declare that the endpoint type is survival. Again by default all covariates are included as main effects and interactions with the treatment in a linear Weibull model.

In the default implementation for the survival outcome, improvement is defined as median survival difference of personalization vs. standard of care. The median difference can be changed via the user passing in a new function with the `difference_function` argument. The median difference results are below.

```
R> pte_results
  I_random observed_est = 0.148,  p val = 0.017,
    95% CI's: basic = [0.011, 0.296], pctile = [0, 0.285],
      BCa = [-0.001, 0.274],
  I_best observed_est = -0.041,  p val = 0.669,
    95% CI's: basic = [-0.134, 0.082], pctile = [-0.164, 0.052],
      BCa = [-0.195, 0.013]
```

It seems that the personalized medicine model increases median survival by 0.148 vs. d_0 being the **random** allocation of the two treatments. If survival was measured in years (the typical unit), this would be about 2 months. However, it cannot beat the d_0 being the **best** of the two treatments. Remember, this is a much more difficult improvement metric to estimate as we are really comparing two cells in **Table 2** to another two cells, one of which is shared. Thus the sample size is low and power suffers. This difficulty is further compounded in the survival case because censored observations add little information.

6 DISCUSSION

We have provided a methodology to test the effectiveness of personalized medicine models. Our approach combines RCT data with a statistical model f of the response for estimating improved outcomes under different treatment allocation protocols. Using the non-parametric bootstrap and cross-validation, we are able to provide confidence bounds for the improvement and hypothesis tests for whether the personalization performs better compared to a business-as-usual procedure. We demonstrate the method's performance on simulated data and on data from a clinical trial on depression. We also present our statistical methods in an open source software package in R named PTE which is available on CRAN. Our package can be used to evaluate personalization models generally e.g. in heart disease, cancer, etc. and even outside of medicine e.g. in Economics and Sociology.

6.1 Limitations and Future Directions

Our method and corresponding software have been developed for a particular kind of RCT design. The RCT must have two arms and one endpoint (continuous, incidence or survival). An extension to more than two treatment arms is trivial as **Eq. 2** is already defined generally. Implementing extensions to longitudinal or panel data are simple within the scope described herein. And extending the methodology to count endpoints would also be simple.

Although we agree that a "once and for all" treatment strategy [may be] suboptimal due to its inflexibility" (Zhao et al., 2015), this one-stage treatment situation is still common in the literature and the real world and this is the setting we chose to research. We consider an extended implementation for dynamic treatment regimes on multi-stage experiments fruitful future work. Consider being provided with RCT data from sequential multiple assignment randomized trials ("SMARTs," Murphy, 2005b) and an a priori response model f. The estimate of $\widehat{V}[\widehat{d}]$ (**Eq. 5**) can be updated for a SMART with k stages (Chakraborty and Murphy, 2014) where our **Table 2** is a summary for only a single stage. In a SMART with k stages, the matrix becomes a hypercube of dimension k. Thus, the average of diagonal entries in the multi-dimensional matrix is the generalization of the estimate of $\widehat{V}[\widehat{d}]$ found in **Eq. 6**. Many of the models for dynamic treatment regimes found in Chakraborty and Moodie (2013) can then be incorporated into our methodology as d, and we may be able to provide many of these models with valid statistical inference. Other statistics computed from this multi-dimensional matrix may be generalized as well.

Our choices of d_0 explored herein were limited to the **random** or the **best** procedures (see **Section 3.3**). There may be other business-as-usual allocation procedures to use here that make for more realistic baseline comparisons. For instance, one can modify **best** to only use the better treatment if a two-sample t-test rejects at prespecified Type I error level and otherwise default to **random**. One can further set d_0 to be a regression model or a physician's decision tree model and then use our framework to pit two personalized medicine models against each other.

It might also be useful to consider how to extend our methodology to observational data. The literature reviewed in **Section 2** generally does not require RCT data but "only" a model that accurately captures selection into treatments e.g. if "the [electronic medical record] contained all the patient information used by a doctor to prescribe treatment up to the vagaries and idiosyncrasies of individual doctors or hospitals" (Kallus, 2017, Section 1). This may be a very demanding

requirement in practice. In this paper, we do not even require valid estimates of the true population response surface. In an observational study one would need that selection model to be correct and/or a correct model of the way in which subjects and treatments were paired (Freedman and Berk, 2008). Although assuming one has a model that captures selection, it would be fairly straightforward to update the estimators of **Section 3.4** to inverse weight by the probability of treatment condition (the "IPWE") making inference possible for observational data (Zhang et al., 2012b; Chakraborty and Murphy, 2014; Kallus, 2017).

Another extension would be to drop the requirement of specifying the model f whose specification is a tremendous constraint in practice: what if the practitioner cannot construct a suitable f using domain knowledge and past research? It is tempting to use a machine learning model that will both specify the structure of f and provide parameter estimates within e.g. Kallus's personalization forests (Kallus, 2017) or convolutional neural networks (LeCun and Bengio, 1998) if the raw subject-level data had images. We believe the bootstrap of **Section 3.5** will withstand such a machination but are awaiting a rigorous proof. Is there a solution in the interim? As suggested as early as Cox (1975), we can always pre-split the data in two where the first piece can be used to specify f and the second piece can be injected into our procedure. The cost is less data for estimation and thus, less power available to prove that the personalization is effective.

If we do not split, all the data is to be used and there are three scenarios that pose different technical problems. Under one scenario, a researcher is able to specify a suite of possible models before looking at the data. The full suite can be viewed as comprising a single procedure for which nonparametric bootstrap procedures may in principle provide simultaneous confidence intervals (Buja and Rolke, 2014). Under the other two scenarios, models are developed inductively from the data. This problem is more acute for instance in Davies (2015) where high-dimensional genomic data is incorporated for personalization (e.g. where there are many more SNPs than patients in the RCT). If it is possible to specify exactly how the model search is undertaken (e.g., using the lasso), some forms of statistical inference may be feasible. This is currently an active research area; for instance, Lockhart et al. (2014) and Lee et al. (2016) develop a significance test for the lasso and there is even some evidence to suggest that the double-peeking is not as problematic as the community has assumed (Zhao et al., 2020).

Our method's generalizability to future patients is also in question as our validation was done within the patients of a RCT. The population of future patients is likely not the same as the population of patients in the RCT. Future patients will likely have wider distributions of the p covariates as typical RCTs feature strict inclusion criteria sometimes targeting high risk patients for higher outcome event rates. A good discussion of these issues is found in Rosenberger and Lachin (2016), Chapter 6. The practitioner will have to draw on experience and employ their best judgment to decide if the estimates our methodology provides will generalize.

And of course, the method herein only evaluates if a personalization scheme works on average over an entire population. "Personalized medicine" eponymously refers to personalization for an individual. Ironically, that is not the goal herein, but we do acknowledge that estimates and inference at an individual level coupled to valid inference for the improvement score is much-needed. This is not without difficulty as clinical trials are typically not powered to examine subgroup effects. A particularly alarming observation is made by Cuijpers et al. (2012), page 7, "if we want to have sufficient statistical power to find clinically relevant differential effect sizes of 0.5, we would need . . . . about 23,000 patients".

DATA AVAILABILITY STATEMENT

The datasets presented in this article in **Section 4.3** is not readily available because the clinical data cannot be anonymized sufficiently according to the IRB guidelines of the institutions where the study was performed. For more information, contact the authors of the study.

AUTHOR CONTRIBUTIONS

All authors were responsible for development of methodology and drafting the manuscript. Kapelner and Bleich did the data analysis of **Section 4**.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Agresti, A. (2018). *An Introduction to Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons.

Bagby, R. M., Quilty, L. C., Segal, Z. V., McBride, C. C., Kennedy, S. H., and Costa, P. T. (2008). Personality and Differential Treatment Response in Major Depression: a Randomized Controlled Trial Comparing Cognitive-Behavioural Therapy and Pharmacotherapy. *Can. J. Psychiatry* 53, 361–370. doi:10.1177/070674370805300605

Barrett, J. K., Henderson, R., and Rosthøj, S. (2014). Doubly Robust Estimation of Optimal Dynamic Treatment Regimes. *Stat. Biosci.* 6, 244–260. doi:10.1007/s12561-013-9097-6

Berger, J. O., Wang, X., and Shen, L. (2014). A Bayesian Approach to Subgroup Identification. *J. Biopharm. Stat.* 24, 110–129. doi:10.1080/10543406.2013.856026

Berk, R. A., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013b). Valid Post-selection Inference. *Ann. Stat.* 41, 802–837. doi:10.1214/12-aos1077

Berk, R., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K., et al. (2014). Misspecified Mean Function Regression. *Sociological Methods Res.* 43, 422–451. doi:10.1177/0049124114526375

Berk, R., Olson, M., Buja, A., and Ouss, A. (2020). Using Recursive Partitioning to Find and Estimate Heterogenous Treatment Effects in Randomized Clinical Trials. *J. Exp. Criminol.*, 1–20. doi:10.1007/s11292-019-09410-0

Berk, R., Pitkin, E., Brown, L., Buja, A., George, E., and Zhao, L. (2013a). Covariance Adjustments for the Analysis of Randomized Field Experiments. *Eval. Rev.* 37, 170–196. doi:10.1177/0193841x13513025

Bernard, C. (1957). *An Introduction to the Study of Experimental Medicine.* New York, NY: Dover Publications.

Box, G. E. P., and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces.* New York, NY: Wiley.

Brinkley, J., Tsiatis, A., and Anstrom, K. J. (2010). A Generalized Estimator of the Attributable Benefit of an Optimal Treatment Regime. *Biometrics* 66, 512–522. doi:10.1111/j.1541-0420.2009.01282.x

Buja, A., and Rolke, W. (2014). *Calibration for Simultaneity: (Re)sampling Methods for Simultaneous Inference with Applications to Function Estimation and Functional Data.* University of Pennsylvania working paper.

Byar, D. P. (1985). Assessing Apparent Treatment-Covariate Interactions in Randomized Clinical Trials. *Statist. Med.* 4, 255–263. doi:10.1002/sim.4780040304

Byar, D. P., and Corle, D. K. (1977). Selecting Optimal Treatment in Clinical Trials Using Covariate Information. *J. chronic Dis.* 30, 445–459. doi:10.1016/0021-9681(77)90037-6

Chakraborty, B., Laber, E. B., and Zhao, Y.-Q. (2014). Inference about the Expected Performance of a Data-Driven Dynamic Treatment Regime. *Clin. Trials* 11, 408–417. doi:10.1177/1740774514537727

Chakraborty, B., and Moodie, E. E. M. (2013). *Statistical Methods for Dynamic Treatment Regimes.* New York, NY: Springer.

Chakraborty, B., and Murphy, S. A. (2014). Dynamic Treatment Regimes. *Annu. Rev. Stat. Appl.* 1, 447–464. doi:10.1146/annurev-statistics-022513-115553

Chakraborty, B., Murphy, S., and Strecher, V. (2010). Inference for Non-regular Parameters in Optimal Dynamic Treatment Regimes. *Stat. Methods Med. Res.* 19, 317–343. doi:10.1177/0962280209105013

Cohen, Z. D., and DeRubeis, R. J. (2018). Treatment Selection in Depression. *Annu. Rev. Clin. Psychol.* 14, 209–236. doi:10.1146/annurev-clinpsy-050817-084746

Collins, L. M., Murphy, S. A., and Bierman, K. L. (2004). A Conceptual Framework for Adaptive Preventive Interventions. *Prev. Sci.* 5, 185–196. doi:10.1023/b:prev.0000037641.26017.00

Cox, D. R. (1975). A Note on Data-Splitting for the Evaluation of Significance Levels. *Biometrika* 62, 441–444. doi:10.1093/biomet/62.2.441

Cox, D. R. (1958). *Planning of Experiments.* New York, NY: Wiley.

Cuijpers, P., Reynolds, C. F., III, Donker, T., Li, J., Andersson, G., and Beekman, A. (2012). Personalized Treatment of Adult Depression: Medication, Psychotherapy, or Both? a Systematic Review. *Depress. Anxiety* 29, 855–864. doi:10.1002/da.21985

Davies, K. (2015). *The $1,000 Genome: The Revolution in DNA Sequencing and the New Era of Personalized Medicine.* New York, NY: Simon & Schuster.

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application.* 1st Edn. Cambridge, UK: Cambridge University Press.

Dawes, R. M. (1979). The Robust Beauty of Improper Linear Models in Decision Making. *Am. Psychol.* 34, 571–582. doi:10.1037/0003-066x.34.7.571

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L., and Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: Translating Research on Prediction into Individual Treatment Recommendations. *A. Demonstration. PLoS One* 9, e83875. doi:10.1371/journal.pone.0083875

DeRubeis, R. J., Hollon, S. D., Amsterdam, J. D., Shelton, R. C., Young, P. R., Salomon, R. M., et al. (2005). Cognitive Therapy vs Medications in the Treatment of Moderate to Severe Depression. *Arch. Gen. Psychiatry* 62, 409–416. doi:10.1001/archpsyc.62.4.409

DiCiccio, T. J., and Efron, B. (1996). Bootstrap Confidence Intervals. *Stat. Sci.* 11, 189–212. doi:10.1214/ss/1032280214

Dusseldorp, E., Doove, L., and Mechelen, I. v. (2016). QUINT: An R Package for the Identification of Subgroups of Clients Who Differ in Which Treatment Alternative Is Best for Them. *Behav. Res.* 48, 650–663. doi:10.3758/s13428-015-0594-z

Dusseldorp, E., and Van Mechelen, I. (2014). Qualitative Interaction Trees: a Tool to Identify Qualitative Treatment-Subgroup Interactions. *Statist. Med.* 33, 219–237. doi:10.1002/sim.5933

Efron, B. (1987). Better Bootstrap Confidence Intervals. *J. Am. Stat. Assoc.* 82, 171–185. doi:10.1080/01621459.1987.10478410

Efron, B., and Tibshirani, R. (1994). *An Introduction to the Bootstrap.* Boca Raton, FL: CRC Press.

Evans, W. E., and Relling, M. V. (2004). Moving towards Individualized Medicine with Pharmacogenomics. *Nature* 429, 464–468. doi:10.1038/nature02626

Faraway, J. J. (2016). Does Data Splitting Improve Prediction?. *Stat. Comput.* 26, 49–60. doi:10.1007/s11222-014-9522-9

Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., and Berk, M. (2017). The New Field of "precision Psychiatry". *BMC Med.* 15, 1–7. doi:10.1186/s12916-017-0849-x

Foster, J. C. (2013). Subgroup Identification and Variable Selection from Randomized Clinical Trial Data. PhD thesis. Ann Arbor, MI: The University of Michigan.

Fournier, J. C., DeRubeis, R. J., Shelton, R. C., Hollon, S. D., Amsterdam, J. D., and Gallop, R. (2009). Prediction of Response to Medication and Cognitive Therapy in the Treatment of Moderate to Severe Depression. *J. consulting Clin. Psychol.* 77, 775–787. doi:10.1037/a0015401

Freedman, D. A., and Berk, R. A. (2008). Weighting Regressions by Propensity Scores. *Eval. Rev.* 32, 392–409. doi:10.1177/0193841x08317586

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33, 1–22. doi:10.18637/jss.v033.i01

Gail, M., and Simon, R. (1985). Testing for Qualitative Interactions between Treatment Effects and Patient Subsets. *Biometrics* 41, 361–372. doi:10.2307/2530862

Goldberg, Y., and Kosorok, M. R. (2012). Q-learning with Censored Data. *Ann. Stat.* 40, 529. doi:10.1214/12-aos968

Gunter, L., Chernick, M., and Sun, J. (2011a). A Simple Method for Variable Selection in Regression with Respect to Treatment Selection. *Pakistan J. Stat. Operations Res.* 7, 363–380. doi:10.18187/pjsor.v7i2-sp.311

Gunter, L., Zhu, J., and Murphy, S. (2011b). Variable Selection for Qualitative Interactions in Personalized Medicine while Controlling the Family-wise Error Rate. *J. Biopharm. Stat.* 21, 1063–1078. doi:10.1080/10543406.2011.608052

Hastie, T., Tibshirani, R., and Friedman, J. H. (2013). *The Elements of Statistical Learning.* 10th Edn. Berlin, Germany: Springer Science.

Henderson, N. C., Louis, T. A., Rosner, G. L., and Varadhan, R. (2020). Individualized Treatment Effects with Censored Data via Fully Nonparametric Bayesian Accelerated Failure Time Models. *Biostatistics* 21, 50–68. doi:10.1093/biostatistics/kxy028

Henderson, R., Ansell, P., and Alshibani, D. (2010). Regret-regression for Optimal Dynamic Treatment Regimes. *Biometrics* 66, 1192–1201. doi:10.1111/j.1541-0420.2009.01368.x

Hood, L., and Friend, S. H. (2011). Predictive, Personalized, Preventive, Participatory (P4) Cancer Medicine. *Nat. Rev. Clin. Oncol.* 8, 184–187. doi:10.1038/nrclinonc.2010.227

Horowitz, J. L. (2001). Chapter 52 - the Bootstrap (Elsevier), of *Handbook of Econometrics,* Vol. 5, 3159–3228. doi:10.1016/s1573-4412(01)05005-xThe Bootstrap

Hosmer, D. W., and Lemeshow, S. (1999). *Applied Survival Analysis: Time-To-Event.* Hoboken, NJ: Wiley.

Imai, K., and Ratkovic, M. (2013). Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation. *Ann. Appl. Stat.* 7, 443–470. doi:10.1214/12-aoas593

Kallus, N. (2017). Recursive Partitioning for Personalization Using Observational Data. *Proc. 34th Int. Conf. on Machine Learning* 70, 1789–1798.

Kang, C., Janes, H., and Huang, Y. (2014). Combining Biomarkers to Optimize Patient Treatment Recommendations. *Biom* 70, 695–707. doi:10.1111/biom.12191

Kapelner, A., Bleich, J., Levine, A., Cohen, Z. D., DeRubeis, R. J., and Berk, R. (2014). Inference for the Effectiveness of Personalized Medicine with Software. Available at: http://arxiv.org/abs/1404.7844.

Kapelner, A., and Krieger, A. (2020). A Matching Procedure for Sequential Experiments that Iteratively Learns Which Covariates Improve Power. Available at: http://arxiv.org/abs/2010.05980.

Laber, E. B., Lizotte, D. J., and Ferguson, B. (2014). Set-valued Dynamic Treatment Regimes for Competing Outcomes. *Biom* 70, 53–61. doi:10.1111/biom.12132

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *Am. Econ. Rev.* 76, 604–620.

Lamont, A., Lyons, M. D., Jaki, T., Stuart, E., Feaster, D. J., Tharmaratnam, K., et al. (2018). Identification of Predicted Individual Treatment Effects in Randomized Clinical Trials. *Stat. Methods Med. Res.* 27, 142–157. doi:10.1177/0962280215623981

LeCun, Y., and Bengio, Y. (1998). "Convolutional Networks for Images, Speech, and Time Series," in *The Handbook of Brain Theory and Neural Networks*. Editor M. A. Arbib (Cambridge: MIT Press), 255–258.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact Post-selection Inference, with Application to the Lasso. *Ann. Stat.* 44, 907–927. doi:10.1214/15-aos1371

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A Significance Test for the Lasso. *Ann. Stat.* 42, 413–468. doi:10.1214/13-aos1175

Lu, W., Zhang, H. H., and Zeng, D. (2013). Variable Selection for Optimal Treatment Decision. *Stat. Methods Med. Res.* 22, 493–504. doi:10.1177/0962280211428383

Ma, J., Stingo, F. C., and Hobbs, B. P. (2019). Bayesian Personalized Treatment Selection Strategies that Integrate Predictive with Prognostic Determinants. *Biometrical J.* 61, 902–917. doi:10.1002/bimj.201700323

McGrath, C. L., Kelley, M. E., Holtzheimer, P. E., Dunlop, B. W., Craighead, W. E., Franco, A. R., et al. (2013). Toward a Neuroimaging Treatment Selection Biomarker for Major Depressive Disorder. *JAMA psychiatry* 70, 821–829. doi:10.1001/jamapsychiatry.2013.143

McKeague, I. W., and Qian, M. (2014). Estimation of Treatment Policies Based on Functional Predictors. *Stat. Sin* 24, 1461–1485. doi:10.5705/ss.2012.196

Moodie, E. E. M., and Richardson, T. S. (2009). Estimating Optimal Dynamic Regimes: Correcting Bias under the Null: [Optimal Dynamic Regimes: Bias Correction]. *Scand. Stat. Theor. Appl* 37, 126–146. doi:10.1111/j.1467-9469.2009.00661.x

Moodie, E. E. M., Richardson, T. S., and Stephens, D. A. (2007). Demystifying Optimal Dynamic Treatment Regimes. *Biometrics* 63, 447–455. doi:10.1111/j.1541-0420.2006.00686.x

Murphy, S. A. (2005a). A Generalization Error for Q-Learning. *J. Mach Learn. Res.* 6, 1073–1097.

Murphy, S. A. (2005b). An Experimental Design for the Development of Adaptive Treatment Strategies. *Statist. Med.* 24, 1455–1481. doi:10.1002/sim.2022

Murphy, S. A. (2003). Optimal Dynamic Treatment Regimes. *J. R. Stat. Soc. Ser. B (Statistical Methodology)* 65, 331–355. doi:10.1111/1467-9868.00389

Nemeroff, C. B., Heim, C. M., Thase, M. E., Klein, D. N., Rush, A. J., Schatzberg, A. F., et al. (2003). Differential Responses to Psychotherapy versus Pharmacotherapy in Patients with Chronic Forms of Major Depression and Childhood Trauma. *Proc. Natl. Acad. Sci.* 100, 14293–14296. doi:10.1073/pnas.2336126100

Orellana, L., Rotnitzky, A., and Robins, J. M. (2010). Dynamic Regime Marginal Structural Mean Models for Estimation of Optimal Dynamic Treatment Regimes, Part I: Main Content. *The Int. J. biostatistics* 6, 1–47. doi:10.2202/1557-4679.1200

Pan, G., and Wolfe, D. A. (1997). Test for Qualitative Interaction of Clinical Significance. *Statist. Med.* 16, 1645–1652. doi:10.1002/(sici)1097-0258(19970730)16:14<1645::aid-sim596>3.0.co;2-g

Paul, G. L. (1967). Strategy of Outcome Research in Psychotherapy. *J. consulting Psychol.* 31, 109–118. doi:10.1037/h0024436

Qian, M., and Murphy, S. A. (2011). Performance Guarantees for Individualized Treatment Rules. *Ann. Stat.* 39, 1180–1210. doi:10.1214/10-aos864

Rice, J. A. (1994). *Mathematical Statistics and Data Analysis*. 2nd Edn. Belmont, CA: Duxbury Press.

Robins, J. M. (2004). "Optimal Structural Nested Models for Optimal Sequential Decisions," in Proceedings of the Second Seattle Symposium on Biostatistics. (Springer), 189–326. doi:10.1007/978-1-4419-9076-1_11

Robins, J., Orellana, L., and Rotnitzky, A. (2008). Estimation and Extrapolation of Optimal Treatment and Testing Strategies. *Statist. Med.* 27, 4678–4721. doi:10.1002/sim.3301

Rolling, C. A., and Yang, Y. (2014). Model Selection for Estimating Treatment Effects. *J. R. Stat. Soc. B* 76, 749–769. doi:10.1111/rssb.12043

Rosenbaum, P. R. (2002). *Observational Studies*. New York, NY: Springer.

Rosenberger, W. F., and Lachin, J. M. (2016). *Randomization in Clinical Trials: Theory and Practice*. 2nd Edn. Hoboken, NJ: John Wiley & Sons.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *J. Educ. Psychol.* 66, 688–701. doi:10.1037/h0037350

Rubin, D. B., and van der Laan, M. J. (2012). Statistical Issues and Limitations in Personalized Medicine Research with Clinical Trials. *Int. J. biostatistics* 8, 1–20. doi:10.1515/1557-4679.1423

Salazar de Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., et al. (2020). Implementing Precision Psychiatry: A Systematic Review of Individualized Prediction Models for Clinical Practice. *Schizophrenia Bulletin* 47, 284–297. doi:10.1093/schbul/sbaa120

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q-and A-Learning Methods for Estimating Optimal Dynamic Treatment Regimes. *Stat. Sci. a Rev. J. Inst. Math. Stat.* 29, 640–661. doi:10.1214/13-sts450

Shao, J. (1994). Bootstrap Sample Size in Nonregular Cases. *Proc. Amer. Math. Soc.* 122, 1251. doi:10.1090/s0002-9939-1994-1227529-8

Shen, Y., and Cai, T. (2016). Identifying Predictive Markers for Personalized Treatment Selection. *Biom* 72, 1017–1025. doi:10.1111/biom.12511

Shuster, J., and van Eys, J. (1983). Interaction between Prognostic Factors and Treatment. *Controlled Clin. trials* 4, 209–214. doi:10.1016/0197-2456(83)90004-1

Silvapulle, M. J. (2001). Tests against Qualitative Interaction: Exact Critical Values and Robust Tests. *Biometrics* 57, 1157–1165. doi:10.1111/j.0006-341x.2001.01157.x

Smith, R. (2012). British Medical Journal Group Blogs. *Stratified, Personalised or Precision Medicine*. Available at: https://blogs.bmj.com/bmj/2012/10/15/richard-smith-stratified-personalised-or-precision-medicine/. Online. (Accessed July 14, 2021).

Su, X., Tsai, C. L., and Wang, H. (2009). Subgroup Analysis via Recursive Partitioning. *J. Machine Learn. Res.* 10, 141–158.

van der Laan, M. J., and Luedtke, A. R. (2015). Targeted Learning of the Mean Outcome under an Optimal Dynamic Treatment Rule. *J. causal inference* 3, 61–95. doi:10.1515/jci-2013-0022

Weitz, E. S., Hollon, S. D., Twisk, J., Van Straten, A., Huibers, M. J. H., David, D., et al. (2015). Baseline Depression Severity as Moderator of Depression Outcomes Between Cognitive Behavioral Therapy vs Pharmacotherapy. *JAMA psychiatry* 72, 1102–1109. doi:10.1001/jamapsychiatry.2015.1516

Weston, A. D., and Hood, L. (2004). Systems Biology, Proteomics, and the Future of Health Care: toward Predictive, Preventative, and Personalized Medicine. *J. Proteome Res.* 3, 179–196. doi:10.1021/pr0499693

Yakovlev, A. Y., Goot, R. E., and Osipova, T. T. (1994). The Choice of Cancer Treatment Based on Covariate Information. *Statist. Med.* 13, 1575–1581. doi:10.1002/sim.4780131508

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012a). Estimating Optimal Treatment Regimes from a Classification Perspective. *Stat* 1, 103–114. doi:10.1002/sta.411

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012b). A Robust Method for Estimating Optimal Treatment Regimes. *Biometrics* 68, 1010–1018. doi:10.1111/j.1541-0420.2012.01763.x

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust Estimation of Optimal Dynamic Treatment Regimes for Sequential Treatment Decisions. *Biometrika* 100, 681–694. doi:10.1093/biomet/ast014

Zhao, S., Witten, D., and Shojaie, A. (2020). Defense of the Indefensible: A Very Naive Approach to High-Dimensional Inference. *arXiv*.

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015). New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes. *J. Am. Stat. Assoc.* 110, 583–598. doi:10.1080/01621459.2014.937488

Zhao, Y., and Zeng, D. (2013). Recent Development on Statistical Methods for Personalized Medicine Discovery. *Front. Med.* 7, 102–110. doi:10.1007/s11684-013-0245-7

Zhou, Y., Wilkinson, D., Schreiber, R., and Pan, R. (2008). Large-scale Parallel Collaborative Filtering for the Netflix Prize. *Lecture Notes Comput. Sci.* 5034, 337–348. doi:10.1007/978-3-540-68880-8_32