Machine Learning for Treatment Assignment: Improving Individualized Risk Attribution

Jeremy Weiss, PhD[1], Finn Kuusisto, MS[1], Kendrick Boyd, PhD[1], Jie Liu, PhD[2], David Page, PhD[1]
[1] University of Wisconsin, Madison, WI [2] University of Washington, Seattle, WA

Abstract

Clinical studies model the average *treatment effect (ATE), but apply this population-level effect to future individuals. Due to recent developments of machine learning algorithms with useful statistical guarantees, we argue instead for modeling the* individualized *treatment effect (ITE), which has better applicability to new patients. We compare ATE-estimation using randomized and observational analysis methods against ITE-estimation using machine learning, and describe how the ITE theoretically generalizes to new population distributions, whereas the ATE may not. On a synthetic data set of statin use and myocardial infarction (MI), we show that a learned ITE model improves true ITE estimation and outperforms the ATE. We additionally argue that ITE models should be learned with a consistent, non-parametric algorithm from unweighted examples and show experiments in favor of our argument using our synthetic data model and a real data set of D-penicillamine use for primary biliary cirrhosis.*

Introduction

Estimation of the risk of a disease attributable to an exposure or treatment is an important task in epidemiology and is typically determined using randomized controlled trials (RCTs). The *average* treatment effect (ATE)–the primary outcome of an RCT–is the average difference between treatment arms in the probability of the outcome, which is then used to recommend future treatments for individual patients. While ATEs are indicative of true treatment effects even in the presence of confounders, they have limited applicability for individual patients because we do not expect the same treatment effect in every person and diversity of effects goes beyond a population's nonuniform prior risk. In addition, the ATE is population-distribution dependent, so it inherently lacks generalizability to alternative test distributions. Therefore, we consider modeling the *individualized* treatment effect (ITE), which is the effect of administering the treatment to a particular patient specified by a set of recorded features. Access to the ITE, in addition to the ATE, has many important clinical applications, with just one illustrated below.

Medication Use: Medications almost certainly have different effects in different individuals. For example, hormone replacement therapy treatment effect findings in RCTs and observational studies were of opposite sign for coronary heart disease, and advocacy of their use was rescinded when the RCT findings were released[1]. Yet, estrogen therapy is still the first line treatment among women experiencing hot flushes. This raises the question of whether ITE modeling can help determine subsets of patients who are still likely to receive benefit. Similarly, many drugs are taken off the market due to excess harm from adverse drug effects. Accurate ITE estimation could bring such drugs safely back to market for select subpopulations.

Currently, non-randomized epidemiological studies adopt classical statistical procedures, such as logistic regression (LR), in seeking to improve patient outcomes. However, machine learning has developed many alternative models for conditional probability distributions (CPDs) with numerous advances achieved. These advances should be leveraged in estimation of the treatment effect–a crucial epidemiological outcome of interest. Our work proposes the use of non-parametric algorithms possessing consistency results in place of logistic regression because of their theoretical ability to accurately recover the CPDs.

In this work, we show the value of ITE over ATE as well as the use of conditional probability models over logistic regression, using both synthetic and real data. We demonstrate our ability to recover the true ITE in synthetic data, and we show the generalizability of the conditional probability model to alternative population distributions of increasing Kullback-Leibler (KL) divergences. We also show that a conditional probability model learned with a consistent, non-parametric algorithm achieves a lower mean squared error (MSE) estimate of the ITE than logistic regression. Furthermore, we show that the conditional probability model produces a better estimate of the ITE than logistic regression on a real RCT data set of D-penicillamine use for primary biliary cirrhosis. Additionally, we show that learning from propensity-score matched examples and stable inverse probability of treatment-weighted examples do not improve over unweighted examples for making ITE prediction when only observational data is available. Thus, by casting treatment effect estimation in a machine learning framework, we introduce ways machine learning can be

used to develop improved, personalized-risk estimates and treatment recommendations.

Background

Randomized controlled trials are the gold standard for estimating the average treatment effect. They randomize patients to different treatment arms and measure the rate or probability of an outcome. The treatment arm with the highest success rate determines the preferred treatment, and the conclusion is that future patients who fit the entry criterion of the study should all get the preferred treatment. Randomization is crucial to balance confounders, which are covariates that lead to the outcome and are associated with the treatment. Randomization also balances confounders not measured in the study, so the conclusion is free of confounding bias in expectation.

In general, one cannot observe both "potential outcomes", *i.e.*, know what will happen to a specific patient under each treatment arm. The treatment that is given elicits the "true" outcome, and the treatment(s) not given elicits the "counterfactual" outcome. The counterfactual outcome is impossible to measure, but with randomization and the assumption that patients are drawn from an underlying population distribution, the expected outcome of patients assigned to a treatment arm is the same as the expected outcome of patients with the same treatment, true or counterfactual. Thus, RCTs provide a recommendation about the treatment effect for every treatment arm in the study for every patient.

Confounding in Observational Studies: The RCT, however, is impractical or infeasible for many exposure-outcome pairs. For example, randomization to a harmful treatment, such as smoking, is unethical. In such cases, observational studies are used to derive risk attribution statements. These include studies that use known-confounder-modeling[2], propensity scoring[3,4], inverse probability of treatment-weighting[5], and doubly-robust estimators[6]. The two main ideas in these methods are to (1) adjust for confounders by modeling them, and (2) manipulate the population distribution so that the treatment is independent of confounders given the outcome. These methods rely upon modeling, but cannot do so effectively if they are missing important contributors to their model: the unobserved confounders. Thus, one key assumption in all of these methods is that there are no unobserved confounders (NUCA), which is difficult to determine in practice. Also, in most of these approaches, a model is assumed for the CPD of the outcome given the exposure and covariate. In these cases, the counterfactual outcomes, which are never observed, are assumed to follow the model CPD.

A second assumption made in clinical studies is the exclusion of intermediate variables–covariates that are on the causal pathway from the treatment to the disease. If included, the treatment effect is underestimated because the effect can be "explained away" by the intermediate variable. The exclusion of intermediate variables decreases the richness of the model, as the intermediate variable may also modify the treatment effect, and analyses that acknowledge and integrate this information exist[7].

Individualized Treatment Effect: One critical drawback of all these methods is that they seek to calculate the average treatment effect, when most applications of risk attribution really desire the ITE. The ITE provides the effect per individual instead of a population-level effect, and information about future individuals can be leveraged in determining optimal treatment choices. Unlike in ATE estimation though, acquiring sufficient counts to estimate the counterfactual ITE outcome is unachievable for any moderate-sized feature vector because the number of possible feature states is exponentially large. Therefore, a modeling approach to estimate the counterfactual outcome becomes necessary. These can be the same CPD models used in pseudo-randomized ATE estimation, *e.g.* logistic regression, but in the Methods section we will discuss two reasons to adopt other machine learning models: non-uniform treatment recommendations and non-parametric consistency.

The call for adoption of the ITE is not new, and the limitations of applying population-average effects on individuals has been noted[8,9]. The ATE or relative risk is stated as the primary outcome, usually followed by a secondary analysis of the heterogeneity of treatment effect. As mentioned in Hayward et al.[10], performing subgroup treatment effects is usually more effective in risk-stratified subgroups derived from multivariate analyses than in subgroups defined by individual covariates, and these methods have been adopted for approximating individualized treatment effects[11]. While these methods do improve finer-grained treatment effect estimates, factors beyond the outcome risk may influence the treatment effect and can be utilized when modeling the ITE.

Modeling of the individualized treatment has been implemented in several studies. Qian et al.[12] develop the framework of reward modeling and using model predictions to estimate individualized treatment rules (ITRs). Our work is related but instead makes statements about the utility of the ITE, the generalizability of the ITE, and the preference for using unweighted observational data for ITE estimation, all with simulations to illustrate these advantages. Our simulations

based on synthetic data have access to a ground truth ITE, which we use to assess our ITE estimations.

However, it is possible to assess the benefit of ITE without access to ground truth. Vickers et al.[13] provides an unbiased method to estimate the advantages of using the ITE recommendation over the ATE recommendation using existing RCT data. They show that by counting outcomes in the subset of patients where ITE- and ATE- treatment recommendations disagree, the expected difference in treatment recommendations is estimated. Our experiments include such analyses to show that the ITE-recommendation can be estimated without access to the counterfactual outcomes. Unfortunately, this method can be severely underpowered in the case that the ITE- and ATE- treatment recommendations are highly similar, and a power calculation analysis to determine recruitment size would be helpful. Alternatively, a new RCT study could be implemented by randomizing to ITE- and ATE- treatment arms.

The methods we adopt do not directly optimize the individualized treatment recommendation. Instead, we model the conditional probability distribution, and then the differences in probability are determined using the estimates for the treatment effect of the true and counterfactual treatments. Zhao et al.[14] develop a method to directly optimize for the ITR under a surrogate loss function from RCT data. While this method produces individualized recommendations, we believe a model should also provide treatment effect estimates under each treatment arm, because the treatment effect itself is critical information clinically. Also our methods do not require RCT data and scale to multiple treatment arms and factorial designs.

Methods

We describe ITE modeling below. Let the ITE for an outcome $y \in \{0, 1\}$ of a patient with features v given treatment $u \in \{0, 1\}$ be the difference in estimates: $p(y = 1|u = 1, v) - p(y = 1|u = 0, v)$. The key assumption made in these modeling approaches is that both potential outcomes–the true outcomes y_{true} and the counterfactual outcomes y_{cf}–come from the CPD model, that is, $p(y_{\text{cf}}|u, v) = p(y_{\text{true}}|u, v) = p(y|u, v)$ for all u and v. The interpretation of the ITE is only causal if the no unmeasured confounders assumption (NUCA) is made; otherwise, it is just a statement about the difference in outcome probability given a new patient described by (u, v).

If we have a correctly specified model and NUCA holds, for any new patient, we have their ITE that guides our treatment choice. This statement is notably population-distribution free and thus can generalize to arbitrary population distributions of (u, v). The ATE does not have this characteristic; its calculation is dependent on the distribution of (u, v) so its application should be limited to populations with similar covariate distributions unless the treatment effect is believed to be uniform.

Recalling that the application of the RCT-recommended treatment suggests that every patient should receive that treatment, a logistic-regression-based model similarly provides a uniform decision. Its decision will be in agreement with the sign of the treatment parameter. However, in many cases, and particularly in those where the treatment effect has small magnitude but high variance, the optimal treatment decision is nonuniform. Thus, we adopt machine learning methods which can estimate the CPD while also providing nonuniform treatment choices. In particular, we use AdaBoost because it has consistency results and is a non-parametric learning algorithm[15,16]. In other words, AdaBoost will recover the correct CPD given enough examples, and will do so regardless of the train (u, v) distribution provided proper support.

Table 1: Discussed methodologies with positive and negative characteristics in green and red respectively. ATE average treatement effect, ITE individualized treatment effect, RCT randomized controlled trial, PSM propensity-score matching, (s)IPT-W (stabilized) inverse probability-of-treatment weighting, LR logistic regression, CPD conditional probability distribution, NUCA no unmeasured confounders assumption.

Topic	Negative	Positive
ATE	applicability, generalizability	clinical trial gold standard
ITE	hard to estimate in RCT	applicability, generalizability
RCT	impractical	balanced confounders
PSM, (s)IPT-W	NUCA, decreased effective sample size	pseudo-randomized
LR	uniform treatment recommendation	log odds interpretation
CPD	potential outcomes follow model	mature machine-learning methodology

With the adoption of a non-parametric learning algorithm comes the parametric/non-parametric learning trade-off. Parametric models may require smaller sample sizes to learn effectively but are not consistent if misspecified; non-parametric models may require larger sample sizes to achieve good CPD estimates but have arbitrary joint distribution consistency results.

Recall that propensity-score matching (PSM) and (stabilized) inverse probability-of-treatment weighting ((s)IPT-W) are methods to produce pseudo-randomized data for the estimation of the ATE[3,4,5]. With ITE as the target statistic, these methods become less desirable. In modeling the CPD, PSM and IPT-W weighting reduce the effective sample size, reducing our numbers for estimation. Thus under the modeling assumption and the goal of modeling ITE, we argue for unweighted CPD estimation. Table 1 compares and summarizes the advantages and disadvantages of study methods related to ATE and ITE estimation.

Experimental Approach

In this section, we restate the claims and reasoning in support of the individualized risk framework and provide ways to confirm them experimentally using synthetic data with access to ground truth, or using observational or RCT data.

As already noted, there is a strong argument for estimating the ITE over the ATE because the ITE is applicable for patient-specific recommendations as opposed to ATE-based, population-average recommendations. With correct specification and NUCA, the ITE is also generalizable to arbitrary population distributions, though it is harder to estimate than the ATE. The value of the ITE recommendation can be compared against an alternative–for example, ATE recommendation–using the subset of randomized patients where treatment recommendations differ[13]. With these methods, we test the hypothesis of ITE superiority and illustrate the benefits of ITE estimation on synthetic data.

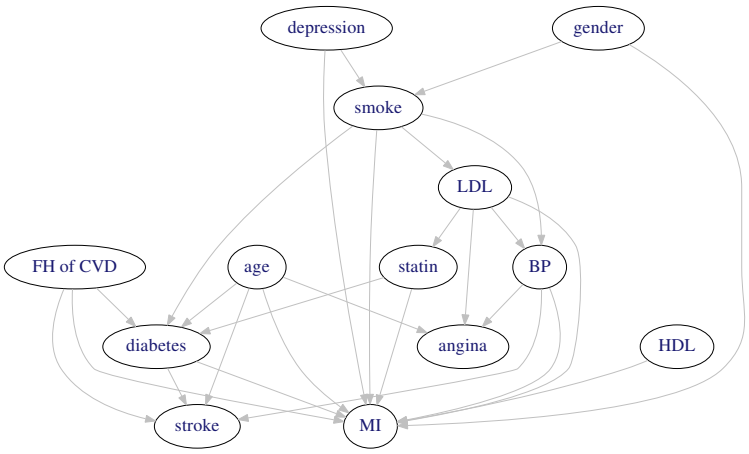
We suggest that, in preference for generalizability of study outcome, the conditional probability distribution $p(y|u, v)$ should be modeled with non-parametric learning algorithms. That is, our goal should be to learn the correct $p(y|u, v)$ irrespective of the distribution $p(u, v)$ because future data distributions $p'(u, v)$ may be different. Non-parametric learning algorithms achieve independence from $p(u, v)$ in the limit of increasing data. We use a synthetic dataset to empirically characterize the recovery of the ITE varying the training set data size and compare the performance of parametric and non-parametric learners varying the similarity of train and test set population distributions. We also use a real RCT data set to compare the treatment assignment policies of parametric and non-parametric learners in the presence of a substantial average treatment effect. Additionally, we use synthetic data to compare ITE estimation generalizability for parametric and non-parametric learners when the test set distribution $p(u, v)$ varies from the training set distribution, though the conditional probability distribution $p(y|u, v)$ remains the same. We also show experimentally that estimating $p(y|u, v)$ directly from the original data distribution outperforms analogous estimators from propensity-score-weighting and similarly to stabilized inverse probability-of-treatment weighting.

Finally, we discuss applications of the conditional probability distribution modeling approach. Numerous concerns have been voiced about the appropriateness of observational data as a data source for the effect of treatments because confounding can bias the statistical interpretation. With free reign on the covariate definitions in observational studies, we may have access to highly-correlated or even logically-related covariates, such as "ever smoked" and "current smoker," in our analysis. We opt to include such covariates for richness of representation that can lead to better estimates of $p(y|u, v)$, but must adapt our interpretation of "intervention" to specified multivariate changes instead of a (univariate) change of treatment state. We discuss several desired conditions when defining the set of "treatment" states and propose methods to provide interpretable recommendations when the space of "treatment" states is large.

Experiments

We first define a synthetic model of myocardial infarction (MI) with thirteen variables: age, gender, smoking status, HDL level, LDL level, diabetes, family history of cardiovascular disease (CVD), blood pressure, history of angina, history of stroke, history of depression, statin use, and MI. For simplicity, each variable is binary. The joint probability distribution is defined by the causal Bayesian network in Figure 1 with hand-crafted conditional probability distributions for each variable informed by medical expertise. Observational data are sampled directly from this Bayesian network, while interventions can be simulated by removing incoming edges to the intervention variable and specifying the Bernoulli distribution parameter. Thus, we simulate data from an RCT of statin use by removing the edge from LDL to statin and using a CPD for statins with equal probability of "yes" and "no" to define our RCT data distribution.

The question we seek to answer for our synthetic model is the effect of statin use on heart attack or MI. We test



(a) Causal network

Variable	Obs.	RCT
age (older)	55	55
smoke	28	28
gender (male)	48	48
HDL	27	27
LDL	39	39
diabetes	39	42
family history of CVD	27	27
blood pressure	30	30
history of angina	35	35
history of stroke	5	6
history of depression	27	27
statin use	25	50
MI	8	9

(b) Marginals

Figure 1: Causal Bayesian network for myocardial infarction (MI) and related variables used for synthetic data in our experiments in (a) and marginals for each variable in (b), reported as the probability of a "yes".

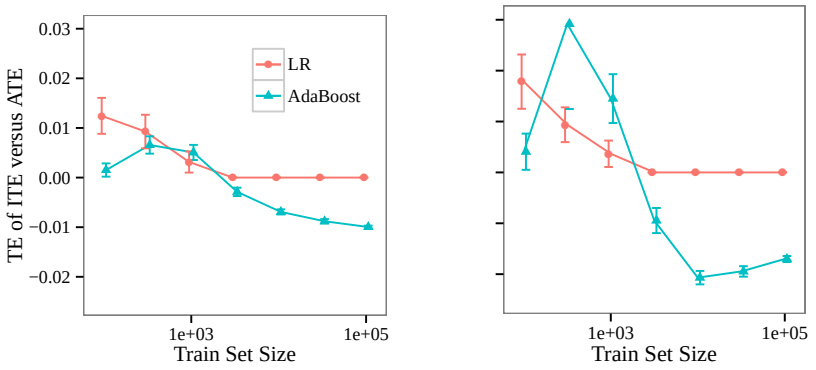
the per-patient recommendations for or against statin use from logistic regression and boosted trees against the ATE recommendation of always prescribing statins. Testing uses data generated from our synthetic, randomized distribution and evaluation is performed using both the RCT method of Vickers et al.[13] and by comparing the predicted ITE to the ground truth ITE calculated exactly from the Bayesian network. Unless otherwise specified, train and test data are generated from the RCT distribution where the LDL to statin edge is removed from the network. When learning our AdaBoost and logistic regression models, we need to ensure the intermediate and confounder assumptions described in the Background section are met. Since we are using synthetic data, there are no unobserved confounders outside our model. However, diabetes is on a causal pathway from statins to MI, so we exclude it from the features available to our models.

With our synthetic model we seek to characterize estimation of the ITE for each method by looking at error modes of each model and producing learning curves for the models as a function of training set size. To test for applicability to an arbitrary test population distribution, we alter the test distributions by changing CPDs for variables with no parents in the causal DAG. Finally, to evaluate ITE estimation from observational data, we use training set data from the observational Bayesian network and compare the estimation from the unweighted training set with estimation using altered data sets via propensity-score matching and stabilized inverse probability-of-treatment weighting.

To validate our claims on real data, we run experiments using trial data for the treatment of primary biliary cirrhosis (PBC) of the liver from the Mayo Clinic[17]. The trial covered a ten-year period and randomized patients across treatment with a placebo versus treatment with D-penicillamine. The data set includes 16 variables, including demographic

Table 2: Statistics for the primary biliary cirrhosis (PBC) dataset censored to a three-year survival period.

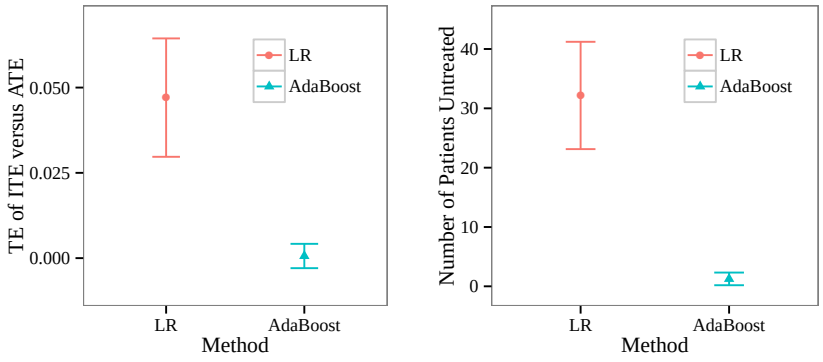
Variable		Count (n=288)	%	Variable	Mean	Std. Dev.
Sex female		255	88.5	Age in years	50.6	10.5
Ascites		24	8.3	Serum bilirunbin (mg/dL)	3.3	4.7
Hepatomegaly		148	51.4	Serum cholesterol (mg/dL)	364.4	224.8
Spider angiomas		84	29.2	Serum albumin (g/dL)	3.5	0.4
Edema	*none*	241	83.7	Urine copper (μg/day)	96.1	84.9
	untreated or successful	27	9.4	Alkaline phosphate (U/L)	2021.1	2213.2
	present despite therapy	20	6.9	Aspartate aminotransferase (U/L)	122.2	58.1
Histologic stage	*1*	16	5.6	Triglycerides (mg/dL)	124.4	65.4
	2	60	20.8	Platelet count	259.1	96.4
	3	112	38.9	Prothrombin time	10.8	1
	4	100	34.7			



(a) Entire population

(b) Subpopulation with differing recommendations

Figure 2: Average difference in treatment effect using the ITE recommendation in place of the ATE recommendation as a function of training set size. The estimated difference in the population is shown in (a); the estimated difference in the subpopulation where treatment recommendations differ is shown in (b). The difference in treatment effect is estimated by the Vickers et al.[13] method with a test set of 100,000 examples. 95% confidence intervals are shown calculated over 100 replications with different training sets.



(a) Entire population ITE vs ATE

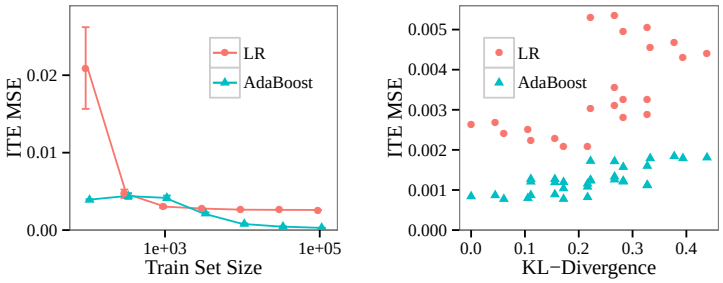
(b) Patients untreated

Figure 3: Average difference in treatment effect using the ITE recommendation in place of the ATE recommendation on the PBC dataset. The estimated difference in the population is shown in (a); the number of patients given a recommendation to not treat is shown in (b). 95% confidence intervals are shown calculated over 100 replications with different sampled training sets.

information like age and sex, as well as various lab tests such as serum albumin, serum cholesterol, and triglycerides (see Table 2). The question we seek to answer for this RCT dataset is the effect of D-penicillamine use on three-year survival. For the three-year survival period, the data set is censored to 288 patients, with 146 in the treatment group and 142 in the placebo group. At the end of three years, the treatment group experienced 27 deaths out of 146, whereas the placebo group experienced 32 out of 142. The trial thus demonstrates an average treatment effect of around 4% reduction, indicating a number needed to treat (NNT) of 25, in death rate at three years.

With the PBC trial data, we seek to compare the estimation of the ITE for each method against the ATE recommendation to treat all patients. Furthermore, given the strength of the average treatment effect, we compare the number of times each method suggests that a patient receive no treatment as opposed to the ATE recommendation. We estimate the average ITE versus ATE and number of untreated patients for each method by running 100 replicates.

We use the 'ada' package implementation of AdaBoost in R to learn the boosted forest[15,16]. Though the consistency guarantees for AdaBoost require the number of iterations to grow linearly with the training set size[18], we use the



(a) ITE learning curves

(b) ITE generalization to different test distributions

Figure 4: Learning curves for logistic regression and AdaBoost showing test set ITE mean-squared error. 95% confidence intervals are calculated from 100 replications. Test set ITE MSE is shown as a function of training set size in (a)nd as a function of KL-divergence between the training set distribution and test set distribution in (b)

square root of the training set size to reduce the computational burden. Otherwise, we use the default settings from the 'ada' package. Logistic regression models are trained using the 'glm' function of R.

Results

Figure 2 shows the utility of adopting the ITE recommendation over the ATE recommendation on our synthetic model. We want to lower the risk of MI, so a negative difference between ITE and ATE is desirable. From Figure 2a, we see that the adoption of the ITE recommendation, as calculated from the AdaBoost model, lowers the probability of MI by 0.01 on average, provided sufficient training data. That is, the number needed to treat is about 100, so treating 100 patients with the ITE-recommended treatment results in one less MI on average than the ATE-recommended treatment of giving everyone statins. Only AdaBoost is able to provide an improved recommendation because of its ability to accurately estimate the conditional probability distribution. Since there are no interaction terms in our logistic regression model, the recommendation converges to the ATE recommendation of giving everyone statins, resulting in the observed difference of 0 for larger training set sizes in Figure 2.

Figure 2b shows the estimated expected difference in probability of MI between ITE- and ATE- recommended treatments among patients where the recommendations disagree on treatment choice. We see that the ITE recommendation lowers the probability of MI in this subset by about 0.02, or a NNT of 50. The upturn for AdaBoost as the training set size approaches 100,000 is likely due to correctly identifying more patients with small benefits from not taking statins. This dilutes the ITE- and ATE- difference among those patients where the recommendations disagree, but the population-wide probability of MI, which is the primary value of interest, continues to decrease.

Figure 3 shows the utility of adopting the ITE recommendation over the ATE recommendation on the PBC trial data. We want to lower the rate of death over a three-year period, so a negative difference between ITE and ATE is desirable, just as it is with our synthetic model. In Figure 3a we see that neither the AdaBoost model nor the logistic regression model outperform the ATE recommendation to treat all patients. While we would prefer to see the ITE outperforming the ATE, this result is not altogether surprising given the effectiveness of treatment. We do, however, see that the AdaBoost model roughly matches the ATE and outperforms the logistic regression model as we hypothesize and demonstrate repeatedly with our synthetic data. In Figure 3b we show the average number of patients for which each model recommends no treatment. The AdaBoost model rarely recommends no treatment, showing that it has effectively learned the treat-all policy, whereas the logistic regression model frequently recommends no treatment.

Learning curves for logistic regression and AdaBoost are shown in Figure 4a. These curves show mean-squared error of the ITE predictions versus training set size. As we expect, the parametric logistic regression converges to a non-zero error because the model is misspecified (since the ground truth model is not log-linear in the exposure and covariates). The error of AdaBoost, however, continues to decrease towards 0 as training set size increases, showing very accurate estimation of the ITE is possible with sufficient data. AdaBoost's approach toward 0 error is in line with the non-parametric consistency results of Bartlett et al.[18].

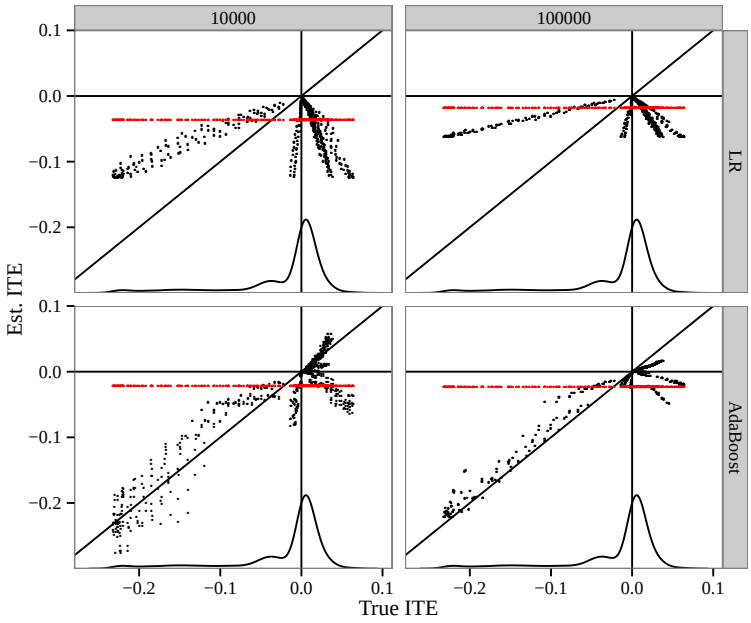


Figure 5: Estimated ITE (black) and ATE (red) versus the ground truth ITE for logistic regression (top) and AdaBoost (bottom) for training set sizes of 10,000 (left) and 100,000 (right). Optimal estimation (*i.e.*, mean-squared error of 0) is given by the line $y = x$. Smoothed, empirical density of the true ITE is shown at the bottom of each plot.

To further investigate the errors in ITE prediction, we show the predicted ITE versus the ground truth ITE (as calculated from our Bayesian network) in Figure 5. Additionally, we plot the ATE which effectively predicts an identical treatment effect for all patients. The goal is to have predictions as close as possible to the true value, *i.e.*, to have points as close to the $y = x$ line as possible. In agreement with the results of Figure 4a, AdaBoost makes better ITE predictions than logistic regression and improves noticeably from 10,000 to 100,000 examples. For logistic regression (top), all ITE estimates will be above or below the line $y = 0$ because the model assumes that a single coefficient determines the direction of the effect. The four groupings of points extending down (at various angles) from the origin correspond to various settings of the variables LDL, HDL, and smoking. This suggests that including interaction terms among statins, LDL, HDL, and smoking in the logistic regression would improve its performance. AdaBoost has the capability to learn arbitrary interactions and can provide individualized recommendations, though the errors remain greater than zero as shown in the bottom of Figure 5. Indeed, some of the groupings of points lie off the $y = x$ line also correspond to certain patient subpopulations for which the ITE is consistently misestimated. We expect, due to the consistency of AdaBoost, that these errors will decrease as more training data is available.

The effect of different data-weighting and matching schemes is shown in Figure 6. The recovery of the CPD model, and thus the ITE, requires the fewest examples when leaving the examples unweighted or using stabilized inverse weighting. While propensity score matching produces worse estimates of ITE, there is no benefit for using stabilized inverse weighting over no weighting for this task. One important consideration is that our data set includes some patients without elevated LDL who take statins, motivated by the suggestion that there could be therapeutic benefit of statins even in borderline hypercholesterolemia. However, in a data set with fewer normal-LDL statin users, propensity-score matching and particularly stabilized inverse weighting will impair the CPD model, because it will attach large excess weight to few examples.

Our final experiment is on the generalizability of ITE predictions for AdaBoost compared to logistic regression. We simulate applying results to different populations by adjusting the prevalence of the five variables in our Bayesian network (see Figure 1) with no parents: age, gender, HDL, depression, family history of CVD. We train on our default RCT data, with marginals shown in Figure 1, but test on modified RCT data with different prevalences of the aforementioned variables. Thus, we have changed $p(v)$, but $p(y|u, v)$ remains the same, so an accurate prediction of CPD, which is exactly $p(y|u, v)$, should handle the changing test distribution gracefully. Figure 4b shows that the MSE does

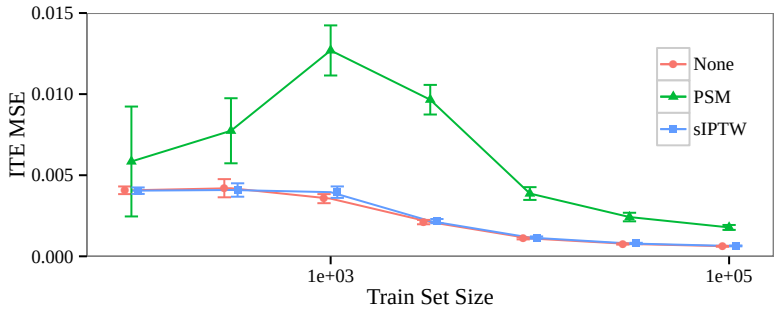


Figure 6: Learning curves for AdaBoost trained on observational data. Test set ITE mean-squared error as a function of training set size is shown comparing training from unweighted examples (None), propensity-score matched samples with a 1:1 ratio (PSM), and stabilized inverse probability of treatment weighted examples (sIPTW). 95% confidence intervals are calculated over 100 replications.

tend to increase for both logistic regression and AdaBoost as the KL-divergence between the train and test distribution increases. However, AdaBoost decreases more slowly, demonstrating that learning a non-parametric, consistent model provides better generalization to other populations.

Discussion and Future Work

Validation on Real Data: Our work highlights the machine learning potential for individualization in clinical scenarios. However, the work we have presented requires empirical justification in several directions. First, while we have provided empirical analysis on one real RCT data set we seek to apply our framework to more clinical data: both RCT and observational. In these settings, we do not have access to the ground truth ITE. Nevertheless, as in Figures 2 and 3, we can adopt the approach in Vickers et al.[13] to compare outcomes from ITE and ATE recommendations. For rigorous evaluation, we must evaluate on several populations with varying covariate distributions. A characterization of which $p(u, v)$ distributions provide reliable ITE recommendations is critical as well. For example, there may be more uncertainty for patients under-represented in the training set, especially with limited data.

Statistical Theory: In addition, a deeper theoretical understanding of the algorithms and evaluations is necessary. For example, investigation of required training set sizes for non-parametric learning algorithms to outperform parametric algorithms, and logistic regression in particular, would be valuable. A characterization of the number of examples needed to move past the mean squared error of the logistic regression for a given task is important as a factor in determining when we should recommend the non-parametric model ITE instead of the ATE. Additionally, improved statistical understanding of the Vickers et al.[13] method is critical. Two such statistical properties of particular interest are (1) a measure of the uncertainty in the differences calculated by the Vickers et al[13] method and (2) power calculations to determine the necessary sample sizes to detect meaningful differences between ATE and ITE.

Epidemiology Methods: Critical factors preventing the automatic use of EHR data for risk attribution are variable definitions, confounders, and intermediate variables. For a pure prediction problem, precisely-defined variables that carry relevant information can improve performance, provided a large enough training set. For risk attribution, however, exclusion of a confounder or inclusion of an intermediate variable can result in biased estimation of both the ITE and ATE. This occurs with ITE prediction from logistic regression or non-parametric learners because the risk is, often arbitrarily, divided between the exposure and the intermediate variable causing inaccurate estimation of the exposure risk. However, we should not simply accept a regime that excludes intermediate variables because inclusion of intermediate variables may enhance our modeling of the conditional probability distribution.

Epidemiological study design often opts for the removal of variables that would actually improve the conditional probability model. We suggest two approaches to explore. First, for interventions, we can define the scope of their influence as a probabilistic change to multiple variables. For example, a diabetes intervention could be represented by the replacement of variable values for “rosaglitazone”, “fasting blood sugar,” and “HbA1c”. Then we can model the effect of intervention by comparing probability of outcomes under intervention or no intervention while richly modeling the conditional probability distribution. Second, a timeline-based analysis where the effect of the intervention

on apparent intermediate variables, as well as the outcome of interest, can be modeled over time might allow all variables to be leveraged for prediction without hindering risk attribution.

Conclusion

In this work, we illustrated the parallels between the standard clinical study framework designed to determine the ATE and the burgeoning clinical study framework designed to determine the ITE. We highlighted shortcomings of the ATE; first, that the ATE is an average outcome, when in practice we usually care about the ITE for future patients, and second, that the ATE is population-distribution dependent. We then discussed modeling of the ITE. Notably, logistic regression can only recommend one treatment arm if we exclude non-linear and exposure-covariate interaction terms because the coefficient for exposure is either negative or non-negative. Furthermore, unless correctly-specified, logistic regression is not a consistent learning algorithm, so we cannot always recover the true conditional probability distribution, even from large populations. Instead, we adopted another popular framework, boosted trees. We showed that the forest-based ITE outperformed the ATE on a synthetic problem of MI prediction using binary variables, and that the forest-based ITE outperformed logistic regression-based ITE on a real problem of primary biliary cirrhosis treatment with D-penicillamine. Additionally, we showed that the use of propensity-score matching and inverse-probability-of-treatment weighting failed to improve the learning of the conditional probability distribution, suggesting that unweighted samples should be used for learning a model of the ITE. Finally, we showed that the forest-based ITE better generalized to different test set distributions than the logistic regression-based ITE. Modeling of the ITE has large theoretical advantages, though robustness guarantees and validation of its performance should be established before large-scale clinical deployment. A few sources of validation include replication studies and heterogeneity of treatment effect analyses using ITE model strata.

References

[1] Manson JE, Chlebowski RT, Stefanick ML, Aragaki AK, Rossouw JE, Prentice RL, et al. Menopausal hormone therapy and health outcomes during the intervention and extended poststopping phases of the Women's Health Initiative randomized trials. JAMA. 2013;310(13):1353–1368.

[2] Prentice R. Use of the logistic model in retrospective studies. Biometrics. 1976;p. 599–606.

[3] Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research. 2011;46(3):399–424.

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

[5] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11(5):550–560.

[6] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005;61(4):962–973.

[7] Robins J. The control of confounding by intermediate variables. Statistics in Medicine. 1989;8(6):679–701.

[8] Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA. 2007;298(10):1209–1212.

[9] Rothwell PM. Can overall results of clinical trials be applied to all patients? The Lancet. 1995;345(8965):1616–1619.

[10] Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Medical Research Methodology. 2006;6(1):18.

[11] Dorresteijn JA, Visseren FL, Ridker PM, Wassink AM, Paynter NP, Steyerberg EW, et al. Estimating treatment effects for individual patients based on the results of randomised clinical trials. BMJ. 2011;343.

[12] Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Annals of Statistics. 2011;39(2):1180.

[13] Vickers AJ, Kattan MW, Sargent DJ. Method for evaluating prediction models that apply the results of randomized trials to individual patients. Trials. 2007;8(1):14.

[14] Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association. 2012;107(499):1106–1118.

[15] Freund Y, Schapire R. Experiments with a new boosting algorithm. In: ICML; 1996. .

[16] Culp M, Johnson K, Michailidis G. ada: An R package for stochastic boosting. Journal of Statistical Software. 2006;17(2):9.

[17] Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. Springer Science & Business Media; 2000.

[18] Bartlett PL, Traskin M. AdaBoost is Consistent. Journal of Machine Learning Research. 2007;8:2347–2368.