Circulation: Cardiovascular Quality and Outcomes

ORIGINAL ARTICLE

# Clinical Value of Predicting Individual Treatment Effects for Intensive Blood Pressure Therapy

## A Machine Learning Experiment to Estimate Treatment Effects from Randomized Trial Data

Tony Duan, BS
Pranav Rajpurkar, MS
Dillon Laird, BS
Andrew Y. Ng, PhD
Sanjay Basu, MD, PhD

**BACKGROUND:** The absolute risk reduction (ARR) in cardiovascular events from therapy is generally assumed to be proportional to baseline risk—such that high-risk patients benefit most. Yet newer analyses have proposed using randomized trial data to develop models that estimate individual treatment effects. We tested 2 hypotheses: first, that models of individual treatment effects would reveal that benefit from intensive blood pressure therapy is proportional to baseline risk; and second, that a machine learning approach designed to predict heterogeneous treatment effects—the X-learner meta-algorithm—is equivalent to a conventional logistic regression approach.

**METHODS AND RESULTS:** We compared conventional logistic regression to the X-learner approach for prediction of 3-year cardiovascular disease event risk reduction from intensive (target systolic blood pressure <120 mm Hg) versus standard (target <140 mm Hg) blood pressure treatment, using individual participant data from the SPRINT (Systolic Blood Pressure Intervention Trial; N=9361) and ACCORD BP (Action to Control Cardiovascular Risk in Diabetes Blood Pressure; N=4733) trials. Each model incorporated 17 covariates, an indicator for treatment arm, and interaction terms between covariates and treatment. Logistic regression had lower C statistic for benefit than the X-learner (0.51 [95% CI, 0.49–0.53] versus 0.60 [95% CI, 0.58–0.63], respectively). Following the logistic regression's recommendation for individualized therapy produced restricted mean time until cardiovascular disease event of 1065.47 days (95% CI, 1061.04–1069.35), while following the X-learner's recommendation improved mean time until cardiovascular disease event to 1068.71 days (95% CI, 1065.42–1072.08). Calibration was worse for logistic regression; it over-estimated ARR attributable to intensive treatment (slope between predicted and observed ARR of 0.73 [95% CI, 0.30–1.14] versus 1.06 [95% CI, 0.74–1.32] for the X-learner, compared with the ideal of 1). Predicted ARRs using logistic regression were generally proportional to baseline pretreatment cardiovascular risk, whereas the X-learner observed—correctly—that individual treatment effects were often not proportional to baseline risk.

**CONCLUSIONS:** Predictions for individual treatment effects from trial data reveal that patients may experience ARRs not simply proportional to baseline cardiovascular disease risk. Machine learning methods may improve discrimination and calibration of individualized treatment effect estimates from clinical trial data.

**CLINICAL TRIAL REGISTRATION:** URL: https://www.clinicaltrials.gov. Unique identifiers: NCT01206062; NCT00000620.

**Key Words:** blood pressure ■ cardiovascular disease ■ calibration ■ machine learning ■ risk factors

https://www.ahajournals.org/journal/circoutcomes

## WHAT IS KNOWN

- Modeling individualized treatment effects can help clinicians identify which patients are more or less likely to benefit from treatment, as compared to the average result from a randomized trial. It is often assumed that the absolute risk reduction an individual will experience from a cardiovascular treatment will be proportional to their baseline cardiovascular risk.
- We explored how alternative machine learning methods may help to identify individualized treatment effects from randomized trial data and whether such effects were necessarily proportional to baseline risk.

## WHAT THE STUDY ADDS

- Machine learning methods revealed that absolute risk reductions from treatment were often not simply proportional to baseline cardiovascular risk, meaning that calculation of individualized treatment effects rather than simply calculating baseline risk would potentially help direct therapy to people most likely to benefit.
- A machine learning method called the X-learner may improve discrimination and calibration of individualized treatment effect estimates from clinical trial data as compared to other machine learning methods and standard logistic regression modeling.

Although models to predict the absolute risk of a disease have been developed for over 5 decades, the increasing availability of individual participant data from randomized trials has led to a new type of modeling: the development of benefit/risk models designed to help clinicians predict whether an individual patient is more likely to experience meaningful benefits or risks from a therapy (ie, whether they may be at the left tail or right tail of the distribution around the average treatment effect in the trial).[1–6] Identifying such heterogeneous treatment effects (HTEs)—or systematically different benefits or risks from a medical therapy among some participants in a study, as compared to the study average result—is clinically important for the application of randomized trial data to patient care.[7] Identifying HTEs can provide clinicians with critical information about whether a therapy with positive average benefit in a trial would be expected to have a similarly positive benefit in a given patient, even greater benefit, no benefit at all, or even harm. Conversely, identifying HTEs can help identify whether a therapy found to have no significant average benefit in a trial may, nevertheless, be useful for a subset of patients.

Numerous articles have highlighted that the absolute risk reduction (ARR) in cardiovascular events from therapy is generally assumed to be proportional to a person's baseline risk—such that high-risk patients will benefit most.[8,9] In this context, it is questionable whether models for individual treatment effects are necessary at all, or whether clinicians can safely assume that a baseline risk model and the average treatment effect reported in the trial will generally predict the absolute benefit of a treatment for their patient.

If an individual treatment effect model has value, a methodological question is what approach can best estimate HTEs to incorporate into an individual treatment effect model. Analysts have recommended developing a model by regressing the primary outcome against trial participant characteristics (demographics, biomarkers, etc), an indicator variable for treatment arm (capturing the average treatment effect), and interaction terms between treatment arm and characteristics (identifying HTEs). Identifying relative HTEs using this multivariable approach has been observed to increase the power of subgroup analyses, as well as to enhance the clinical utility of randomized trial results.[10] Hence, benefit/risk equations for therapeutic decision-making, derived using the multivariable approach, have been increasingly published in high-profile medical journals for treatments such as statin therapy, intensive blood pressure therapy, and antiplatelet therapy.[1,3,5] Yet, HTE effects may be biased if equations are overfit to a single trial population or if they erroneously include irrelevant interaction terms because of collinearity.[11,12] Important covariates may not be known in advance, and complex interactions between many subtle covariates may be missed when performing a standard linear regression, which tends to privilege simple combinations of well-characterized risk factors over complex, nonlinear, or subtle combinations of factors.[12]

To overcome these limitations, machine learning (ML) approaches have been increasingly considered for identifying HTEs. ML approaches can not only identify complex combinations of nonlinear interactions that may predispose to an outcome in a subtle way but also use estimation and cross-validation methods to avoid overfitting. Although ML approaches have been increasingly applied to predict overall disease risk using large datasets (eg, early disease detection through analysis of electronic medical records or registries),[12–16] they have not been rigorously evaluated for specifically detecting HTEs from clinical trial data.

Here, we tested 2 hypotheses: first, that models of individual treatment effects would reveal that benefits from intensive blood pressure therapy are simply proportional to baseline risk—such that individual risk models would not offer clinical value that routine cardiovascular risk models do not; and second, that an ML approach specifically designed to estimate HTEs from individual participant data in clinical trials—a meta-algorithm known as the X-learner[17]—would not be more accurate than a conventional regression approach typically used to create clinical decision rules.

# METHODS

Our code and analytic methods have been made available to other researchers for purposes of replicating the procedure.[18] The data has been made available to other researchers for purposes of reproducing the results, at https://biolincc.nhlbi.nih.gov/home.

We compared metrics of HTE discrimination (for higher versus lower treatment ARR) and calibration (for degree of estimated treatment ARR versus observed treatment ARR) when estimating the cardiovascular disease (CVD) risk reduction over 3 years from intensive blood pressure treatment (targeting a systolic blood pressure <120 mm Hg) versus standard treatment (targeting systolic pressure <140 mm Hg). To do so, we developed HTE models from the SPRINT (Systolic Blood Pressure Intervention Trial) and ACCORD BP (Action to Control Cardiovascular Risk in Diabetes Blood Pressure) randomized trials.[19,20] We focused on a comparison between a conventional logistic regression approach and the X-learner ML approach, although in sensitivity analyses (Methods in the Data Supplement), we explored Cox regression as well as alternative ML methods.

## Data

We used individual participant data from the SPRINT and ACCORD BP trials. The SPRINT trial (N=9361) was a randomized, controlled, open-label trial of intensive versus standard blood pressure treatment among adults without type 2 diabetes mellitus, conducted at 102 clinical sites in the United States between November 2010 and August 2015[19]. The trial was stopped early after a median follow-up of 3.3 years because of a significantly lower rate of the primary composite CVD outcome in the intensive treatment group than in the standard treatment group. Inclusion criteria for the SPRINT trial included: age at least 50 years, systolic blood pressure 130 to 180 mm Hg, and increased CVD event risk (defined as either clinical or subclinical CVD other than stroke; chronic kidney disease, excluding polycystic kidney disease, with an estimated glomerular filtration rate of between 20 and 60 mL/[min·1.73 m$^2$]; a 10-year Framingham risk score of at least 15%; or age at least 75 years). Exclusion criteria included having diabetes mellitus or a prior stroke. By contrast, the ACCORD BP trial (N=4733) was a randomized, controlled, open-label trial of intensive versus standard blood pressure treatment among adults with type 2 diabetes mellitus, conducted at 77 clinical sites in the United States and Canada between January 2003 and June 2009, with a mean follow-up of 4.7 years.[20] Inclusion criteria for the ACCORD BP trial included: age at least 40 years with CVD or at least 55 years with anatomic evidence of substantial atherosclerosis, albuminuria, left ventricular hypertrophy or at least 2 additional CVD risk factors (dyslipidemia, hypertension, smoking, or obesity); systolic blood pressure 130 to 180 mm Hg taking ≤3 blood pressure agents and having a 24-hour protein excretion rate <1 g; and type 2 diabetes mellitus with a hemoglobin A1c level of at least 7.5%. Exclusion criteria included having a body mass index >45 kg/m$^2$, serum creatinine >1.5 mg/dL, or other serious illness. Because the published composite primary outcomes differed between the SPRINT and ACCORD BP trials, we utilized the disaggregated outcome variables in the ACCORD BP dataset to construct the CVD composite outcome matching the SPRINT definition, which was myocardial infarction, other acute coronary syndromes, stroke, heart failure, or death from cardiovascular causes.[19] For the purposes of derivation and validation we used a combined dataset of SPRINT and ACCORD-BP, with an indicator variable for trial equal to 1 for the ACCORD BP participants and 0 for SPRINT participants; summary statistics of the combined and individual datasets are available in Table I in the Data Supplement. We also repeated derivation and validation using just one or the other of the trial datasets, as a robustness check.

## Model Development

### Inverse Probability of Censorship Weighting

To handle the time-to-event nature of the trial data, we dichotomized outcomes as occurrence of CVD event within 3 years and used inverse propensity of censorship weighting.[21,22] Specifically, a Cox regression model was fit using all covariates to estimate participant-specific distributions over time to censoring. Participants whose outcomes were censored before the time of interest were excluded from the dataset used to fit models. Remaining participants were weighted by the inverse of their estimated probability of not being censored by the time of interest (if CVD event did not occur before the time of interest) or by the time of event (if CVD event did occur before the time of interest).

### Conventional Approach

The conventional approach involved a logistic regression model which included all 17 potential covariates, a dummy variable for treatment arm, and all possible interaction terms between the dummy variable for treatment arm and each covariate. This approach was chosen to reflect the strategy adopted by numerous recent HTE modeling analyses using clinical trial data.[1,2,23] It is distinct from the alternative risk-based multivariable HTE assessment approach that involves only a single composite interaction term between a well-established risk score for a primary outcome (eg, the Framingham risk score for CVD, or its more recent variants) and the treatment.[24,25]

### ML Approach

We compared the conventional approach of developing a logistic regression model, above, to a range of alternative ML algorithms adopted in the literature (Methods in the Data Supplement). The primary algorithm we focus on in the main text is the X-learner approach with random forest base learners.[17,26] The X-learner is a meta-algorithm specifically designed for estimating individual treatment effects. We trained the learner through a 3-step process: we first estimated expectations of outcomes given predictors under control and treatment separately, then estimated imputed treatment effects as the difference between expected outcomes and actual outcomes for each individual, and finally predicted the treatment effects as weighted averages of the estimated imputed treatment effects (Figure 1).[17]

These estimation steps could be performed using any ML or regression method as base learners. We chose to use random forests, in which decision trees were built that separate the population into subsets based on combinations of characteristics that help identify those with higher versus lower treatment effects and then an individual's treatment effect was the average of thousands of trees built from subsamples of the dataset. We chose the X-learner with random forests as base learners because random forests have been known to outperform older methods by inherently accounting for interactions among multiple variables as branches of each decision tree in the forest.[27]

In sensitivity analyses (Methods in the Data Supplement), we explored alternative common ML methods and validated the superiority of the X-learner approach. We note as well that the X-learner has been proven to be unbiased in predicting the difference in treatment effect between study arms, unlike older ML methods that can be biased by focusing on variables that predict on the absolute rate of events (eg, risk of CVD) rather than HTEs (eg, how individual features affect the treatment's ability to reduce the risk of CVD).[17,27]

## Outcome Metrics

Our prespecified outcome metrics included: (1) the C statistic for benefit,[28] which is a variant of the common C statistic (area under the receiver operating characteristic curve) designed to specifically calculate the ability for a model to discriminate between people having more benefit versus less benefit from a treatment (rather than just higher versus lower overall CVD risk), (2) the decision value of the 3-year restricted mean survival time (RMST),[29] which is an unbiased estimate of the RMST for patients under the policy implied by the model (ie, treat patients with predicted ARR suggesting lower CVD risk with intensive treatment and do not treat patients with zero or negative predicted ARR), and (3) the slope of the calibration line between predicted ARR (in quintiles) and observed ARR.

The C statistic for benefit was calculated by first matching patient pairs across the 2 study arms on predicted risk reduction at 3 years, then calculating the proportion of matched pairs with unequal observed benefit, in which the pair receiving greater treatment benefit was observed to do so. The decision value of the 3-year RMST was calculated by computing the Kaplan-Meier estimates of survival curves for subgroups of predicted benefit (predicted ARR suggesting lower CVD risk with intensive treatment) and no benefit (zero or negative predicted ARR). The area under each Kaplan-Meier curve, up to 3 years, is defined as the RMST for each subgroup. We then take the average RMST over all participants of the subgroup to which they were assigned—this is an unbiased estimate of the RMST under the treatment choice recommended by the model. The calibration curve was constructed by aggregating participants into quintiles of predicted ARR (where quintiles were chosen to ensure adequate sample size for stable curve estimates[28]). For each quintile, observed ARR was defined as the difference between the intensive versus standard treatment arm CVD rate at 3 years, using Kaplan-Meier estimates to account for censoring. Discrimination was further assessed by comparing bootstrap CIs for observed ARRs in each of the subgroups of predicted benefit and no benefit. Lastly, predicted treatment effects in both the logistic regression approach and the X-learner approach were compared with deciles of baseline pretreatment cardiovascular risk, where the latter was estimated using the American College of Cardiology/American Heart Association atherosclerotic cardiovascular disease risk score estimator[25] to assess whether treatment effects were simply proportional to baseline CVD risk (higher benefit for participants with higher baseline risk), thus potentially obviating the need for an HTE model. We also compared predicted treatment effects to deciles of baseline pretreatment cardiovascular risk where the latter was estimated using 10-year Framingham risk scores.[24]

It is well known that models validated using the same data on which they were trained will be too optimistic due to overfitting.[30] To account for this, the logistic regression model was

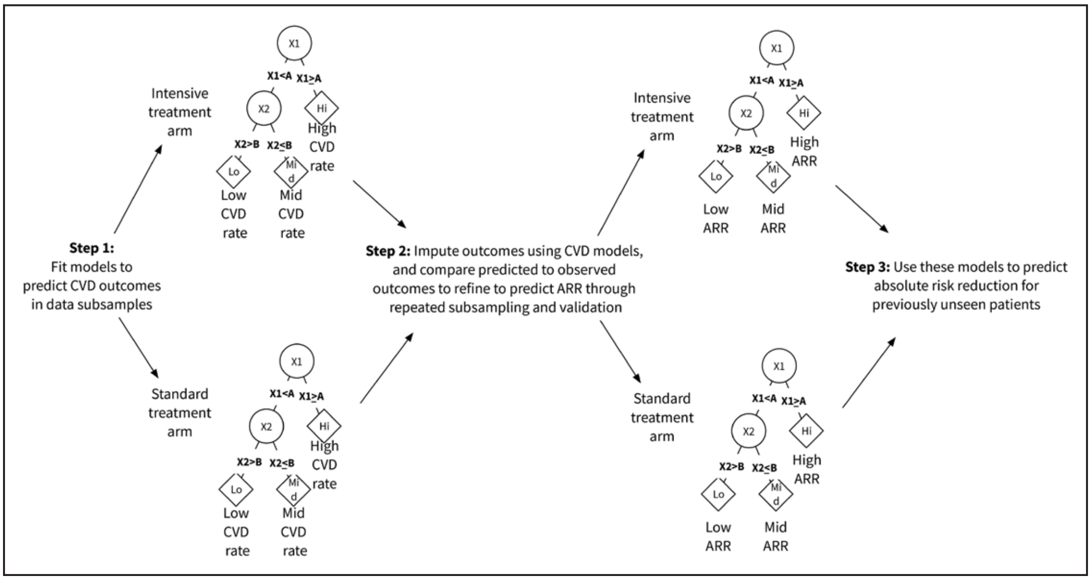

**Figure 1. Conceptualization of the X-learner approach to detect heterogeneous treatment effects from trial data.**
We first estimate expectations of outcomes given predictors under control and treatment separately, then estimate imputed treatment effects (defined as the difference between expected outcomes and actual outcomes), and finally predict individual treatment effects as weighted averages of the estimated imputed treatment effects. For each step in the process, we use random forests with 1000–2000 trees as base learners and predict out-of-bag samples to prevent overfitting. ARR indicates absolute risk reduction; and CVD, cardiovascular disease.

corrected for optimism by calculating C-for-benefit and decision value RMST over 250 bootstrap samples of the dataset, stratified by treatment arm and presence of CVD event. The average difference between performance on the bootstrap dataset (on which the logistic regression model was derived) and the original dataset (on which the model was validated) was defined as the optimism,[28] and subtracted from apparent performance estimates. In the X-learner approach, we calculated predicted risk reduction for each participant only using trees in the random forest base learners for which that participant was out-of-sample, so no participant was used for both derivation and validation of any individual tree. As a result, no optimism correction was necessary for the X-learner approach. All CIs were constructed by bootstrap resampling of the dataset, stratified by treatment arm and presence of CVD event.

Analyses were performed in Python (version 3.6.2; The Python Software Foundation, Wilmington, Delaware) and R (version 3.5.1; The R Foundation for Statistical Computing, Vienna, Austria).

# RESULTS

## Conventional Approach

The conventional logistic regression model produced a model with 17 main terms, a dummy variable for the treatment arm, and 17 HTE interaction terms between the main terms and the treatment arm variable. In descending order, the interaction terms with largest HTE effect sizes (based on the absolute change in effect size based on a 1 unit change in covariate *Z* score) were the interaction term between treatment and statin usage (with statin usage corresponding to less benefit), sex (with being female corresponding to less benefit), HDL (high-density lipoprotein) with higher HDL corresponding to more benefit), and diastolic blood pressure (with higher diastolic blood pressure corresponding to more benefit). No single variable in isolation clearly predicted higher or lower treatment benefit; rather, the groups with benefit and with no benefit had similar demographic and clinical characteristics in univariate analysis, except for non-significantly higher baseline systolic blood pressure and fewer ACCORD BP participants in the group predicted to benefit from intensive treatment (Table 1).

The conventional logistic regression model had a corrected (bootstrap-adjusted) C statistic for benefit of 0.51 (95% CI, 0.49–0.53; Table 2), meaning that it was only slightly better than a flip of a coin for correctly discriminating which of 2 participants in the dataset would have more versus less ARR from intensive blood pressure therapy. Note that the C-for-benefit statistic is in general much more conservative than the traditional C statistic, which only assesses the ability of a model to detect higher versus lower absolute CVD risk, not ARR; the traditional C statistic is determined primarily by the main effect terms (eg, age, sex, tobacco smoking) rather than the HTE interaction terms of a model (eg, interaction between treatment arm and HDL or treatment arm and statin use). The corrected (bootstrap-adjusted) decision value of the 3-year RMST was 1062.86 days (95% CI, 1058.43–1066.74), which is the estimate of the 3-year RMST under the treatment recommended by the model (ie, treat participants with positive predicted risk reduction under treatment and do not treat patients with zero or negative predicted risk reduction). For comparison, a baseline policy of prescribing intensive treatment for participants in the SPRINT trial and standard treatment for those in the ACCORD BP trial results in 3-year decision RMST of 1061.24 days (95% CI, 1057.37–1064.10).

Discrimination was further assessed by bucketing participants into subgroups of predicted benefit (those with predicted ARR suggesting lower CVD risk with intensive treatment) and no benefit (those with zero or negative predicted ARR). We found that the conventional logistic regression model was able to have a positive observed risk reduction of 0.0225 (a 2.3% ARR; 95% CI, 1.8%–2.7%) among the group predicted to have benefit. However, the logistic regression model was not as accurate in predicting the no-benefit group, which had an observed risk reduction CI crossing the zero observed risk reduction line and an average risk increase of 0.0107 (1.1% absolute risk increase) but wide 95% CIs spanning from a 2.7% increase to a 0.8% decrease (Figure 2).

The conventional logistic regression model also had worse calibration, by over-estimating the ARR attributable to intensive blood pressure treatment (Table 2; Figure 3). The slope of the calibration line between the predicted ARR (*x* axis on Figure 3) and observed ARR (*y* axis on Figure 3, difference between intensive and standard treatment arm CVD event rates in quantiles of predicted risk reduction) was 0.73 (95% CI, 0.30–1.14) as compared to the ideal of 1 (Table 2).

Predicted treatment effects using the traditional logistic regression approach were generally proportional to baseline pretreatment cardiovascular risk, with the range of predicted treatment benefit being proportionately higher with higher baseline risk (Figure 4, Table II in the Data Supplement). When compared with baseline pretreatment cardiovascular risk as defined by Framingham risk scores instead of American College of Cardiology/American Heart Association atherosclerotic cardiovascular disease risk scores, results did not meaningfully differ (Figure I in the Data Supplement).

When the discrimination and calibration outcome metrics were computed separately on the SPRINT and ACCORD BP datasets, rather than on the combined set alone, results did not meaningfully differ between the trials (Tables III and IV and Figures II and III in the Data Supplement). When outcome metrics were computed for 5-year outcomes instead of 3-year outcomes using

**Table 1. Summary Statistics of Participants in the Combined Dataset of the SPRINT and ACCORD BP Trials, Partitioned into Predicted Subgroups of Benefit or No Benefit as Determined by Machine Learning (Left) and Conventional (Right) Methods**

| | Mean [SD] using Machine Learning | | Mean [SD] using Conventional | |
|---|---|---|---|---|
| Covariates | Benefit (N=9763) | No benefit (N=3841) | Benefit (N=11029) | No benefit (N=2575) |
| Age, y | 66.67 (9.13) | 65.34 (8.07) | 67.51 (8.85) | 61.08 (6.79) |
| Female, fraction | 0.39 (0.49) | 0.43 (0.50) | 0.34 (0.47) | 0.66 (0.47) |
| Black, fraction | 0.30 (0.46) | 0.26 (0.44) | 0.29 (0.45) | 0.28 (0.45) |
| Hispanic, fraction | 0.09 (0.29) | 0.10 (0.30) | 0.06 (0.23) | 0.26 (0.44) |
| Systolic blood pressure, mm Hg | 141.20 (16.06) | 135.67 (13.75) | 140.66 (15.61) | 135.25 (15.01) |
| Diastolic blood pressure, mm Hg | 78.93 (11.23) | 73.58 (11.16) | 78.15 (11.62) | 74.30 (10.19) |
| No. of blood pressure treatment classes | 1.78 (1.06) | 1.81 (1.05) | 1.80 (1.06) | 1.76 (1.02) |
| Current smoker, fraction | 0.10 (0.30) | 0.07 (0.26) | 0.10 (0.30) | 0.07 (0.26) |
| Former smoker, fraction | 0.44 (0.50) | 0.46 (0.50) | 0.47 (0.50) | 0.35 (0.48) |
| Aspirin, fraction | 0.51 (0.50) | 0.53 (0.50) | 0.53 (0.50) | 0.47 (0.50) |
| Statin, fraction | 0.45 (0.50) | 0.65 (0.48) | 0.43 (0.50) | 0.82 (0.39) |
| Serum creatinine, mg/dL | 1.01 (0.31) | 1.04 (0.35) | 1.03 (0.29) | 1.00 (0.43) |
| Total cholesterol, mg/dL | 191.70 (41.51) | 189.33 (43.54) | 192.00 (41.64) | 186.86 (43.83) |
| High-density lipoprotein cholesterol, mg/dL | 51.78 (14.29) | 48.27 (14.47) | 52.39 (14.50) | 43.94 (11.91) |
| Triglycerides, mg/dL | 133.13 (79.20) | 180.02 (191.34) | 127.95 (73.82) | 225.29 (223.00) |
| Body mass index, kg/m$^2$ | 30.69 (5.79) | 30.58 (5.73) | 30.59 (5.88) | 30.93 (5.31) |
| ACCORD BP participants, fraction | 0.27 (0.44) | 0.50 (0.50) | 0.26 (0.44) | 0.65 (0.48) |

The benefit bucket consists of participants predicted to have ARR >0, and the no-benefit bucket consists of participants predicted to have ARR ≤0. ACCORD BP indicates Action to Control Cardiovascular Risk in Diabetes Blood Pressure; ARR, absolute risk reduction; and SPRINT, Systolic Blood Pressure Intervention Trial.

the ACCORD BP dataset (the trial for which the majority of participants had follow-up times through 5 years, unlike SPRINT), results did not meaningfully differ (Table V in the Data Supplement).

## ML Approach

The X-learner ML approach predicted more benefit from intensive blood pressure treatment for older participants, those not on statins, and those with lower triglycerides, but few measures differed remarkably between the predicted benefit and predicted no-benefit subgroups (Table 1), suggesting complex interactions rather than simple univariate subgroups were identified by the approach.

The X-learner ML approach had a corrected C statistic for benefit of 0.60 (95% CI, 0.58–0.63); Table 2, which was significantly better than the conventional logistic regression approach for correctly discriminating which of 2 participants in the dataset would have more versus less ARR from intensive blood pressure therapy. The corrected 3-year decision value of RMST was 1068.71 days (95% CI, 1065.42–1072.08), which is higher than the corresponding statistic using the logistic regression approach. For comparison, a treatment approach of prescribing intensive treatment for those in the SPRINT trial and standard treatment for those in the ACCORD BP trial resulted in a 3-year decision RMST of 1061.24 days (95% CI, 1057.37–1064.10).

The X-learner approach was able to produce a positive observed ARR of 0.0356 (a 3.6% ARR; 95% CI, 2.9%–4.2%) among the group predicted to have benefit. Unlike the logistic regression approach, the ML approach was also accurate in predicting the no-benefit group, which had an observed risk reduction of −0.0313 (3.1% absolute risk increase [95% CI, 1.6%–4.7%]; Figure 2).

The ML approach also had better calibration than the logistic regression model (Table 2; Figure 3). The slope of the calibration line between the predicted ARR (*x* axis on Figure 3) and observed ARR (*y* axis on Figure 3) was 1.06 (95% CI, 0.74–1.32; Table 2).

Predicted treatment effects using the ML approach were not generally proportional to baseline pretreatment cardiovascular risk, with the range of predicted treatment benefit not being proportionately increased with higher baseline risk (Figure 4). When compared with baseline pretreatment cardiovascular risk as defined by Framingham risk scores, results did not meaningfully differ (Figure I in the Data Supplement).

When the discrimination and calibration outcome metrics were computed separately on the SPRINT and on the ACCORD BP datasets, rather than on the combined set alone, results did not meaningfully differ (Tables III and IV and Figures II and III in the Data Supplement).

To aid interpretability of the X-learner approach, we calculated variable importance plots[26] (Figure IV in the Data Supplement) and partial dependence plots[31]

**Table 2. Discrimination and Calibration Metrics for Risk Reduction Predictions (95% CIs)**

| | Machine Learning | Conventional |
|---|---|---|
| Discrimination | | |
| Apparent C-for-benefit (higher is better) | 0.60 (0.58 to 0.63) | 0.54 (0.52 to 0.56) |
| C-for-benefit optimism | 0.00 | 0.03 |
| Corrected C-for-benefit | 0.60 (0.58 to 0.63) | 0.51 (0.49 to 0.53) |
| Apparent decision value RMST, d (higher is better) | 1068.71 (1065.42 to 1072.08) | 1065.47 (1061.04 to 1069.35) |
| Decision value RMST optimism, d | 0.00 | 2.61 |
| Corrected decision value RMST, d | 1068.71 (1065.42 to 1072.08) | 1062.86 (1058.43 to 1066.74) |
| Calibration | | |
| Slope (ideally 1) | 1.06 (0.74 to 1.32) | 0.73 (0.30 to 1.14) |
| Intercept (ideally 0) | −0.00 (−0.01 to 0.00) | 0.00 (−0.01 to 0.01) |

For discrimination, we calculated the C-for-benefit statistic and adjusted for optimism with bootstrap samples. The conventional method required correction because it was both trained and validated on the same dataset, whereas the machine learning method does not because we evaluated out-of-sample predictions. We also calculated the corrected 3-year decision value RMST (in days), and note that a baseline policy of prescribing intensive treatment for participants in the SPRINT trial and standard treatment for those in the ACCORD-BP trial results in 3-year decision RMST of 1061.24 days (95% CI, 1057.37 to 1064.10). For calibration, we recorded the slope and intercept of the calibration curve fitted to quintiles of predicted risk reduction. CIs and optimism were calculated through 250 bootstrap samples. ACCORD BP indicates Action to Control Cardiovascular Risk in Diabetes Blood Pressure Trial; RMST, restricted mean survival time; and SPRINT, Systolic Blood Pressure Intervention Trial.

(Figure V in the Data Supplement), which help identify which variables were most critical in estimating the HTEs (and the directionality of the coefficient in terms of increasing or decreasing benefit) and show how much the ARR estimates change across the range of each variable's values (displaying any nonlinearities). As shown in the Methods in the Data Supplement, the X-learner approach using random forest base learners was superior to alternative ML methods, including an X-learner with a linear base learners, causal forests, random survival forests, and Cox regression with interaction terms (Table VI in the Data Supplement).

## DISCUSSION

We compared a conventional logistic regression approach for modeling HTEs to the X-learner ML approach, applying both approaches to individual participant data from randomized trials of intensive blood pressure treatment. The ML approach revealed correctly that an individual patient's predicted absolute benefit from intensive treatment was not necessarily proportional to their baseline CVD risk. This contradicts prior hypotheses that simply calculating baseline risk will be sufficient to guide therapy,[9] highlighting the clinical importance of HTE risk estimation for making individual treatment effect estimates. We also observed that the ML approach had significantly better discriminative ability, evident in higher C-for-benefit and decision value RMST statistics. The ML approach also partitioned participants into a benefit subgroup that observed a higher empirical ARR than the no-benefit subgroup, whereas the difference between subgroups was more modest for the logistic regression model. Finally, the ML approach had better calibration than the logistic regression model, which over-estimated the ARR attributable to intensive blood pressure treatment.

Our specific modeling approach poses several advances over prior literature, particularly in rigorously comparing the predicted versus observed ARR from therapy, rather than simply calculating discrimination or calibration statistics on overall CVD event rates.[1] This is important because outcome metrics related to overall absolute CVD risk will be driven by major well-known CVD risk factors (eg, age, sex, tobacco smoking), and smaller interaction terms that critically define HTEs (eg, interactions between treatment arm and bio-

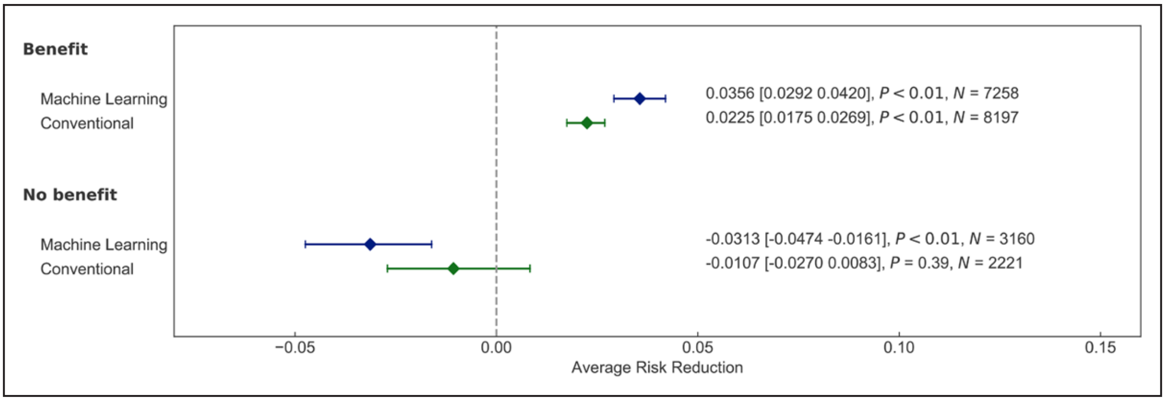

**Figure 2. Participants with uncensored outcomes at 3 y are grouped into predicted buckets of benefit (absolute risk reduction [ARR] >0) and no benefit (ARR ≤0) via machine learning and conventional methods.**

Using these buckets, we bootstrap 95% CIs for the observed ARR in each bucket and calculate corresponding *P* values via the Wald test. The X-learner machine learning approach yielded more discriminative buckets than the conventional logistic regression approach, with the benefit bucket exhibiting higher observed ARR and the no-benefit bucket exhibiting lower observed ARR.

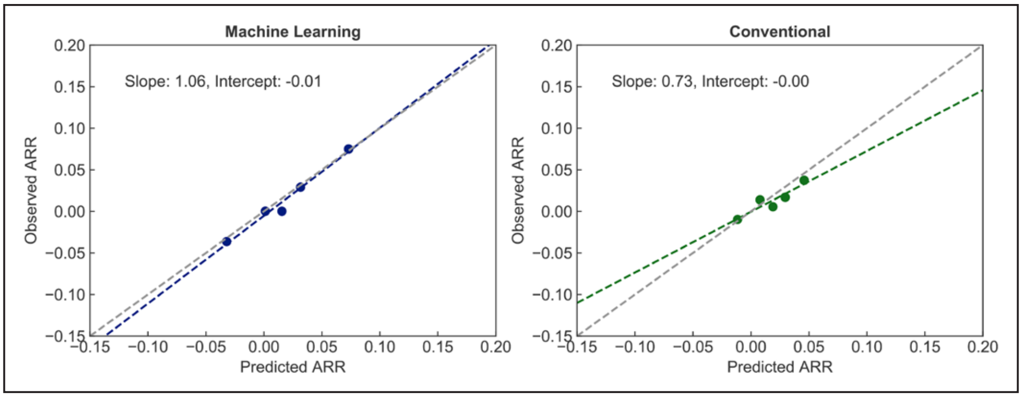

**Figure 3. Calibration plots for predicted absolute risk reduction (ARR) vs observed absolute risk reduction (difference between intensive and standard treatment arms in cardiovascular event rates) using the X-learner machine learning approach and the conventional logistic regression approach, evaluated at quintiles of predicted absolute risk reduction, and using Kaplan-Meier estimates to adjust for censoring.**

markers) will be lost amidst the more dominant overall CVD risk terms when calculating conventional discrimination and calibration metrics. Hence, an HTE model may seem good by traditional C statistics and calibration curves but actually have incorrect HTE terms. We found that direct assessments of observed CVD event rate reductions in held-out samples of data and plotting a calibration curve between predicted and observed risk reduction (rather just predicted versus observed CVD event rates) revealed benefits of the ML approach tested here. Our approach also advanced beyond prior articles modeling the SPRINT and ACCORD BP datasets, in that we found the X-learner approach captured complex interactions and had improved discrimination and calibration as compared to older approaches using logistic regression, which can capture interaction terms but suffer from the limitations of linear modeling, the risk of overfitting, and the risk of false positive results with more complex interactions are tested.[30] By contrast, the X-learner approach used here uses out-of-sample testing to reduce the risk of overfitting and a tree-based approach to enable complex nonlinear interaction terms to be incorporated.[11] We present our decision value RMST as a generic outcome metric for evaluating HTE predictions on clinical trial datasets. In future analyses, our method of evaluation can help determine when the X-learner method (or any other HTE estimation method) would provide improvement over the standard assumption that the average result was generalizable to everyone, indicating in which cases HTE analysis is useful or not.

There are, nevertheless, important limitations to our analysis. First, the ML approach applied here, like many ML approaches, is more complex and difficult to visualize than standard regression modeling. Hence, we communicated the results of our approach by sharing the statistical code to apply to any given patient or other datasets, as well as by plotting variable importance and partial dependence plots to visualize which variables the learner is using and how it is transforming those variables into estimated ARRs. The results cannot be easily captured in a single equation because the X-learner forest modeling approach assembles thousands of decision trees to produce a prediction rather than a single set of

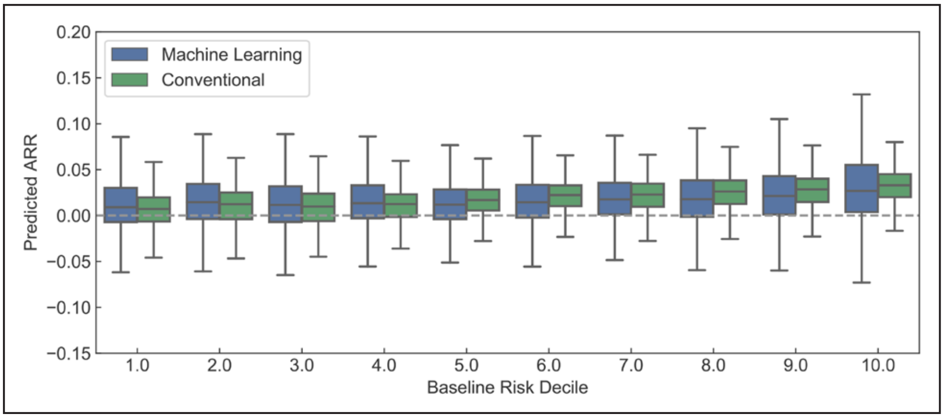

**Figure 4. Predicted risk reduction across deciles of predicted baseline cardiovascular disease risk.**

We used predictions of baseline risk calculated by American College of Cardiology/American Heart Association atherosclerotic cardiovascular disease risk score estimates to group trial participants into baseline risk deciles and compare the median and quartiles of predicted absolute risk reduction (ARR) from intensive blood pressure treatment across each decile. The logistic regression approach generally predicted absolute risk reduction to be proportional to baseline risk, whereas the X-learner machine learning approach predicted wider ranges of risk reduction per decile, not necessarily proportional to baseline risk.

coefficients in a linear model, but the X-learner model can be automatically built into websites and electronic medical records for implementation. Ongoing efforts to improve the communication and transparency of ML methods are needed as part of further research. Additionally, the SPRINT and ACCORD BP datasets remain selective populations of trial participants and are, therefore, not as varied as patients in real clinical settings. Hence, a necessary next step would be to prospectively compare individuals predicted as benefitting or having no benefits from the various models derived here and observe their real-world outcomes to assess the generalizability and importance of our findings.

Nevertheless, in the meantime, our results from this study suggest that the X-learner ML approach may be superior to conventional logistic regression modeling for estimating HTEs, particularly from the perspective of avoiding miscalibration of HTE models. Using methods sensitive to the interactions between treatment arm and covariates, and derived through repeated cross-validation, may be critical to ensuring accurate models of HTEs, and simply assuming that ARR is necessarily proportional to baseline disease risk may not accurately capture true variations in patient risk or benefit from a therapy.

## ARTICLE INFORMATION

Received June 25, 2018; accepted January 16, 2019.

The Data Supplement is available at https://www.ahajournals.org/doi/suppl/10.1161/CIRCOUTCOMES.118.005010.

### Correspondence

Tony Duan, BS, Gates Computer Science, 353 Serra Mall, Stanford University, Stanford, CA 94305. Email tonyduan@cs.stanford.edu

### Affiliations

Department of Computer Science (T.D., P.R., D.L., A.Y.N.), and Center for Primary Care and Outcomes Research and Center for Population Health Sciences, Departments of Medicine and of Health Research and Policy (S.B.), Stanford University, Stanford, CA.

### Sources of Funding

Research reported in this publication was supported by the National Heart, Lung and Blood Institute of the National Institutes of Health under Award Numbers R01HL144555 and R21MD012867. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

### Disclosures

None.

## REFERENCES

1. Patel KK, Arnold SV, Chan PS, Tang Y, Pokharel Y, Jones PG, Spertus JA. Personalizing the intensity of blood pressure control: modeling the heterogeneity of risks and benefits from SPRINT (Systolic Blood Pressure Intervention Trial). *Circ Cardiovasc Qual Outcomes*. 2017;10:e003624. doi: 10.1161/CIRCOUTCOMES.117.003624
2. Burke JF, Hayward RA, Nelson JP, Kent DM. Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. *Circ Cardiovasc Qual Outcomes*. 2014;7:163–169. doi: 10.1161/CIRCOUTCOMES.113.000497
3. Yeh RW, Secemsky EA, Kereiakes DJ, Normand SL, Gershlick AH, Cohen DJ, Spertus JA, Steg PG, Cutlip DE, Rinaldi MJ, Camenzind E, Wijns W, Apruzzese PK, Song Y, Massaro JM, Mauri L; DAPT Study Investigators. Development and validation of a prediction rule for benefit and harm of dual antiplatelet therapy beyond 1 year after percutaneous coronary intervention. *JAMA*. 2016;315:1735–1749. doi: 10.1001/jama.2016.3775
4. Yeboah J, Erbel R, Delaney JC, Nance R, Guo M, Bertoni AG, Budoff M, Moebus S, Jöckel KH, Burke GL, Wong ND, Lehmann N, Herrington DM, Möhlenkamp S, Greenland P. Development of a new diabetes risk prediction tool for incident coronary heart disease events: the Multi-Ethnic Study of Atherosclerosis and the Heinz Nixdorf Recall Study. *Atherosclerosis*. 2014;236:411–417. doi: 10.1016/j.atherosclerosis.2014.07.035
5. Dorresteijn JA, Visseren FL, Ridker PM, Wassink AM, Paynter NP, Steyerberg EW, van der Graaf Y, Cook NR. Estimating treatment effects for individual patients based on the results of randomised clinical trials. *BMJ*. 2011;343:d5888. doi: 10.1136/bmj.d5888
6. Baum A, Scarpa J, Bruzelius E, Tamler R, Basu S, Faghmous J. Targeting weight loss interventions to reduce cardiovascular complications of type 2 diabetes: a machine learning-based post-hoc analysis of heterogeneous treatment effects in the Look AHEAD trial. *Lancet Diabetes Endocrinol*. 2017;5:808–815. doi: 10.1016/S2213-8587(17)30176-6
7. Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials*. 2010;11:85. doi: 10.1186/1745-6215-11-85
8. Perkovic V, Rodgers A. Redefining blood-pressure targets–SPRINT starts the marathon. *N Engl J Med*. 2015;373:2175–2178. doi: 10.1056/NEJMe1513301
9. Phillips RA, Xu J, Peterson LE, Arnold RM, Diamond JA, Schussheim AE. Impact of cardiovascular risk on the relative benefit and harm of intensive treatment of hypertension. *J Am Coll Cardiol*. 2018;71:1601–1610. doi: 10.1016/j.jacc.2018.01.074
10. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Med Res Methodol*. 2006;6:18. doi: 10.1186/1471-2288-6-18
11. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci U S A*. 2016;113:7353–7360. doi: 10.1073/pnas.1510489113
12. Crown WH. Potential application of machine learning in health outcomes research and some statistical cautions. *Value Health*. 2015;18:137–140. doi: 10.1016/j.jval.2014.12.005
13. Gibbons C, Richards S, Valderas JM, Campbell J. Supervised machine learning algorithms can classify open-text feedback of doctor performance with human-level accuracy. *J Med Internet Res*. 2017;19:e65. doi: 10.2196/jmir.6533
14. Deo RC, Nallamothu BK. Learning about machine learning: the promise and pitfalls of big data and the electronic health record. *Circ Cardiovasc Qual Outcomes*. 2016;9:618–620. doi: 10.1161/CIRCOUTCOMES.116.003308
15. Rose S. A machine learning framework for plan payment risk adjustment. *Health Serv Res*. 2016;51:2358–2374. doi: 10.1111/1475-6773.12464
16. Dugan TM, Mukhopadhyay S, Carroll A, Downs S. Machine learning techniques for prediction of early childhood obesity. *Appl Clin Inform*. 2015;6:506–520. doi: 10.4338/ACI-2015-03-RA-0036
17. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Meta-Learners for estimating heterogeneous treatment effects using machine learning. ArXiv.org. 2017. http://arxiv.org/abs/1706.03461. Accessed April 27, 2018.
18. Duan T. Predicting Individual Patient Treatment Effects From Randomized Trial Data: Tonyduan/hte-prediction-rcts. 2019. https://github.com/tonyduan/hte-prediction-rcts. Accessed January 17, 2019.
19. Wright JT Jr, Williamson JD, Whelton PK, Snyder JK, Sink KM, Rocco MV, Reboussin DM, Rahman M, Oparil S, Lewis CE, Kimmel PL, Johnson KC, Goff DC Jr, Fine LJ, Cutler JA, Cushman WC, Cheung AK, Ambrosius WT; SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. *N Engl J Med*. 2015;373:2103–2116. doi: 10.1056/NEJMoa1511939
20. Cushman WC, Evans GW, Byington RP, Goff DC Jr, Grimm RH Jr, Cutler JA, Simons-Morton DG, Basile JN, Corson MA, Probstfield JL, Katz L, Peterson KA, Friedewald WT, Buse JB, Bigger JT, Gerstein HC, Ismail-Beigi F; ACCORD Study Group. Effects of intensive blood-pressure control in type 2 diabetes mellitus. *N Engl J Med*. 2010;362:1575–1585. doi: 10.1056/NEJMoa1001286
21. Wooldridge JM. Inverse probability weighted estimation for general missing data problems. *J Econom*. 2007;141:1281–1301.

22. Vock DM, Wolfson J, Bandyopadhyay S, Adomavicius G, Johnson PE, Vazquez-Benitez G, O'Connor PJ. Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *J Biomed Inform*. 2016;61:119–131. doi: 10.1016/j.jbi.2016.03.009
23. Kent DM, Nelson J, Dahabreh IJ, Rothwell PM, Altman DG, Hayward RA. Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. *Int J Epidemiol*. 2016;45:2075–2088. doi: 10.1093/ije/dyw118
24. D'Agostino RB Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*. 2008;117:743–753. doi: 10.1161/CIRCULATIONAHA.107.699579
25. Goff DC Jr, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, Robinson JG, Schwartz JS, Shero ST, Smith SC Jr, Sorlie P, Stone NJ, Wilson PW, Jordan HS, Nevo L, Wnek J, Anderson JL, Halperin JL, Albert NM, Bozkurt B, Brindis RG, Curtis LH, DeMets D, Hochman JS, Kovacs RJ, Ohman EM, Pressler SJ, Sellke FW, Shen WK, Smith SC Jr, Tomaselli GF; American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association task force on practice guidelines. *Circulation*. 2014;129(25 suppl 2):S49–S73. doi: 10.1161/01.cir.0000437741.48606.98
26. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
27. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. ArXiv.org. 2015. http://arxiv.org/abs/1510.04342. Accessed March 19, 2018.
28. van Klaveren D, Steyerberg EW, Serruys PW, Kent DM. The proposed 'concordance-statistic for benefit' provided a useful metric when modeling heterogeneous treatment effects. *J Clin Epidemiol*. 2018;94:59–68. doi: 10.1016/j.jclinepi.2017.10.021
29. Schuler A, Shah N. General-purpose validation and model selection when estimating individual treatment effects. ArXiv.org. 2018. http://arxiv.org/abs/1804.05146. Accessed May 19, 2018.
30. Basu S, Sussman JB, Rigdon J, Steimle L, Denton BT, Hayward RA. Benefit and harm of intensive blood pressure treatment: derivation and validation of risk models using data from the SPRINT and ACCORD trials. *PLoS Med*. 2017;14:e1002410. doi: 10.1371/journal.pmed.1002410
31. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: visualizing statistical learning with plots of individual conditional Expectation. ArXiv.org. 2013. http://arxiv.org/abs/1309.6392. Accessed April 27, 2018.