

## ORIGINAL ARTICLE

# Ensemble Machine Learning for Personalized Antihypertensive Treatment

Dimitris Bertsimas<sup>1</sup> | Alison Borenstein<sup>2</sup> | Antonin Dauvin<sup>2,3</sup> | Agni Orfanoudaki<sup>\*3</sup>

<sup>1</sup>Sloan School of Management,  
Massachusetts Institute of Technology,  
MA, USA

<sup>2</sup>Operations Research Center and Sloan  
School of Management, Massachusetts  
Institute of Technology, MA, USA

<sup>3</sup>Applied Mathematics, Ecole Polytechnique,  
France

<sup>4</sup>Saïd Business School, University of  
Oxford, UK

## Correspondence

\*Agni Orfanoudaki,  
Saïd Business School,  
University of Oxford,  
Park End St, Oxford, UK, OX1 1HP.  
Email: Agni.Orfanoudaki@sbs.ox.ac.uk

## Summary

Due to its prevalence and association with cardiovascular diseases and premature death, hypertension is a major public health challenge. Proper prevention and management measures are needed to effectively reduce the pervasiveness of the condition. Current clinical guidelines for hypertension provide physicians with general suggestions for first-line pharmacologic treatment, but do not consider patient-specific characteristics. In this study, longitudinal electronic health record data are utilized to develop personalized predictions and prescription recommendations for hypertensive patients. We demonstrate that both binary classification and regression algorithms can be used to accurately predict a patient's future hypertensive status. We then present a prescriptive framework to determine the optimal antihypertensive treatment for a patient using their individual characteristics and clinical condition. Given the observational nature of the data, we address potential confounding through generalized propensity score evaluation and optimal matching. For patients for whom the algorithm recommendation differs from the standard of care, we demonstrate an approximate 15.87% decrease in next blood pressure score based on the predicted outcome under the recommended treatment. An interactive dashboard has been developed to be used by physicians as a clinical support tool.

## KEYWORDS:

Prescriptive Analytics; Hypertension; Machine Learning; Electronic Health Records

## 1 | INTRODUCTION

Hypertension, a medical condition associated with high or elevated blood pressure, affects an estimated 1.13 billion people worldwide (WHO 2019a). Left untreated, hypertension can increase a patient's risk of developing heart, brain, kidney, and other diseases (WHO 2019b). Untreated hypertension also increases the risk of stroke (WHO 2019b), which is considered a major cause of premature deaths, and a prevalent co-morbidity of COVID-19 (Guan et al. 2020).

In 2016, the World Health Organization (WHO) and the United States Centers for Disease Control and Prevention (CDC) launched the Global Hearts Initiative, aimed at a 25% reduction in hypertension prevalence by 2025. To reach this objective, appropriate guidelines are needed for the physician community. In conjunction with the Global Hearts Initiative, the WHO released a set of evidence-based protocols (WHO 2018) that designate who should be treated for hypertension. These protocols recommend first-line treatments from any one of four main classes of antihypertensive medications: angiotensin converting enzyme (ACE) inhibitors, angiotensin receptor blockers (ARB), calcium channel blockers (CCB), and thiazide or thiazide-like diuretics. The guidelines state that proper management of the disease typically requires a combination of medications. These

treatment recommendations are population-wide, with the exception of pregnant women, for whom ACE inhibitors, ARBs, and thiazide or thiazide-like diuretics are not recommended (WHO 2018).

The 2014 Evidence-Based Guidelines for the Management of High Blood Pressure in Adults, developed by the Eighth Joint National Committee (JNC 8), use a similar approach for managing hypertension in adults (James et al. 2014). These guidelines recommend the use of the same four classes of possible initial treatments, with the main objective of attaining and then maintaining a targeted blood pressure value. If the blood pressure goal is not achieved within a month of starting a single treatment, the dosage is typically increased or a second drug is added to the patient’s regimen. If the goal cannot be reached with two drugs, a third drug may be introduced. Sub-population considerations are noted for the black population, for whom a thiazide-like diuretic or CCB is recommended as a first-line treatment (James et al. 2014).

Despite such guidelines being strongly supported by evidence resulting from randomized controlled trials (RCTs), current treatment decision protocols are not highly personalized (Byrd 2016). Both discussed guidelines advise on admissible medications and target blood pressure values. Nevertheless, they state that these recommendations should not be substituted for clinical judgement and the physician’s consideration of the individual characteristics of each patient. In practice, a trial-and-error approach is most commonly adopted, whereby physicians use their experience to prescribe an initial treatment and then refine treatment based on the patient’s health trajectory. However, in the presence of more than 100 drugs, alone or in predefined combinations, finding the correct combination of treatments and dosages is typically a lengthy, iterative process (Epstein 2014). Given the challenges in identifying the optimal treatment strategy at the individual patient level, the prevalence of hypertension is projected to rise rather than to fall in coming years (Burnier & Egan 2019).

An added barrier to the successful management of hypertension is that blood pressure often fluctuates constantly in response to physical and mental activities. It is therefore often characterized by oscillations over short and long-term periods (Chadachan, Ye, Tay, Subramaniam, & Setia 2018). This presents a challenge for both diagnosis and treatment. Given the prevalence of the disease and the severity of its effects on global human health, the development of a personalized approach to treatment of hypertension would assist providers in improved disease management for their patients (Turner, Schwartz, & Boerwinkle 2007). Physicians would greatly benefit from risk calculators that can predict whether the future hypertensive status of patients will improve, deteriorate, or remain stable given their condition, as well as from an interpretable tool that uses patient-specific characteristics to recommend a treatment.

## 1.1 | Literature Review

The ultimate aim of this study is to create models that, given a choice of options, can determine for an individual patient with hypertension their future blood pressure trajectory and identify the best treatment for the disease. Our dataset is comprised of  $n$  observations of the form  $\{(\mathbf{x}_i, y_i, z_i)\}_{i=1}^n$ , where  $\mathbf{x}_i \in \mathbb{R}^p$  are the features of the  $i$ th observation,  $z_i \in [T] = \{1, \dots, T\}$  is the assigned treatment, and  $y_i \in \mathbb{R}$  is the corresponding outcome under the treatment. We use  $y(1), \dots, y(T)$  to denote the  $T$  “potential outcomes” that result from assigning each of the  $T$  respective treatments.

Several approaches have been suggested for solving variants of this problem, both from a causal inference perspective and a machine learning perspective. The Potential Outcomes Framework, also known as the Rubin-Neyman Causal Model, describes how patients are given treatment through a probabilistic assignment mechanism (D. B. Rubin 1974). This framework allows for possible dependence of the mechanism process on potential outcomes (Angrist, Imbens, & Rubin 1996; D. B. Rubin 1990). Under this model, each individual has two potential outcomes,  $y(1)$  and  $y(0)$ , and the causal effect of the treatment is denoted by the difference between the two. The fundamental problem of causal inference, however, is that only one of the two potential outcomes can ever actually be observed. For this reason, causal approaches typically concentrate on determining aggregated causal effects—treatment effects on a population rather than on an individual.

The question of determining heterogeneous treatment effects must be addressed using patient-level characteristics to determine the impact of a treatment for each individual in isolation. In high-dimensional settings where large amounts of data are available, utilizing machine learning for this purpose seems like a natural approach. An approach commonly referred to in literature as “Regress and Compare” involves regressing the outcomes against the covariates of samples who received each treatment separately, predicting the individual’s outcome under each treatment, and recommending the treatment with the best outcome (Stoehlmacher et al. 2004). Several studies have utilized such methods to predict patient-level responses to treatment (Feldstein, Savlov, & Hilf 1978), as well as to compare different treatments (Qian & Murphy 2011). Although intuitive, this approach is subject to prediction errors associated with using only a single method. Also, without necessary adjustment (i.e., through covariate adjustment, inverse probability of treatment weighting, matching, etc.) (Imbens & Rubin 2019) such a

method may suffer from bias, limiting the ability of the results to be viably integrated into clinical practice. Certain studies on heterogeneous treatment effects have made use of various methods to account for bias (Athey & Imbens 2015; Hassanpour & Greiner 2019), though most of them focus on the comparison of treatments in a binary setting.

Alternative machine learning approaches to this problem include extensions of the “Regress and Compare” methodology using a  $k$ -nearest neighbors method ( $k$ -NN) (Bertsimas, Kallus, Weinstein, & Zhuo 2017), as well as tree-based methods that involve recursive partitioning (Kallus 2016), causal trees (Athey & Imbens 2016), causal forests (Wager & Athey 2018), and optimal prescriptive trees (Bertsimas, Dunn, & Mundru 2019). Recently, a machine learning based framework was introduced to identify the best therapy for patients with Coronary Artery Disease (Bertsimas, Orfanoudaki, & Weiner 2020) and was recently extended to address the management of hypertensive patients with COVID-19 (Bertsimas et al. 2021). In the former investigation, a series of regression models were created for each treatment alternative to predict the time from diagnosis to a potential adverse event (TAE) and the therapy with the best expected outcome was selected through a voting mechanism that considered the predicted outcome from each model. This work first demonstrated how machine learning methods could be utilized to create tailored prescriptions for patients with certain diseases. We draw from the work of many of these studies and from Bertsimas et al. (2020) in particular to address the main challenges that persist in the field of personalized medicine, including: counterfactual estimation, confounding and selection bias, and multi-treatment comparison. We follow the framework of Bertsimas et al. (2020) in combining several machine learning methods for prediction to improve outcome estimation confidence. Moreover, we apply this methodology in a concentrated setting to solve the problem of antihypertensive treatment selection.

Several authors have called attention to the need for personalized management of hypertension (Byrd 2016; Mancina & Grassi 2013; Savoia et al. 2017; Turner et al. 2007). Steps toward this goal have included individual patient data meta-analysis to understand the combined effects of self-monitoring and treatment (Tucker et al. 2017) and using randomized trial data to predict absolute risk reduction (ARR) in cardiovascular events from intensive blood pressure therapy (Duan, Rajpurkar, Laird, Ng, & Basu 2019). Application of Electronic Health Records (EHRs) to individualized treatment decision rules remains relatively unexplored.

## 1.2 | Contributions

In this paper, we propose an analytics-based approach for estimating the future blood pressure trajectory of, and for prescribing personalized treatments to, hypertensive patients. Our work is composed of two main parts, which are treated independently. First, we introduce a combination of linear and non-linear classification models that are able to distinguish between hypertensive patients whose blood pressure levels will improve, worsen, or remain stable. In the prescriptive setting, we extend the ensemble approach proposed by Bertsimas et al. (2020). We combine several machine learning model predictions and a voting mechanism to arrive at the optimal treatment at the individual patient level. The main contributions of this paper include:

- **Creation of binary classification models to detect improvement or worsening of hypertension states:** We define a metric that summarizes the blood pressure status of a patient over time. We then develop two series of binary classification models to predict (1) whether a patient’s status will improve significantly and (2) whether a patient’s status will worsen significantly (defined as a change of  $\pm 1.5$  in the BP metric).
- **Design of a quasi-experiment from observational data:** Given the observational nature of the data and, therefore, the risk of selection bias, we simulate a quasi-randomized experiment using matching techniques prior to creating predictive models for each treatment option. We extend matching methods that are typically applied when comparing two treatments to the multiple treatment case. As a result, we ensure that the populations receiving each of the multiple treatments under consideration have similar pre-treatment covariate distributions.
- **Development of regression models to estimate counterfactuals under different treatment regimens:** For each patient, we observe the blood pressure status metric introduced above conditional on the treatment they actually received but not under any of the alternative treatment options. To estimate the unobserved counterfactual outcomes, patients are divided into cohorts based on their inclusion in six mutually exclusive treatment regimens. Separate machine learning models are trained for each of the treatments and validated through out-of-sample testing. After training, counterfactual predictions of the outcome under each treatment are produced for each patient in the test set.
- **Application and extension** of a prescriptive methodology specifically for antihypertensive treatment: We follow the prescription algorithm set forth by (Bertsimas et al. 2020). We tailor it to the condition in which a chronic disease (in

this case, hypertension) is being treated. Whereas in (Bertsimas et al. 2020) a one-off decision is being made to select a surgical intervention or treatment believed to result in the greatest amount of time to a TAE, we aim to choose the best combination of medication and dosage for a continuous treatment regimen. Therefore, departing from the original framework, if majority agreement on the best treatment option is not attained, we defer to the physician as to whether or not the patient's treatment regimen should be altered. We impose a minimum expected improvement threshold over the regimen currently prescribed to the patient to avoid unnecessary changes in prescription.

- **Creation of an online dashboard for clinician support:** An online application is developed as a decision support tool for clinicians. The application allows the physician to visualize, for an individual patient, predicted outcomes under different treatment regimens. The application also includes a measure of agreement between independent models in determining the optimal treatment. The following link may be used to access the application: <http://alisonrb.shinyapps.io/PersonalizedAntihypertension>.

## 2 | METHODS

### 2.1 | Dataset Characteristics

This study utilizes longitudinal EHR data from Boston Medical Center (BMC) patients. BMC is an academic medical center in Boston, MA, that provides pediatric and adult primary care, specialty care, and trauma and emergency services. The raw dataset consists of 145,386 patients, comprising more than 9 million observations corresponding to patient visits between 2000 and 2017. Patients in the raw dataset had at least two records of blood pressure measurements in distinct visits and met at least one of the following inclusion criteria:

- Were administered antihypertensive medications;
- Had EHR observations with ICD9/10 hypertension diagnosis codes;
- Had systolic blood pressure measurements higher than 140 mm Hg or diastolic blood pressure measurements higher than 90 mm Hg (AHA 2018).

Visits to the emergency department were excluded from the sample population. For each patient, the EHR included demographic data, systolic and diastolic blood pressure values, drug prescription descriptions, dosages, and duration. We were also able to retrieve information regarding the patient height, weight, and body mass index (BMI) measurements, history of medical events, and lab value measurements.

### 2.2 | Data Preprocessing

Due to the noise in blood pressure measurements, individual patient observations were aggregated into three-month time interval summaries. The three month window was selected in consultation with a physician and based on a sensitivity analysis on the amount of missing values and blood pressure measurements that we observed in the final dataset. For each summary observation, minimum, median, and maximum systolic and diastolic blood pressure measurements were extracted. Lab values in each summary observation correspond to the median value of all measurements included in the time interval. The final dataset was comprised of patients with three visit summaries such that, for each current observation, we had previous visit information and subsequent visit information. Patients in the final dataset had at least six observations in the current time interval. Furthermore, patients without a nine-month follow-up visit were excluded. Thus, the final cohort ( $N = 19,926$ ) was intended to capture individuals who visit their doctor regularly.

After obtaining a final set of patient observations, additional preprocessing was required to handle outliers and missing values. To account for outlying lab values, upper and lower bounds were imposed based on established reference intervals for laboratory measurements. Missing values were imputed using MedImpute, a recently developed imputation method that leverages the fact that the same patient could be included multiple times in the dataset based on multiple visits (Bertsimas, Orfanoudaki, & Pawlowski 2018). Tables S2-S3 at the Supplemental Material outline the percent of missing values for all features included. Figure S10 summarizes our analysis for the selection of the missing data imputation algorithm. It shows a detailed comparison of five imputation methods considered, in terms of both the imputation accuracy and downstream predictive performance, for the worsening task.

## 2.3 | Blood Pressure Score Definition

To produce a set of covariates for analysis, several features were engineered from the information included in the EHR data. One such feature, referred to herein as blood pressure score, was developed to obtain a de-noised metric that defines a patient's blood pressure status. The American Heart Association (AHA) recognizes five blood pressure categories based on systolic and diastolic blood pressure measurements that range from normal (Class 0) to hypertensive crisis (Class 4) (AHA 2018). The ranges of blood pressure values associated with each category are shown in Table 1.

Each unique record of systolic and diastolic blood pressure was assigned to one of the five blood pressure categories. Therefore, each aggregated observation included the frequency at which the patient had blood pressure recordings in each of the five categories. The final hypertensive status of a patient was defined by the blood pressure score metric given by Eq (1). Thus, if for a given patient 10 measurements were available out of which three belonged in the "Normal" category, two in the "Elevated," and five in "Hypertension Stage 1," then the corresponding score would be:  $1.2 = (0 \cdot 3 + 1 \cdot 2 + 2 \cdot 5)/10$ .

Based on this definition, a patient that transitions from one category to the next is associated with a score increase of one point. After defining this metric, for each patient, we could then utilize the previous, current, and next blood pressure score, as well as the blood pressure category frequencies, as continuous features. These features acted as summary functions that could encapsulate the state of the patient over time with a low-dimensional representation.

$$\text{score} = 0f_0 + 1f_1 + 2f_2 + 3f_3 + 4f_4 \quad (1)$$

We recommend the use of this metric to capture in a continuous and more holistic fashion the hypertensive status of the patient. The proposed blood pressure score carries medical intuition as it is based on the widely accepted hypertensive status categories that are established by the AHA. Thus, it carries similar benefits and limitations to those associated with the AHA definition. According to the latter, a diastolic blood pressure measurement of 79 is considered elevated while a measurement of 81 is classified as Hypertensive Stage 1. Though in some cases this dichotomy is counter-intuitive, the AHA has advocated for this definition to assist clinicians in navigating the noise associated with blood pressure measurements. We believe that the score introduced in this manuscript remedies some of these limitations as it requires at least six measurements to be defined (six for the systolic blood pressure and six for the diastolic blood pressure) and consequently it is more robust to outliers.

In addition, each observation corresponds to a three-month time interval. As a result, though each measurement in a single period carries the same weight, we distinguish between more recent and less recent visits in a long term horizon. In other words, measurements that are apart for more than three months are not included in the same metric.

The choice of weights reflects our intention to highlight abnormally high measurements in the score. The metric penalizes proportionally to the category risk the relative frequency of measurements in a linear fashion. Thus, this metric can be viewed as a weighted average of the AHA categories, adjusted by the severity of the measurement. We conducted sensitivity analysis to set the corresponding thresholds of change for the predictive and prescriptive problem that would signify an improvement and worsening of a patient's status. Further details regarding these results are available in Section 2.5 and the Supplemental Material (Figures S6-S9).

## 2.4 | Feature Engineering

In total, 90 variables were compiled for analysis. All continuous and categorical variables that were included are summarized in Tables S2-S3 of the Supplemental Material, respectively. Predictor variables consisted of those directly recorded in the EHR, as well as those that were derived from the raw data. Demographic variables included patient gender, ethnicity, religion, and native language. For each observation, we utilized the current visit summary age, height, weight, BMI, minimum, maximum, and median blood pressure measurements, as well as the median of the lab values (Table 4). Lab-related covariates, for which more than 30% of the values of the aggregated observations, were missing were excluded. The patient's history of cardiovascular disease and type II diabetes was also included.

Each observation summarized the patient's blood pressure trajectory through inclusion of previous measurements, such as previous visit blood pressure values and category frequencies. Furthermore, categorical variables capturing previous dosage levels of medications were extracted from patient records. For previous treatment variables, the following treatment types were considered: ACE inhibitors, blockers (alpha, beta, and/or calcium channel), angiotensin II inhibitors, diuretics, others, and none.

For our first predictive task, determining the trajectory of a patient's blood pressure status, our dataset included only the last visit for every patient in the system. For the task of determining optimal treatment, we created separate datasets for each

treatment-dosage option. Prior to the separation of the full dataset into cohorts based on treatment type, current treatment-dosage categories were encoded as binary variables. The following types were included: low dosage of a blockers and diuretics combination, high dosage of a blockers and diuretics combination, low dosage of blockers, high dosage of blockers, and low dosage of diuretics. Note that blockers may include alpha, beta, and/or calcium channel blockers treatments. Table S1 at the Supplemental Material provides further information regarding the classification of the patient daily prescriptions into high or low treatment regimens. These treatment-dosage categories were chosen with the intention of capturing a large portion of the patient population while also limiting the number of potential treatment-dosage options for comparative analysis.

## 2.5 | Outcomes of Interest

For all predictive modeling tasks, the outcome variable of interest was the patient's next blood pressure score, as defined by Eq (1), using measurements associated with the patient's next three-month aggregated visit summary.

For the task of predicting blood pressure status trajectory, the outcome of interest was transformed to enable application of binary classification models. Ultimately, we considered two outcome variables: (1) Improvement Task: Reduction of the blood pressure score greater than the threshold  $\delta$  of 1.5 between two visits signifies improvement; (2) Worsening Task: Increase of the blood pressure score greater than  $\delta = 1.5$  over two consecutive visits is considered deterioration. The choice of the threshold was based on a sensitivity analysis (Figures S6-S9). We set the  $\delta$  threshold such that resulted in at least a 0.9 AUC ROC for the improvement task. Note that as we increase the value of  $\delta$ , the AUC ROC is improving but the AUC precision recall (PR) is decreasing. Thus, the choice of the threshold depends on the importance and the balance that the researcher wishes to achieve between the two metrics. The choice of  $\delta = 1.5$  is also validated from a clinical perspective, as it considers only significant changes in the blood pressure status, being more robust to noisy measurements.

## 2.6 | Treatment Options

In order to compare multiple treatments, separate regression models were trained for each treatment option. The final cohort of 12,561 observations was divided into five mutually exclusive subsets based on the current treatment-dosage regimen. Percentages of the final cohort belonging to each type of treatment are listed in Table 2. We limited the sample population to the five categories of treatment-dosage regimen which corresponds to 63.04% of the original sample ( $N_{orig} = 19,926$ ). Blockers and diuretics were selected as the most commonly prescribed classes of antihypertensive medication. The former class includes calcium channel, alpha, and beta blockers prescriptions. Dosage was also included since it has been shown to be highly significant in predicting blood pressure reduction effects (Law, Morris, & Wald 2009). The combination of both monotherapy classes and dosage level allows us to detect the true differences between the efficacy of the treatment regimens. It also allows us to identify the therapy which would lead to an improvement of a patient's hypertensive status with the lowest amount of medication possible.

## 2.7 | Debiasing Approach for Determining Optimal Treatments

As stated, the primary objective of this study was to determine the optimal treatment for a patient, given his or her individual characteristics. To accomplish this, each treatment option had to be compared to assess which would result in the most favorable outcome for the patient.

Comparison of treatments is typically achieved through RCTs, which represent the gold standard for determining treatment effects (Pearl 2009). In a typical RCT, patients are randomly assigned to a treatment group and a control group. Each unit in the trial,  $x_i$ , has two potential outcomes:  $Y_0(x_i)$  is the potential outcome had the unit not been treated, and  $Y_1(x_i)$  is the potential outcome had the unit been treated. From these values, we can estimate the conditional average treatment effect (CATE) for unit  $i$  according to Eq (2) which, mathematically, corresponds to the difference in expectations of outcomes under treatment and control.

$$CATE(x_i) = \mathbb{E}_{Y_1 \sim (y_1|x_i)}[Y_1 | x_i] - \mathbb{E}_{Y_0 \sim (y_0|x_i)}[Y_0 | x_i] \quad (2)$$

However, in reality, only one of these two values can be observed, which is the fundamental problem of causal inference. Therefore, in order to estimate the CATE for an individual, we must impute the unobserved counterfactual outcome and compare it with the observed factual outcome. The strength of RCTs stem from the treatment assignment mechanism being random.

In many cases however, and especially in the context of medicine, random assignment may be either prohibitively expensive, unethical, or infeasible (Pearl [2009]). For this reason, observational data is often used to estimate causal effects.

With respect to our study, which utilizes observational data, it is reasonable to assume that patients were not likely to have received a random assignment of treatments. For this reason, inferring causality from the data can be challenging due to the presence of confounding variables and selection bias. Specifically, when attempting to make causal inferences through comparison of groups that are different not only in terms of treatment but also with respect to predictors that are related to both the treatment and the outcome, we can be misled by the results (Gelman & Hill [2018]). Matching studies are designed to minimize imbalances on measured preintervention characteristics, thereby reducing bias in estimates of treatment effects.

Several methods have been proposed to adjust for bias; the most commonly considered are covariate adjustment, inverse probability of treatment weighting (IPTW), stratification, and matching, each of which are detailed in (Austin [2011]). In this study, we utilized generalized propensity scores, which were developed as an extension of the propensity score measure for binary treatment (Imai & van Dyk [2004]; Imbens [2000]; Rosenbaum & Rubin [1983]), to confirm common support existed among all treatment options. We then used matching methods and extended them to the multi-treatment case.

Before matching, it is worthwhile to confirm that there is substantial overlap of the propensity score distributions among the different treatment groups—this is known as the notion of common support. Identification of common support is a critical aspect of the strong ignorability assumption for identifying causal effects from observational data (Rosenbaum & Rubin [1983]). We verified that common support existed through the use of generalized propensity score (GPS) (Imbens [2000]). After determining pre-treatment covariates believed to affect both treatment assignment and the outcome (possible confounders), we estimated the GPS for each patient observation using both multinomial logistic regression and the SuperLearner framework (Van der Laan, Polley, & Hubbard [2007]; Zhou, Tong, Li, & Thomas [2020]). Through this effort, we ascertained that the data was composed of units that are eligible to receive all of the treatments.

By randomly assigning units to receive or not receive a treatment, one can ensure that there are no systematic differences between treatment groups before the treatment is assigned. In observational studies, random treatment assignment is not possible and there are several variables, commonly referred to as confounding variables, that may affect both the treatment assignment and the outcome (Gelman & Hill [2018]). For example, a patient's age might dictate which treatment options they are eligible to receive and might also affect how that patient responds to the received treatment. Thus, it is possible that the age distribution of the population receiving one treatment may differ from the age distribution of the population receiving a different treatment. Such differences in covariate distributions give rise to a problem known as selection bias which, if not properly accounted for, can lead to biased estimates of the effect of a treatment (Gelman & Hill [2018]). In these instances, it is imperative to separate the causal effect of the treatment from the effect of preexisting differences between patients belonging to different treatment groups.

One common approach to adjust for selection bias is to use GPS-reweighting to account for confoundedness in the treatment selection that we observe at the standard of care. This approach can be quite accurate in estimating the true CATE and removing potential observational bias from the data. The literature, though, suggests the exclusion of units from the analysis with extremely high or low GPS as they may lead to poor balance or large variance in the resulting population (Busso, DiNardo, & McCrary [2014]). As a result, depending on the treatment assignment mechanism observed in the population of interest, a GPS-reweighted loss function can lead to a limited sample size in the resulting processed dataset.

In our effort to control for selection bias, we utilized matching techniques. The overall goal of matching is to replicate a RCT by forming groups, without using the outcome, for which the observed covariate distributions are alike (balanced). Thus, we aimed to find populations for each treatment for which the pre-treatment covariate distributions were similar. Achieving such balance allows the initial attribution of the observed difference in outcomes to be an effect of the treatments rather than the differences in covariates. The idea is that for each individual receiving any one treatment, we wish to observe a similar individual who has received each of the other treatments. In the case of a single treatment and control, matching methods developed by Cochran and Rubin are often utilized for this purpose (Cochran & Rubin [1973]; D. B. Rubin [1973a] [1973b]).

There are a number of algorithms, including nearest neighbor matching and optimal matching, that have been developed for matching in the single treatment-control case. While nearest neighbor matching is more commonly used, optimal matching has been shown in many instances to achieve better balance on the confounders (Harder, Stuart, & Anthony [2010]). In this study, we used a form of optimal matching known as cardinality matching (Zubizarreta, Paredes, & Rosenbaum [2014]). With cardinality matching, a linear integer programming problem is solved, where the objective is to maximize the size of the matched sample subject to constraints on covariate balance. Specifically, we sought to minimize the differences in means between the pre-treatment covariates across all pairwise comparisons of treatment groups while also maximizing the number of matched units.

Matching methods have generally been developed for the binary treatment case; when considering more than two treatments, many of these methods become computationally intractable. Following the work of Silber et al. (Silber et al. 2014) and Bennett et al. (Bennett, Vielma, & Zubizarreta 2020) to overcome this limitation, we matched individuals from each treatment group to the treatment group with the fewest number of observations, which we considered as our representative sample. Matching was implemented with the `designmatch` package in R (Bennett et al. 2020). The `designmatch` function for optimal cardinality matching in observational studies was used.

Statistical literature has warned that regression analysis cannot reliably adjust for differences in observed covariates if substantial differences in the covariate distributions exist (Cochran 1957 1965). Post-matching, we assessed the balance between the treatment groups using pairwise standardized absolute mean differences, as suggested by (D. Rubin 2001). Typically, different groups are considered balanced if the standardized absolute mean differences between the groups are less than 0.25 (Cohen 2009; D. Rubin 2001). Once this level of balance is achieved, outcome analysis can be performed.

## 2.8 | Predictive Models

### 2.8.1 | Classification models for blood pressure status trajectory.

For the binary prediction problem, our goal was to create accurate classification models for both the improvement and the worsening task. The population was randomly divided into 80% training and 20% testing sets, ensuring that the ratio of prevalence for each outcome remained the same between the two (see Table 3). We developed and tested prediction algorithms for hypertensive status trajectory using five commonly used machine learning approaches (Dauvin et al. 2019): multivariate logistic regression (MLR) (Hastie, Friedman, & Tibshirani 2017), classification and regression trees (CART) (Breiman 2017), random forests (RF) (Breiman 2001), gradient boosting machines (GBM) (Chen & Guestrin 2016), and optimal classification trees (OCT) (Bertsimas & Dunn 2017 2019). We included a combination of linear and tree-based methods, along with ensemble algorithms. A brief summary of the methods is provided in Table S4.

All predictive models were evaluated based on their ability to discriminate between the two outcomes of interest. We report bootstrapped results for the c-statistic on the testing set with the corresponding confidence interval (CI) and standard deviation (SD). This metric, also known as the Area Under the Receiver Operating Characteristic Curve (AUC), measures the ability of a model to discriminate between the outcomes of interest, incorporating both the sensitivity and the specificity of the model. It has been used as a measure of model success in multiple prior risk-scoring development efforts (Hanley & Mcneil 1982).

### 2.8.2 | Regression models for treatment recommendations.

In order to arrive at a medication recommendation, we had to infer the patient's response to each of the treatment-dosage options. For this task, we trained a separate model for each treatment-dosage combination. Thus, rather than adding an indicator of treatment-dosage type as a feature, which would not allow us to ensure the same baseline risk for each regimen, we created a suite of models for each treatment-dosage separately.

We developed predictive models for the unmatched version of the dataset and the matched version of the dataset. We present our findings for both. In both cases, the final cohort for each treatment-dosage option was used to train models that predict the next blood pressure score. For each treatment, we further divided the populations into 75% training and 25% testing sets. We also performed bootstrapping of the results across five random splits of the data to obtain confidence intervals for the evaluation metrics. If there were multiple observations for a single patient, we restricted all observations to either be in only the training set or only the testing set.

We trained a variety of regression models for each treatment-dosage combination, consisting of linear and non-linear methods to learn relationships between the outcome and the covariates, as well as interactions between covariates.

Models leveraged for the regression task included  $l_1$  regularized regression (LASSO) (Hastie et al. 2017), support vector regression (SVR) (Drucker, Burges, Kaufman, Smola, & Vapnik 1997), classification and regression trees (CART) (Breiman 2017), random forests (RF) (Breiman 2001), gradient boosting machines (GBM) (Chen & Guestrin 2016), optimal regression trees (ORT), and optimal prescriptive trees (OPT) (see Table S5) (Bertsimas & Dunn 2017 2019). While the majority of the machine learning methods used are applied for a wide variety of tasks, OPTs were designed with personalized decision making in mind (Bertsimas et al. 2019); thus, their application is highly relevant to this problem.

The OPTs method differs from the other algorithms in that it is a prescriptive method, whereas the others are predictive methods. OPTs utilize joint learning, whereby the entire sample is used for training purposes to predict counterfactuals and to assign the optimal treatment. The objective function introduced in the OPT framework is one that balances optimality and

accuracy through the use of a prescription factor,  $\mu$ , which controls the trade-off between prescription quality and predictive accuracy. In our work, we used a prescription factor of 0.5. The tree-based output of OPTs results in all observations in the same leaf being assigned to the same optimal treatment group. For the predictive models, we use a regress and compare approach to determine the best treatment according to each model. To align the output from OPTs to those from the other predictive models, we extracted the predictions for patients in each treatment group and used the prediction for the actual treatment received to evaluate performance.

Cross-validation was used to select hyperparameters for each of the models. Out-of-sample  $R^2$  and  $MAE$  were used to evaluate model performance. Subsequent to training, for every patient in the combined test set, prediction outcomes were obtained from each of the trained models to utilize in the prescription algorithm.

## 2.9 | Prescriptive Component

After using models trained on each separate treatment-dosage cohort to predict the counterfactual estimations of the next blood pressure score, we obtained a matrix of model/treatment-dosage combinations for each patient in the test set. An example of such a matrix is shown in Table 5. Our algorithmic framework for medication prescription is similar to that described by Bertsimas et al. (Bertsimas et al. 2020), except that we only change the medication regimen if the majority of the models agree that an alternate treatment will result in a better outcome than the current regimen with an expected improvement of at least 0.2 in the blood pressure score. Our methodology is summarized as follows:

1. Using the matrix of counterfactual estimates, we derive which treatment-dosage each regression model selects as the best, based on the lowest expected outcome (next blood pressure score).
2. We also ascertain which of the treatment-dosage options is most frequently chosen as the best among the algorithms considered.
3. Following the approach of Bertsimas et al. (2020), if the majority of the regression models agree that a certain treatment-dosage will result in the lowest outcome, we average the predictions from the models in agreement to obtain a final prediction for the next blood pressure score under the chosen treatment-dosage.
4. If there is not majority agreement among the regression models or if the expected improvement from the change in dosage or medication is less than or equal to 0.2, the algorithm defers to the physician to determine if the patient should remain on their current regimen or if their medication regimen should be refined. The selection of the 0.2 threshold was based on a sensitivity analysis on the treatment allocation distribution and efficacy of the algorithm.

In Table 5, we will walk through the prescription algorithm to demonstrate a full example. For this particular patient, six of the seven models (CART, GBM, OPT, ORT, RF, SVM) agree that a high dosage of blockers with diuretics is the regimen that will result in the lowest blood pressure score for the next three-month period, while one model (LASSO) predicts a low dosage of diuretics to be the best treatment option. Given that model agreement is 85.7% (6 of 7 models agree), a high dosage of blockers with diuretics is selected as the treatment recommendation and the final prediction for the patient's next blood pressure score is calculated by averaging the models that agree, resulting in a final prediction of 1.575 for the next blood pressure score. Supposing the patient's current blood pressure score is 2.5, for example, this implies that the treatment recommendation will result in the patient's AHA category lowering by one class, which is above the 0.2 threshold ( $0.925 \geq 0.2$ ). A blood pressure score of 2.5 indicates that a majority of the patient's blood pressure readings fall within the hypertension stage 1 and hypertension stage 2 categories, and a decrease in score to 1.575 would result in a much lower frequency of hypertension stage 2 readings and, correspondingly, a higher frequency of readings in a category associated with lower systolic and diastolic blood pressure measurements.

## 3 | RESULTS

In this section, we present results from both prediction components as well as the prescriptive component of this study. First, we share the results of the models we trained to predict a patient's hypertensive state trajectory. Then, with respect to our models for treatment-dosage recommendations, we present prediction evaluation metrics for each model/treatment-dosage combination associated with both the full sample (unmatched data) and the matched sample, as well as prescription evaluation metrics for

both samples. We then present the remainder of the results using the matched data, as we believe that this approach is more robust to potential biases in outcomes.

### 3.1 | Binary Prediction Results

GBM and MLR demonstrated an edge in their out-of-sample AUC performance for both tasks. In particular, for the worsening task, the GBM (77.6%, 95% CI, 76.6-78.7) and MLR (78.1%, 95% CI, 76.6-78.7) outperformed the RF (76.8%, 95% CI, 75.6-77.9), OCT (72.0%, 95% CI, 70.3-73.7), and CART (52.9%, 95% CI, 52.5-53.4) methods. Similar ranking of performance was observed for the improvement task but with higher overall results in terms of AUC and smaller discrepancies between the models; GBM (89.9%, 95% CI, 89.2-90.6) and MLR (90.0%, 95% CI, 89.5-90.6), RF (89.8%, 95% CI, 89.2-90.4), OCT (87.9%, 95% CI, 87.3-88.6), and CART (57.9%, 95% CI, 57.3-58.5). Notice in Figures S3-4 that the confidence intervals of the higher performing methods are tighter compared to the rest of the algorithms. The results reveal that interpretable methods such as MLR and OCT achieve comparable performance compared to less transparent algorithms like GBM or RF. Figures S1-S2 summarize the predictive performance of all models based on the AUC for the improvement and the worsening task, respectively. Table 6 presents the average out-of-sample AUC across five splits between the training and the testing set, along with the respective 95% confidence intervals. Table S10 compares the performance of the models by displaying the degree of significance in the performance comparison between the models. In Figure S5 of the Supplemental Material, we have included the resulting feature importance graphs for MLR, RF and GBM, highlighting the most predictive risk factors for both classification tasks.

### 3.2 | Continuous Predictions under Different Treatments

#### 3.2.1 | Matching Results.

In this study, we weighed the need to retain a large enough dataset for each treatment cohort to properly train machine learning models with the need to balance pre-treatment covariates for each of the treatment options. Iterating through the procedure described in the debiasing approach section, we were able to retain 97% of the original final dataset (resulting in  $N_{matched} = 12,178$ ) while achieving pairwise balance across all treatments below 0.25 for all pre-treatment covariates, as shown in Figure 1. Furthermore, 90% of the pre-treatment variables after matching had a standardized absolute mean difference below 0.10. In Supplementary Tables S6-S7, we summarize the pre-treatment variables stratified by treatment before and after the matching procedure.

We conducted an additional analysis to compare matching against a GPS-loss function weighted approach (Li 2019). In the presence of multiple treatments, we computed each individual's weight as the ratio of the probability receiving the target treatment versus the probability of receiving the treatment that is observed at the standard of care (Li, Morgan, & Zaslavsky 2018). The resulting datasets had very similar characteristics to the proposed approach and the distribution of the best predicted treatments was not substantially different from the one based on covariate matching (see Supplementary Table S13).

#### 3.2.2 | Predictive Accuracy of Regression Models

We used  $R^2$  and  $MAE$  metrics to evaluate the out-of-sample performance of the separate treatment models. The  $R^2$ , or coefficient of determination, metric represents the proportional improvement in prediction accuracy compared to a model that predicts the outcome for all samples to be the mean value of all samples in the training set, while the  $MAE$  metric measures the average absolute magnitude of the errors in the predictions. In the full sample dataset, the average  $R^2$  values ranged from 0.24 to 0.45, depending on the treatment subset and model type. The mean  $MAE$  ranged from 0.44 to 0.57. The out-of-sample performance for the blockers models with high dosage were superior to all other treatment types for the LASSO, SVM, CART, RF and GBM algorithms. In terms of predictive accuracy, RF and GBM models outperformed the others. The evaluation metrics for the full, unmatched sample are presented in Tables 7 and S8. We display the mean value as well as the 95% confidence interval (CI) resulting from evaluation on five random splits of the data.

Again for the matched sample, we utilized  $R^2$  and  $MAE$  as performance evaluation metrics. The  $R^2$  values ranged from 0.20 to 0.45, and the  $MAE$  values ranged from 0.46 to 0.59. The predictive accuracy of the individual treatment models is very similar between the unmatched and matched datasets. Also similar to the full sample results, the RF, and GBM models display the highest predictive accuracy in general. Evaluation metrics for the matched sample are presented in Tables 8 and S9.

### 3.3 | Prescription Algorithm Results

In (Bertsimas et al. 2020), the authors propose several methods for prescriptive algorithm evaluation. We adopted a metric similar to their “prediction accuracy of TAE” metric, where we computed the  $R^2$  with patients for whom the prescription algorithm recommendation matched the treatment that the patient actually received. This evaluation procedure was used because we can only ever realize one factual outcome for each observation, and the other four counterfactual outcomes must be imputed. Thus, the evaluation metric considers only instances where we can compare the ground truth to a prediction. This approach, though limited, enables us to infer the strength of our prescription algorithm with respect to recommendation accuracy. The final  $R^2$  obtained from this procedure was 0.58 [95% CI, 0.55-0.61] using the unmatched dataset and 0.57 [95% CI, 0.55-0.59] using the matched dataset.

We also adopted the Prescription Effectiveness (PE) and Prescription Robustness (PR) metrics introduced by (Bertsimas et al. 2020). The goal of these metrics is to consider different predictions of the outcome with respect to a multitude of ground truths. The baseline ground truth corresponds to the outcome that was actually observed in the data and, thus, provides us with the next blood pressure score that was associated with the treatment that was prescribed by the physician. Alternative ground truths refer to predictions of the patient’s next blood pressure score associated with each of the regression models. With the PE metric, we consider each regression model in isolation and compare the effectiveness of the predicted prescription outcome relative to the baseline ground truth outcome observed in the data. The PR metric is determined by generating alternative ground truths assuming that each regression model knew the outcome at the standard of care and comparing the effectiveness of each of the other regression models against that outcome. In this way, we can evaluate the robustness of the treatment effect estimation under different ground truths. To make these metrics more interpretable to our outcome of interest, we transform the raw outcome (which corresponds to the decrease in the next blood pressure score for each model relative to each ground truth) into a percentage decrease in next blood pressure score. We present these results in Table 9. Further details regarding the definition and interpretation of PE and PR are presented in the Supplemental Material.

Table 9 presents the expected decrease in a patient’s next blood pressure score when comparing the current treatment allocation plan with our prescription algorithm plan using different estimation models as the ground truth. Relative to the current allocation plan, our prescription algorithm allocation plan represents a 15.87% expected decrease in next blood pressure score. By observing the first column of Table 9, we find that RF, GBM and ORT are the most optimistic models, estimating an approximate 18% decrease over the baseline ground truth. CART, on the other hand, is the most conservative, estimating an approximate 10.5% decrease in blood pressure score relative to the baseline. The remaining regression models estimate fairly similar decreases, between 16% and 18%. Across all models, we demonstrate an expected benefit from the algorithm allocation relative to the current allocation. Our PR metrics results show consistency in estimations across all models and alternate ground truths considered. Contrary to the PE metric, RF is the least optimistic model across all ground truths and, in terms of PR, OPT is the most optimistic. We observe that OPT estimates higher efficacy for the proposed treatment allocation than our prescription algorithm. However, we find that the final  $R^2$  values of OPT are substantially lower than the final  $R^2$  values from our algorithm, which combines estimations from multiple models. We can see from these metrics that some regression methods overestimate the expected outcome while others underestimate it, and thus that the strongest results are obtained from averaging models that agree on the optimal treatment decision.

### 3.4 | Variable Importance

We observed a high level of agreement between the different regression models as to which of the variables were identified as important to the prediction task. Current and previous blood pressure score were among the variables with highest importance for all treatments across all models, suggesting the usefulness of such a summary function. Variables summarizing the frequency of a patient’s blood pressure measurements falling into each of the AHA categories were identified as important across all models as well. Median value measures of systolic and diastolic blood pressure were also highly predictive of the outcome for many models.

Modeling results also revealed several variables whose importance was particular to a treatment-dosage type. Depending on the treatment, visit summary lab values were indicated by several of the regression methods as important variables. For example, hemoglobin and cholesterol related lab values were identified as important by the models trained on the blockers & diuretics patient subset for both low and high dosage. Additional lab measurements that were frequently identified as significant included creatinine, iron, and triglycerides. Age was identified as an important factor for all therapies considered. Furthermore, duration of drug prescription was important for the diuretics models. As demonstrated by these examples, factors identified as important

for a particular low dosage therapy were typically also identified by the combination of therapies or for therapies with higher dosage. Feature importance was consistent between the unmatched and matches samples for each treatment type.

### 3.5 | Model Agreement

For the prescription algorithm, we recorded the level of agreement among the regression models. In Table 10, we report the percentage of agreement of the machine learning models by treatment type and overall. The results reflect the average agreement to the entire testing set across all five random splits of the data and not to an individual sample patient. Of the testing set observations, majority agreement among the seven models is achieved in approximately 49.32% of instances. Higher level of agreement corresponds to greater confidence in the prescription recommendation, whereas lower level of agreement indicates lower confidence. For this reason, below a majority threshold (corresponding to three or fewer models out of seven in agreement), the prescription algorithm defers to the physician for further evaluation as to whether or not the current regimen should be altered.

### 3.6 | Treatment Recommendation Distributions

In Table 11, we present the results of the prescription algorithm in the context of distribution of final treatment recommendations. In cases where majority agreement is not met or the expected improvement in the blood pressure score is less than 0.2, we assume that the patient will remain on his or her current line of therapy. By looking at the reallocation percentages, the results suggest that, in many cases, lower dosage regimes may be preferable to combination therapies or higher dosages. Furthermore, the combination of blockers and diuretics in high dosage appear to be the most frequently chosen treatment option. High dosage on blockers exclusively, on the other hand, is the least frequent monotherapy option. Among the low-dosage therapy options, diuretics is the most commonly suggested option by the algorithm.

These findings are generally in agreement with current treatment protocols and guidelines, which favor thiazide-like diuretics as a first-line therapy and, generally, do not recommend solely Beta blockers for initial treatment due to complications associated with cardiovascular death, myocardial infarction, or stroke (WHO 2018).

### 3.7 | Improvement Over Standard of Care

Our algorithm agrees with the actual treatment the patient received between 78% and 82% of the time, depending on the split of the data. For patients for whom the algorithm disagrees with the standard of care on an optimal treatment, we evaluate the potential improvement in patient outcomes based on the expected blood pressure score in the subsequent visit.

Table 12 displays the mean actual outcome, mean predicted best outcome, and the corresponding percentage decrease in patients' next blood pressure scores. Patients are grouped by their current treatment regimen. Patients predicted to experience the greatest decrease in their next blood pressure score are those that are currently taking a high dosage of blockers and diuretics. For these patients, the best potential outcome represents a decrease in score of approximately 17%.

We provide a concrete example below to illustrate the potential benefit that is suggested by these results. Let us suppose that patient A has a current blood pressure score of 1.70, based on the following calculation as dictated by Eq 1:

$$\text{score} = 0 \cdot 0\% + 1 \cdot 50\% + 2 \cdot 30\% + 3 \cdot 20\% + 4 \cdot 0\% = 1.70 \quad (3)$$

Based on this score, a majority (50%) of patient A's blood pressure measurements fall into the Elevated Blood Pressure category, while 30% and 20% of the measurements belong to the Stage 1 Hypertension and Stage 2 Hypertension categories, respectively.

An example calculation corresponding to a 18.8% decrease in blood pressure score, hypothetically switching the patient from the current regimen to the predicted best treatment option, might be:

$$\text{score} = 0 \cdot 0\% + 1 \cdot 62\% + 2 \cdot 38\% + 3 \cdot 0\% + 4 \cdot 0\% = 1.38 \quad (4)$$

In this hypothetical instance, the frequency of measurements shifts from higher-risk categories to lower-risk categories. As demonstrated by this example, the prescription algorithm results suggest considerable improvement ( $\geq 0.2$ ) over current standard of care. Improvement among specific sub-groups is detailed in Section 3.8.

### 3.8 | Subgroup Analysis

We expand the results of our study through subgroup analysis, whereby we investigate final out-of-sample  $R^2$ , and potential decrease in next blood pressure score for each subgroup shown in Table 13.

#### Ethnicity

Through subgroup analysis based on patient ethnicity, the highest out-of-sample  $R^2$  values were observed for patients whose ethnicity is Caucasian, for which a mean out-of-sample accuracy of 71.0% was achieved when comparing patients for whom recommended treatment matched actual treatment. The accuracy of the algorithm was higher in this sub-population due to the higher number of samples that belong in this category. The performance of the models would improve across underrepresented ethnicities should more observations be included. Nevertheless, we notice that patients predicted to experience the greatest decrease in their next blood pressure score are those whose ethnicity is Black, Hispanic, or Other.

#### Age

We also grouped patients into four different age buckets to investigate model performance on different age populations. We found that the out-of-sample performance was highest for patients whose age is between 40 and 60. The latter sub-population represents the majority of the observations found in the data. We also found that the patients in the youngest [18,40) and the eldest [80-110) age group had the greatest potential benefit in terms of decrease in blood pressure score.

#### Gender

Outcomes in terms of potential blood pressure score decrease are similar. Minor differences in the algorithm's efficacy are recorded between the two groups. This finding is not surprising, given that gender was not indicated as an important factor by the regression models.

#### Language

We present a subgroup analysis based on the Language of the sample population. The algorithm is associated with the highest accuracy in the English ( $R^2 = 0.65$ ) and Spanish ( $R^2 = 0.55$ ) speaking cohorts as they consist the groups with the highest representation. The algorithm yields the highest benefit in the Chinese (18.95%) and Creole (17.14%) speaking groups.

#### Religion

Correspondingly, we partition the patients in eight categories based on their religion. We notice that there only minor differences between most groups with the exception of the Jewish subpopulation where the algorithm expects a 23.26% decrease in the blood pressure score.

### 3.9 | Online Application for Practitioners

In an effort to provide a useful and interpretable tool for practitioners, we developed an online web application using our prescription recommendation algorithm. Through this application (accessible at: <http://alisonrb.shinyapps.io/PersonalizedAntihypertension/>) a physician could enter new patient health data to obtain personalized treatment recommendations. Once the patient information is entered, the application generates a table of model-treatment predictions, similar to that shown in Table 5. In addition, within our tool we display a plot showing the patient's blood pressure score trajectory, where the last point plotted represents the predicted blood pressure score based on the recommended treatment. The intention of this online application is to provide an example of how physicians may utilize the output of machine learning models as a support tool in their decision-making process. The user interface for our online web application is displayed in Figure 2.

## 4 | DISCUSSION

In this study we leveraged longitudinal EHR from an academic medical center to derive predictive and prescriptive models for hypertensive patients. Utilizing multiple machine learning algorithms, we arrive at personalized treatment recommendations for patients with hypertension. Furthermore, we developed an online tool for physicians that can make direct use of EHR data to provide actionable insights from our predictive models. By harnessing the power of a large database of information, we demonstrate the potential of personalized predictions and prescriptions to improve medical outcomes.

Getting access to a large hospital database and formulating a comprehensive dataset for a chronic disease posed significant challenges to our analysis. The majority of the patient records were very sparse, rendering it hard to create observations that follow the full trajectory of a patient over a six-month period. For this reason, multiple visits were aggregated into observations that cover a three-month time window. Moreover, blood pressure is a very unstable measurement whose value may vary significantly even for the same individual during 24 hours. Depending on the time of the day, the clinical and mental status of the patient as well as their diet, the values of systolic and diastolic blood pressures can be severely impacted. Thus, a continuous metric was created for the quantification of the patient status that uses the AHA defined blood pressure classes and their respective relative frequencies.

For the binary prediction tasks, our objective was to employ state-of-the-art machine learning algorithms as well as traditional statistical methods and create accurate models that can accurately inform the medical decision-making process. We introduce the first high performing prediction models for predicting future hypertensive status using widely established machine learning algorithms. Our results demonstrate the superior predictive power of our methodology which results in out-of-sample performance that reaches 90% average AUC.

Our prescription algorithm and corresponding final treatment recommendations display a high level of accuracy in identifying a patient's next blood pressure score, as calculated according to Eq (1). Moreover, our *MAE* results indicate that our predictions accurately capture which blood pressure category a patient's blood pressure measurements will fall into in the next three-month period. The prescriptive algorithm recommends a change in the treatment regimen in approximately 20% of the cases leading to an average improvement of 15.87%. Our findings highlight that our algorithm performs particularly well on certain subgroups, namely the Caucasian population and the population of patients whose age is between 40 and 60. The latter represent the majority of the recorded observations. With respect to potential improvement over the standard of care, we find that patients aged below 40 and above 80 have the greatest potential benefit based on our recommendations.

The strength of our approach is derived from its capacity to combine predictions from independent machine learning models to increase trust in its final predictions. The minimum improvement threshold and the majority agreement condition ensure that we only suggest changes in medication with high efficacy. Our approach thus avoids the prescription of unnecessarily high dosages in cases where the estimated blood pressure score reduction is not significant. Furthermore, the proposed framework takes into account the biases present in observational data, which gives us greater confidence in our ability to identify causal relationships between different treatments and medical outcomes. This debiasing effort is especially important when applying machine learning methods to observational data where randomization is not possible, and it provides further credibility to our results.

Given that we use a wide range of machine learning methods, investigation of feature importance plays a vital role in interpreting our outcomes. Across all models the most important variables were those that were either direct blood pressure measurements or measurements derived from blood pressure values. For all methods, the patient's blood pressure score was the most significant factor, pointing to the importance of developing a summary function to encapsulate a patient's status over time. As discussed, blood pressure is a noisy, unstable measurement that can vary significantly within even a 24-hour period (Frattola, Parati, Cuspidi, Albini, & Mancia 1993). Thus, an important learning outcome is the need to take multiple measurements into account when making continuous treatment regimen decisions for hypertensive patients.

Our work demonstrates the heterogeneity in treatment responses, resulting in a wide variety of outcomes across the patient population. Our models identify the nuanced interactions between the multitude of risk factors that affect blood pressure trajectories and have the ability to play a crucial role in patient management. Specifically, hypertensive patients face a lot of variation in their blood pressure measurements and quantitative approaches such as those presented here can be proven to be a valuable tool for physicians. We recognize that physicians have domain knowledge that cannot be replaced by algorithms, and thus our intention is to provide methods that support physicians in the process they currently use to determine treatments, and to increase the personalization of the treatment process. Future work could focus on highlighting the drivers of risk that affect the algorithm recommendations at the individual level.

## 4.1 | Limitations

The results of our study are subject to several limitations. To begin, our data source was limited to one medical center, BMC, for which the patient population is not necessarily representative of the general United States. Our analysis used data spanning through the years of 2000-2017. Although emergency visits are excluded, the database did not allow us to distinguish between routine office visits, pre-surgery recording, and in-patient hospital stays. Given these considerations, a prospective study is needed to ensure the generalizability of the proposed models in a distinct healthcare system and with more contemporaneous patient records. Such a study would allow the clinical community to test whether these algorithms would extend well to a different dataset without retraining.

Additionally, the EHR data is limited in its capacity to capture other factors which may be relevant for determining treatment decisions and/or patient outcomes. For example, diet and physical activity, both of which are critical components of hypertension management, are not directly recorded in the EHR. Lab values, however, may inherently capture information related to such factors. Though these confounding factors are not captured by the BMC records, we show that our methodology still achieves high predictive performance and provides a quantitative tool that can guide clinical decision making in practice. The aim of this study is to showcase a general framework by developing a tool that is applicable to clinicians across the country that belong to different systems.

A limitation of our classification models refers to causality between the variables and the outcomes, which is still not proven despite the high degree of correlation between the two. We acknowledge that there might be biases in the data with respect to the prescription patterns observed that are specific to the healthcare organization. While we attempt to address causality considerations in our regression models through our matching methodology, which reduces bias resulting from observed confounding variables, we recognize that unobserved confounders may exist. For example, adherence to treatment, which is also critical for controlling blood pressure, is not captured within the dataset. By imposing requirements on observations for inclusion based on frequency and number of visits, however, we aim to capture patients that adhere to their treatment regimen.

Note that there may be variation across the medication classes considered due to confounding factors that are not recorded in the database and thus not included in the covariate matching process. Our prescriptive analysis excludes 36.96% of the samples from the original cohort due to low representation of dosage combinations from other medication categories, such as ACE and Angiotensin II Inhibitors. Furthermore, we would like to acknowledge that including additional predictor variables not present in our dataset may improve prediction accuracy and decrease variance of the results. Even still, we achieve significant improvement over baseline models with respect to both  $R^2$  and  $MAE$  metrics. Further improvements, especially from a clinical practice perspective, might also be gained by considering additional treatment types, or combinations thereof.

We would also like to point out that while, generally, hypertensive treatment is viewed from the longitudinal perspective, we take a stationary view in our approach. We do, however, summarize historical information such as previous visit information so that we can utilize temporal aspects of the data in our framework. Future work should expand upon this approach by looking at the problem from a temporal rather than stationary perspective.

## 5 | CONCLUSION

This study demonstrates the benefit that machine learning prediction algorithms can bring into the care of chronic disease patients. We demonstrate that accurate predictive models can help physicians identify the future trajectory of patients in practice. Furthermore, through this study we present, to the best of our knowledge, the first prescription algorithm for personalized antihypertensive treatment recommendations. We demonstrate the potential value in leveraging longitudinal EHR data for personalized treatment decisions through prediction of heterogeneous treatment effects. We also display strong evidence for the degree to which personalized treatments can improve patient outcomes, relative to the standard of care. We place particular emphasis on essential components of causal inference by applying generalized propensity scoring and matching methods to confirm common support between various treatments and to reduce selection bias surrounding selection of a patient into a treatment regimen. Based on the predictive performance of our models and prescription algorithm, we show that, in instances where counterfactual outcomes cannot be observed, we can reduce uncertainty and improve confidence in the prediction by aggregating the output of multiple machine learning models. Furthermore, we emphasize the importance of interpretability and transparency through the development of an online dashboard that can be used by physicians as a clinical support tool. As additional medical data become available, we believe there is opportunity to expand upon the framework we have developed to provide an even more robust solution for personalized treatment decision-making for hypertensive patients.

## **ACKNOWLEDGMENTS**

The authors wish to thank Bill Adams, MD and the Boston Medical Center for the use of its i2b2 database. We would also like to thank the reviewers from Naval Research Logistics, whose comments significantly improved the paper. No funding was received for this work.

## **CONFLICT OF INTEREST STATEMENT**

No authors have other potential conflicts of interest.

## References

- AHA. (2018, Oct). *Aha high blood pressure toolkit*. Made with FlippingBook. Retrieved from <http://aha-clinical-review.ascendeventmedia.com/books/aha-high-blood-pressure-toolkit/>
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444–455. Retrieved from <http://www.jstor.org/stable/2291629>
- Athey, S., & Imbens, G. (2015). Machine learning methods for estimating heterogeneous causal effects..
- Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353–7360. Retrieved from <https://www.pnas.org/content/113/27/7353> doi: 10.1073/pnas.1510489113
- Austin, P. (2011, 05). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46, 399–424. doi: 10.1080/00273171.2011.568786
- Bennett, M., Vielma, J. P., & Zubizarreta, J. R. (2020). Building representative matched samples with multi-valued treatments in large observational studies. *Journal of Computational and Graphical Statistics*, 0(0), 1–29. Retrieved from <https://doi.org/10.1080/10618600.2020.1753532> doi: 10.1080/10618600.2020.1753532
- Bertsimas, D., Borenstein, A., Mingardi, L., Nohadani, O., Orfanoudaki, A., Stellato, B., ... others (2021). Personalized prescription of acei/arbs for hypertensive covid-19 patients. *Health care management science*, 1–17.
- Bertsimas, D., & Dunn, J. (2017, 04). Optimal classification trees. *Machine Learning*, 106. doi: 10.1007/s10994-017-5633-9
- Bertsimas, D., & Dunn, J. (2019). *Machine learning under a modern optimization lens*. Dynamic Ideas LLC.
- Bertsimas, D., Dunn, J., & Mundru, N. (2019). Optimal prescriptive trees. *INFORMS Journal on Optimization*, 1(2), 164–183. Retrieved from <https://doi.org/10.1287/ijoo.2018.0005> doi: 10.1287/ijoo.2018.0005
- Bertsimas, D., Kallus, N., Weinstein, A. M., & Zhuo, Y. D. (2017). Personalized diabetes management using electronic medical records. *Diabetes Care*, 40(2), 210–217. Retrieved from <https://care.diabetesjournals.org/content/40/2/210> doi: 10.2337/dc16-0826
- Bertsimas, D., Orfanoudaki, A., & Pawlowski, C. (2018). *Imputation of clinical covariates in time series*.
- Bertsimas, D., Orfanoudaki, A., & Weiner, R. B. (2020). Personalized treatment for coronary artery disease patients: a machine learning approach. *Health Care Management Science*, 1–25.
- Breiman, L. (2001, October). Random forests. *Mach. Learn.*, 45(1), 5–32. Retrieved from <https://doi.org/10.1023/A:1010933404324> doi: 10.1023/A:1010933404324
- Breiman, L. (2017). *Classification and regression trees*. CRC Press.
- Burnier, M., & Egan, B. M. (2019). Adherence in hypertension. *Circulation Research*, 124(7), 1124–1140. Retrieved from <https://www.ahajournals.org/doi/abs/10.1161/CIRCRESAHA.118.313220> doi: 10.1161/CIRCRESAHA.118.313220
- Busso, M., DiNardo, J., & McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics*, 96(5), 885–897.
- Byrd, J. B. (2016). Personalized medicine and treatment approaches in hypertension: current perspectives. *Integr Blood Press Control*, 9, 59–67.
- Chadachan, V. M., Ye, M. T., Tay, J. C., Subramaniam, K., & Setia, S. (2018). Understanding short-term blood-pressure-variability phenotypes: from concept to clinical practice. *Int J Gen Med*, 11, 241–254.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (p. 785–794). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/2939672.2939785> doi: 10.1145/2939672.2939785
- Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics*, 13(3), 261–281. Retrieved from <http://www.jstor.org/stable/2527916>
- Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society: Series A (General)*, 128(2), 234–255. Retrieved from <https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2344179> doi: 10.2307/2344179
- Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 35(4), 417–446. Retrieved from <http://www.jstor.org/stable/25049893>
- Cohen, J. (2009). *Statistical power analysis for the behavioral sciences*. Psychology Press.
- Dauvin, A., Donado, C., Bachtiger, P., Huang, K.-C., Sauer, C. M., Ramazzotti, D., ... Douglas, M. J. (2019). Machine learning

- can accurately predict pre-admission baseline hemoglobin and creatinine in intensive care patients. *npj Digital Medicine*.
- Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems* (pp. 155–161).
- Duan, T., Rajpurkar, P., Laird, D., Ng, A. Y., & Basu, S. (2019). Clinical value of predicting individual treatment effects for intensive blood pressure therapy. *Circulation: Cardiovascular Quality and Outcomes*, 12(3), e005010. Retrieved from <https://www.ahajournals.org/doi/abs/10.1161/CIRCOUTCOMES.118.005010> doi: 10.1161/CIRCOUTCOMES.118.005010
- Epstein, C. L. (2014). *An analytics approach to hypertension treatment* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
- Feldstein, M. L., Savlov, E. D., & Hilf, R. (1978, Aug). A statistical model for predicting response of breast cancer patients to cytotoxic chemotherapy. *Cancer Res.*, 38(8), 2544–2548.
- Frattola, A., Parati, G., Cuspidi, C., Albini, F., & Mancia, G. (1993, Oct). Prognostic value of 24-hour blood pressure variability. *J. Hypertens.*, 11(10), 1133–1137.
- Gelman, A. B., & Hill, J. (2018). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
- Guan, W., Liang, W., Zhao, Y., Liang, H.-r., Chen, Z., Li, Y., ... He, J.-x. (2020). Comorbidity and its impact on 1590 patients with covid-19 in china: A nationwide analysis. *European Respiratory Journal*. Retrieved from <https://erj.ersjournals.com/content/early/2020/03/17/13993003.00547-2020> doi: 10.1183/13993003.00547-2020
- Hanley, J., & McNeil, B. (1982, 05). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143, 29–36. doi: 10.1148/radiology.143.1.7063747
- Harder, V., Stuart, E., & Anthony, J. (2010, 09). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, 15, 234–49. doi: 10.1037/a0019623
- Hassanpour, N., & Greiner, R. (2019, 7). Counterfactual regression with importance sampling weights. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI-19* (pp. 5880–5887). International Joint Conferences on Artificial Intelligence Organization. Retrieved from <https://doi.org/10.24963/ijcai.2019/815> doi: 10.24963/ijcai.2019/815
- Hastie, T., Friedman, J., & Tibshirani, R. (2017). *The elements of statistical learning: data mining, inference, and prediction*. Springer.
- Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes. *Journal of the American Statistical Association*, 99(467), 854–866. Retrieved from <https://doi.org/10.1198/016214504000001187> doi: 10.1198/016214504000001187
- Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3), 706–710. Retrieved from <http://www.jstor.org/stable/2673642>
- Imbens, G. W., & Rubin, D. B. (2019). *Causal inference: for statistics, social, and biomedical sciences: an introduction*. Cambridge Univ. Press.
- James, P. A., Oparil, S., Carter, B. L., Cushman, W. C., Dennison-Himmelfarb, C., Handler, J., ... Ortiz, E. (2014, Feb). 2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the Eighth Joint National Committee (JNC 8). *JAMA*, 311(5), 507–520.
- Kallus, N. (2016). *Recursive partitioning for personalization using observational data*.
- Law, M., Morris, J., & Wald, N. (2009). Use of blood pressure lowering drugs in the prevention of cardiovascular disease: meta-analysis of 147 randomised trials in the context of expectations from prospective epidemiological studies. *Bmj*, 338.
- Li, F. (2019). Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, 13(4), 2389–2415.
- Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521), 390–400.
- Mancia, G., & Grassi, G. (2013). Individualization of antihypertensive drug treatment. *Diabetes Care*, 36(Supplement 2), S301–S306. Retrieved from [https://care.diabetesjournals.org/content/36/Supplement\\_2/S301](https://care.diabetesjournals.org/content/36/Supplement_2/S301) doi: 10.2337/dcS13-2013
- Pearl, J. (2009). *Causality*. Cambridge University Press.
- Qian, M., & Murphy, S. A. (2011, 04). Performance guarantees for individualized treatment rules. *Ann. Statist.*, 39(2), 1180–1210. Retrieved from <https://doi.org/10.1214/10-AOS864> doi: 10.1214/10-AOS864

- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55. Retrieved from <http://www.jstor.org/stable/2335942>
- Rubin, D. (2001, 12). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2, 169–188. doi: 10.1023/A:1020363010465
- Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics*, 29(1), 159–183. Retrieved from <http://www.jstor.org/stable/2529684>
- Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29(1), 185–203. Retrieved from <http://www.jstor.org/stable/2529685>
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701. doi: 10.1037/h0037350
- Rubin, D. B. (1990, 11). [on the application of probability theory to agricultural experiments. essay on principles. section 9.] comment: Neyman (1923) and causal inference in experiments and observational studies. *Statist. Sci.*, 5(4), 472–480. Retrieved from <https://doi.org/10.1214/ss/1177012032> doi: 10.1214/ss/1177012032
- Savoia, C., Volpe, M., Grassi, G., Borghi, C., Rosei, E. A., & Touyz, R. (2017). Personalized medicine—a modern approach for the diagnosis and management of hypertension. *Clinical Science*, 131(22), 2671–2685. doi: 10.1042/cs20160407
- Silber, J. H., Rosenbaum, P. R., Ross, R. N., Ludwig, J. M., Wang, W., Niknam, B. A., ... Fleisher, L. A. (2014). Template matching for auditing hospital cost and quality. *Health Services Research*, 49(5), 1446–1474. doi: 10.1111/1475-6773.12156
- Stoecklacher, J., Park, D. J., Zhang, W., Yang, D., Groshen, S., Zahedy, S., & Lenz, H. J. (2004, Jul). A multivariate analysis of genomic polymorphisms: prediction of clinical outcome to 5-FU/oxaliplatin combination chemotherapy in refractory colorectal cancer. *Br. J. Cancer*, 91(2), 344–354.
- Tucker, K. L., Sheppard, J. P., Stevens, R., Bosworth, H. B., Bove, A., Bray, E. P., ... McManus, R. J. (2017, 09). Self-monitoring of blood pressure in hypertension: A systematic review and individual patient data meta-analysis. *PLOS Medicine*, 14(9), 1–29. Retrieved from <https://doi.org/10.1371/journal.pmed.1002389> doi: 10.1371/journal.pmed.1002389
- Turner, S. T., Schwartz, G. L., & Boerwinkle, E. (2007). Personalized medicine for high blood pressure. *Hypertension*, 50(1), 1–5. Retrieved from <https://www.ahajournals.org/doi/abs/10.1161/HYPERTENSIONAHA.107.087049> doi: 10.1161/HYPERTENSIONAHA.107.087049
- Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1).
- Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242. Retrieved from <https://doi.org/10.1080/01621459.2017.1319839> doi: 10.1080/01621459.2017.1319839
- WHO. (2018). *Technical package for cardiovascular disease management in primary health care: healthy-lifestyle counselling* (Technical documents).
- WHO. (2019a, Sep). *Hypertension*. World Health Organization. Retrieved from <https://www.who.int/news-room/fact-sheets/detail/hypertension>
- WHO. (2019b, May). *World hypertension day 2019*. World Health Organization. Retrieved from <https://www.who.int/news-room/events/world-hypertension-day-2019>
- Zhou, T., Tong, G., Li, F., & Thomas, L. E. (2020). Psweight: An r package for propensity score weighting analysis. *arXiv preprint arXiv:2010.08893*.
- Zubizarreta, J. R., Paredes, R. D., & Rosenbaum, P. R. (2014, 03). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *Ann. Appl. Stat.*, 8(1), 204–231. Retrieved from <https://doi.org/10.1214/13-AOAS713> doi: 10.1214/13-AOAS713

---

**TABLES****TABLE 1** Five blood pressure categories as recognized by the AHA.**TABLE 2** Percentage of final cohort belonging to each treatment option.**TABLE 3** Baseline characteristics of the training and testing set with respect to the binary outcomes.**TABLE 4** Summary of predictor variables measured directly from the EHR of the academic medical center.**TABLE 5** Example predictions of outcome for a single patient. The presented values reflect the predicted blood pressure score of a sample patient across all medication categories and regression methods considered.**TABLE 6** Results on the testing set with respect to the average AUC across five different randomized splits of the population.**TABLE 7**  $R^2$  metrics (mean and 95% CI) for full sample.**TABLE 8**  $R^2$  metrics (mean and 95% CI) for matched sample.

**TABLE 9 PE and PR metrics for all models and ground truths considered, converted to percentage (%) decrease in next blood pressure score. The results reflect only samples for which a change in medication is recommended by the algorithm.**

**TABLE 10 Percentage of agreement of machine learning models.**

**TABLE 11 Distribution of best predicted treatment.**

**TABLE 12 Potential outcome improvement over standard of care for patients for whom the algorithm suggests a shift in the treatment regimen.**

**TABLE 13 Subgroup analysis by ethnicity, age, gender, language, and religion.**

## FIGURES

**FIGURE 1 Pre-Treatment covariate balance after matching.** All variables with the prefix “prev” refer to information that was recorded during the previous to the most recent visit of the patient at the hospital. The abbreviations “t2dm”, “ckd”, and “mi” stand for type 2 diabetes mellitus, chronic kidney disease, and myocardial infarction respectively. The comorbidity’s degree of severity is indicated in increasing order by the abbreviations c1-c3.

**FIGURE 2 Online application prototype.**



## TABLES

**TABLE 1 Five blood pressure categories as recognized by the AHA.**

Class	Blood Pressure Category	Systolic mm Hg		Diastolic mm Hg
0	Normal	less than 120	and	less than 80
1	Elevated	120-129	and	less than 80
2	High Blood Pressure (Hypertension) Stage 1	130-139	or	80-89
3	High Blood Pressure (Hypertension) Stage 2	140 or higher	or	90 or higher
4	Hypertensive Crisis	higher than 180	and/or	higher than 120

**TABLE 2 Percentage of final cohort belonging to each treatment option.**

Treatment	Description	Cohort Percentage
Blockers	May include calcium channel blockers, beta blockers, or alpha blockers. Calcium channel blockers prevent calcium from entering the heart and blood vessel muscle cells, causing the cells to relax. Beta blockers work by blocking the effects of adrenaline, which cause your heart to beat slower and with less force. Alpha blockers relax certain muscles and help small blood vessels remain open.	36.25%
Diuretics	Diuretics remove excess water and sodium from the body, which decreases the amount of fluid flowing through the blood vessels.	17.36%
Blockers & Diuretics	Any combination of the drugs classified as either blockers or diuretics.	46.37%

**TABLE 3 Baseline characteristics of the training and testing set with respect to the binary outcomes.**

	Training Set	Testing Set
Sample Size ( <i>N</i> )	20,904	5,227
Improvement Cases	763	191
Ratio of Improvement	3.65%	3.65%
Worsening Cases	643	173
Ratio of Worsening	3.07%	3.33%

**TABLE 4 Summary of predictor variables measured directly from the EHR of the academic medical center.**

Demographics	Blood Pressure Measurements	Lab Values	Medication	Medical History	Other
Age	Median of Systolic Blood Pressure at current trimester	Oxygen levels	Dosage of ACE Inhibitors	Primary cardiovascular event of myocardial infarction <sup>1</sup>	BMI
Gender	Median Diastolic Blood Pressure at current trimester	Cholesterol serum	Dosage of Alpha Blockers	Secondary cardiovascular event of myocardial infarction	Number of months from last trimester recorded
Race	Frequency of category 0 at current trimester	Cholesterol HDL	Dosage of Angiotensin II Inhibitors	Urgent cardiovascular event of myocardial infarction	
Religion	Frequency of category 1 at current trimester	Cholesterol LDL	Dosage of Beta Blockers	Primary cardiovascular event of stroke	
Language	Frequency of category 2 at current trimester	Triglycerides	Dosage of Calcium Channel Blockers	Secondary cardiovascular event of stroke	
	Frequency of category 3 at current trimester	Blood hemoglobin	Dosage Diuretics	Urgent cardiovascular event of stroke	
	Frequency of category 4 at current trimester	Hemoglobin RBC	Dosage Others	Primary adverse event of chronic kidney disease	
	Frequency of category 0 at previous trimester	Hemoglobin RBCc	Drug Type (path/pills/etc.)	Secondary adverse event of chronic kidney disease	
	Frequency of category 1 at previous trimester	Hemoglobin A1c	Duration of drugs prescription	Urgent adverse event of chronic kidney disease	
	Frequency of category 2 at previous trimester	Albumin		Type II diabetes	
	Frequency of category 3 at previous trimester	Urine ph			
	Frequency of category 4 at previous trimester	Hematocrit			
	Current hypertension Score	Chloride			
		Calcium			
		Billirubin			

<sup>1</sup>Primary (secondary) event refers to patients who have experienced one (at least two) adverse event(s). Urgent refers to events which have been recorded in previous visits by the emergency department of the BMC.

**TABLE 5** Example predictions of outcome for a single patient. The presented values reflect the predicted blood pressure score of a sample patient across all medication categories and regression methods considered.

	Blockers & Diuretics High Dosage	Blockers & Diuretics Low Dosage	Blockers Low Dosage	Blockers High Dosage	Diuretics Low Dosage
CART	<b>1.761</b>	2.206	1.858	1.800	2.089
GBM	<b>1.788</b>	2.053	2.007	1.942	1.993
LASSO	1.929	2.072	2.176	1.842	<b>2.083</b>
OPT	<b>1.200</b>	2.019	2.313	2.108	2.199
ORT	<b>1.551</b>	2.116	2.197	1.775	1.961
RF	<b>1.626</b>	1.941	1.941	1.936	1.961
SVM	<b>1.524</b>	1.669	2.129	1.931	1.780

**TABLE 6** Results on the testing set with respect to the average AUC across five different randomized splits of the population.

Target Outcome	Method	AUC	95% CI	Standard Deviation
<b>Worsening</b>	MLR	0.781	(0.772, 0.790)	0.008
	GBM	0.776	(0.766, 0.787)	0.009
	RF	0.768	(0.756, 0.779)	0.010
	OCT	0.720	(0.703, 0.737)	0.015
	CART	0.529	(0.525, 0.534)	0.004
<b>Improvement</b>	MLR	0.900	(0.895, 0.906)	0.005
	GBM	0.899	(0.892, 0.906)	0.006
	RF	0.898	(0.892, 0.904)	0.005
	OCT	0.879	(0.873, 0.886)	0.006
	CART	0.579	(0.573, 0.585)	0.005

**TABLE 7**  $R^2$  metrics (mean and 95% CI) for full sample.

	LASSO	SVM	CART	Random Forest
Blockers & Diuretics (High)	0.38 (0.33, 0.42)	0.36 (0.33, 0.39)	0.3 (0.24, 0.37)	0.39 (0.34, 0.43)
Blockers & Diuretics (Low)	0.38 (0.36, 0.4)	0.36 (0.34, 0.39)	0.33 (0.31, 0.35)	0.38 (0.36, 0.4)
Blockers (High)	0.45 (0.41, 0.49)	0.41 (0.37, 0.45)	0.41 (0.36, 0.45)	0.44 (0.4, 0.48)
Blockers (Low)	0.42 (0.38, 0.45)	0.39 (0.36, 0.42)	0.35 (0.31, 0.38)	0.42 (0.39, 0.44)
Diuretics (Low)	0.42 (0.39, 0.46)	0.37 (0.33, 0.4)	0.36 (0.34, 0.38)	0.42 (0.39, 0.45)
	GBM	ORT	OPT ( $\mu = 0.5$ )	
Blockers & Diuretics (High)	0.38 (0.34, 0.42)	0.38 (0.35, 0.41)	0.19 (0.16, 0.22)	
Blockers & Diuretics (Low)	0.37 (0.36, 0.39)	0.39 (0.35, 0.43)	0.30 (0.27, 0.33)	
Blockers (High)	0.44 (0.41, 0.48)	0.32 (0.28, 0.36)	0.27 (0.25, 0.29)	
Blockers (Low)	0.4 (0.38, 0.43)	0.31 (0.30, 0.32)	0.25 (0.21, 0.29)	
Diuretics (Low)	0.41 (0.38, 0.44)	0.24 (0.20, 0.28)	0.29 (0.27, 0.31)	

**TABLE 8  $R^2$  metrics (mean and 95% CI) for matched sample.**

	LASSO	SVM	CART	Random Forest
Blockers & Diuretics (High)	0.38 (0.35, 0.41)	0.36 (0.34, 0.38)	0.31 (0.25, 0.37)	0.40 (0.35, 0.45)
Blockers & Diuretics (Low)	0.37 (0.34, 0.39)	0.34 (0.31, 0.37)	0.32 (0.3, 0.35)	0.37 (0.35, 0.39)
Blockers (High)	0.45 (0.4, 0.5)	0.4 (0.37, 0.43)	0.41 (0.38, 0.45)	0.45 (0.4, 0.49)
Blockers (Low)	0.41 (0.38, 0.44)	0.38 (0.34, 0.42)	0.34 (0.29, 0.38)	0.41 (0.37, 0.44)
Diuretics (Low)	0.41 (0.35, 0.47)	0.35 (0.3, 0.41)	0.35 (0.29, 0.4)	0.41 (0.36, 0.46)
	GBM	ORT	OPT ( $\mu = 0.5$ )	
Blockers & Diuretics (High)	0.38 (0.34, 0.42)	0.40 (0.37, 0.43)	0.20 (0.16, 0.24)	
Blockers & Diuretics (Low)	0.36 (0.34, 0.39)	0.39 (0.37, 0.41)	0.31 (0.28, 0.34)	
Blockers (High)	0.45 (0.41, 0.48)	0.33 (0.28, 0.38)	0.26 (0.24, 0.28)	
Blockers (Low)	0.4 (0.37, 0.43)	0.29 (0.27, 0.31)	0.25 (0.22, 0.28)	
Diuretics (Low)	0.41 (0.35, 0.47)	0.25 (0.21, 0.29)	0.28 (0.27, 0.29)	

**TABLE 9 PE and PR metrics for all models and ground truths considered, converted to percentage (%) decrease in next blood pressure score. The results reflect only samples for which a change in medication is recommended by the algorithm.**

Estimation Model	Ground Truth							
	Baseline (PE)	LASSO	SVM	CART	RF	GBM	ORT	OPT
LASSO	16.90	8.18	6.40	9.05	9.38	8.94	10.82	7.95
SVM	17.43	9.79	9.72	10.92	11.43	11.36	12.87	10.94
CART	10.55	13.08	11.76	14.92	14.48	14.25	16.28	13.42
RF	18.17	7.81	6.36	8.94	9.29	8.87	10.73	8.05
GBM	18.67	8.03	6.57	9.68	9.75	9.67	11.41	8.84
ORT	18.81	14.44	13.15	16.10	15.76	15.53	17.55	14.76
OPT	16.24	15.77	14.59	17.69	17.94	18.13	19.66	21.54
Prescription Algorithm	15.71	14.97	14.47	16.98	16.13	15.78	18.58	17.49

**TABLE 10 Percentage of agreement of machine learning models.**

Number of Models in Agreement	Blockers Diuretics (High)	Blockers Diuretics (Low)	Blockers (High)	Blockers (Low)	Diuretics (Low)	Overall
2 of 7	9.56%	13.91%	9.27%	11.07%	7.01%	10.10%
3 of 7	41.67%	51.31%	35.72%	50.35%	27.57%	40.58%
4 of 7	30.93%	23.05%	31.68%	25.22%	35.97%	29.63%
5 of 7	13.81%	9.18%	15.45%	10.27%	19.39%	13.90%
6 of 7	3.62%	2.45%	6.46%	2.60%	7.71%	4.76%
7 of 7	0.41%	0.10%	1.43%	0.47%	2.35%	1.03%

**TABLE 11 Distribution of best predicted treatment.**

Actual \ Predicted					
	Blockers Diuretics (High)	Blockers Diuretics (Low)	Blockers (High)	Blockers (Low)	Diuretics (Low)
Blockers & Diuretics (High)	95.98%	1.36%	1.09%	0.91%	0.66%
Blockers & Diuretics (Low)	5.86%	82.93%	5.63%	4.32%	1.26%
Blockers (High)	1.85%	1.59%	93.53%	1.31%	1.72%
Blockers (Low)	7.75%	6.97%	8.28%	74.56%	2.44%
Diuretics (Low)	5.84%	11.14%	14.34%	6.85%	61.83%

**TABLE 12 Potential outcome improvement over standard of care for patients for whom the algorithm suggests a shift in the treatment regimen.**

<b>Actual Treatment</b>	<b>Average of Actual Next Score</b>	<b>Average of Best Next Score</b>	<b>% Decrease in Next Score</b>
Blockers & Diuretics (High)	1.989	1.650	17.03%
Blockers & Diuretics (Low)	1.823	1.522	16.55%
Blockers (High)	1.431	1.194	16.57%
Blockers (Low)	1.525	1.286	15.67%
Diuretics (Low)	1.843	1.557	15.54%

**TABLE 13 Subgroup analysis by ethnicity, age, gender, language, and religion.**

<b>Subgroup</b>	<b>Out-of-Sample <math>R^2</math></b>	<b>Average of Actual Next Score</b>	<b>Average of Best Next Score</b>	<b>% Decrease in Next Score</b>
<b>Ethnicity</b>				
Black	0.44 (0.40, 0.48)	1.820	1.513	16.87%
Caucasian	0.71 (0.68, 0.74)	1.507	1.303	13.58%
Hispanic	0.55 (0.52, 0.58)	1.618	1.422	12.12%
Other	0.61 (0.56, 0.66)	1.833	1.509	17.63%
<b>Age Group</b>				
[18-40)	0.41 (0.35, 0.47)	1.806	1.448	19.80%
[40-60)	0.63 (0.58, 0.66)	1.718	1.459	15.08%
[60-80)	0.51 (0.45, 0.57)	1.719	1.474	14.26%
[80-110)	0.57 (0.52, 0.62)	1.839	1.498	18.52%
<b>Gender</b>				
Female	0.57 (0.52, 0.62)	1.743	1.463	16.09%
Male	0.57 (0.52, 0.62)	1.732	1.464	15.51%
<b>Language</b>				
Chinese	0.43 (0.38, 0.48)	1.677	1.359	18.95%
Creole	0.47 (0.43, 0.51)	1.959	1.623	17.14%
English	0.65 (0.58, 0.64)	1.729	1.454	15.89%
Spanish	0.55 (0.51, 0.59)	1.651	1.432	13.27%
<b>Religion</b>				
Baptist	0.56 (0.52, 0.60)	1.832	1.503	17.94%
Catholic	0.60 (0.52, 0.68)	1.696	1.449	14.53%
Jehovah witness	0.45 (0.40, 0.50)	1.708	1.441	15.65%
Jewish	0.52 (0.47, 0.57)	1.292	0.990	23.36%
Methodist	0.58 (0.52, 0.64)	1.603	1.314	18.06%
Muslim	0.47 (0.42, 0.52)	1.777	1.500	15.54%
None	0.57 (0.55, 0.59)	1.817	1.520	16.35%
Protestant	0.60 (0.53, 0.67)	1.777	1.479	16.81%

# Covariate Balance: Maximum Pairwise Difference Across Treatment Pairs



