Ensemble machine learning for personalized antihypertensive treatment

Dimitris Bertsimas[1,2], Alison Borenstein[1], Antonin Dauvin [1,3], Agni Orfanoudaki[2]

1 Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA
2 Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA
3 Department of Applied Mathematics, Ecole Polytechnique, Palaiseau, France

Corresponding Author:
Agni Orfanoudaki
Muckley Bldg, 1 Amherst St
Cambridge, MA 02142
Email: agniorf@mit.edu

Abstract

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 take patient-specific characteristics into account. In this study, longitudinal Electronic Health Record (EHR) data are utilized to determine the optimal antihypertensive treatment for a patient using his or her individual characteristics and clinical condition. Given the observational nature of the data, we address potential confounding through generalized propensity score evaluation and optimal matching. We use multiple machine learning algorithms to estimate counterfactual predictions for a patient under each treatment option and then apply a voting mechanism among the different models to recommend a treatment based on the best expected outcome. We report results on both the unmatched version of the dataset and the matched dataset. We obtain final out-of-sample R^2 values of 0.60 [95% CI, 0.56-0.64] and 0.55 [95% CI, 0.52-0.59] on the unmatched and matched data, respectively. The final R^2 metric is based on instances for which the treatment suggested by the algorithm matches the patient's actual treatment, thereby allowing us to know the ground truth outcome for comparison. For patients for whom the algorithm recommendation differs from the standard of care, we demonstrate an approximate 15% decrease in next blood pressure based on the predicted outcome under the recommended treatment. Additionally, we develop an interactive dashboard to be used by physicians as a clinical support tool.

Introduction

Hypertension, a medical condition associated with high or elevated blood pressure, affects an estimated 1.13 billion people worldwide [1]. Left untreated, hypertension can increase a patient's risk of developing heart, brain, kidney, and other diseases [2]. Untreated hypertension also increases the risk of stroke [2], which is considered a major cause of premature deaths, and a prevalent co-morbidity of COVID-19 [3].

In 2016, the World Health Organization (WHO) and 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 [4] that designate who should be treated for hypertension and 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 [4].

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 [5]. These guidelines note the same four classes of possible initial treatments to be used, with the main objective of attaining and then maintaining a goal 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 [5].

Despite such guidelines being strongly supported by evidence resulting from randomized controlled trails (RCTs), current treatment decision protocols are not highly personalized [6]. Both guidelines discussed 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. With respect to treatment in practice, most commonly, a trial-and-error approach is adopted whereby physicians use their experience to prescribe an initial treatment and then refine treatment based on the patient's health trajectory. Finding the correct combination of treatments and dosages is typically a lengthy, iterative process. As a result, the prevalence of hypertension is projected to rise rather than to fall in coming years [7].

An added barrier to successful management of hypertension is that blood pressure often fluctuates constantly in response to physical and mental activities and is therefore often characterized by oscillations over short and long-term periods [8]. 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 [9]. Clinicians could greatly benefit from an interpretable tool that uses patient-specific characteristics to recommend a treatment.

Literature review

The aim of this study is to create a model that, given a choice of options, can determine the best treatment for an individual patient. 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 ith observation, $z_i \in [T] = \{1, ..., T\}$ is the assigned treatment, and $y_i \in \mathbb{R}$ is the corresponding outcome under the treatment. We use $y(1), ..., y(T)$ 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 [10]. This framework allows for possible dependence of the mechanism process on potential outcomes [11, 12]. 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 each 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 [13]. Several studies have utilized such methods to predict patient-level responses to treatment [14], as well as to compare different treatments [15]. 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.) [16] 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 [17, 18], 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) [19], as well as tree-based methods that involve recursive partitioning [20], casual trees [21], causal forests [22], and optimal prescriptive trees [23]. Recently, a machine learning based framework was introduced to identify the best therapy for patients with Coronary Artery Disease [24]. In this work, 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 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 [24] 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 [24] 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 [6, 9, 25, 26]. Steps toward this goal have included individual patient data meta-analysis to understand the combined effects of self-monitoring and treatment [27] and using randomized trial data to predict absolute risk reduction (ARR) in cardiovascular events from intensive blood pressure therapy [28]. Application of Electronic Health Records (EHRs) to individualized treatment decision rules remains relatively unexplored.

Contributions

In this paper, we propose the use of EHRs to develop an analytics-based approach to prescribing antihypertensive treatments to patients. Our primary objective is to utilize the ensemble approach proposed by [24] to 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:

- **Development 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. We extend matching methods that are typically applied when comparing two treatments to the multiple treatment case, ensuring that the populations receiving each of the multiple treatments under consideration have similar pre-treatment covariate distributions.

- **Development of predictive models to detect improvement/worsening of hypertension states estimate counterfactuals**: We define a metric that summarizes the blood pressure status of a patient over time. For each patient, we observe this metric (outcome) 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 of the treatments are produced for each patient in the test set.

- **Development of a prescriptive methodology for antihypertensive treatment**: We follow the prescription algorithm set forth by [24] but tailor it to the situation in which a chronic disease (in this case, hypertension) is being treated. Whereas in [24] 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 option for a continuous treatment regimen. Therefore, departing from [24], 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.

- **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`.

Methods

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 150,776 patients, comprising more than 10 million observations corresponding to patient visits between 1982 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 (which defines hypertension).

For each patient, the EHR included demographic data, systolic and diastolic blood pressure values, drug prescription descriptions, dosages, and duration, height, weight, and body mass index (BMI) measurements, history of medical events, and lab value measurements.

Data preprocessing

Due to the noise in blood pressure measurements, individual patient observations were aggregated into three-month time interval summaries. 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 [29].

Feature engineering

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) [30]. 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. In accordance with the approach developed by Bertsimas et al. in *Machine Learning Identifies Guidelines for Blood Pressure Control*, submitted in the Journal of Hypertension in 2020, the blood pressure score metric was calculated with Eq (1). 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.

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

Table 1. Five blood pressure categories as recognized by the American Heart Association.

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

In total, 90 variables were included in the analysis—continuous and categorical variables are summarized in Table 2 and Table 3, 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, and median lab values. Lab-related covariates for which more than 95% of the 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.

Table 2. Summary of continuous patient features.

Variable	1st Quantile	Median	Mean	3rd Quantile
Previous BP Measurements				
Median Systolic	124.00	136.00	137.09	148.00
Minimum Systolic	112.00	124.00	125.67	138.00
Maximum Systolic	134.00	148.00	148.58	160.00
Median Diastolic	74.00	80.00	81.11	88.00
Minimum Diastolic	68.00	74.00	74.98	80.00
Maximum Diastolic	80.00	86.00	87.11	94.00
Cat. 0 Frequency	0.00	0.00	0.14	0.14
Cat. 1 Frequency	0.00	0.29	0.36	0.60
Cat. 2 Frequency	0.00	0.25	0.32	0.50
Cat. 3 Frequency	0.00	0.00	0.14	0.18
Cat. 4 Frequency	0.00	0.00	0.05	0.00
Score	1.00	1.50	1.60	2.00
Current BP Measurements				
Median Systolic	126.00	137.50	137.90	149.50
Minimum Systolic	110.00	120.00	121.99	132.00
Maximum Systolic	140.00	152.00	153.61	166.00
Median Diastolic	74.00	80.00	81.41	88.00
Minimum Diastolic	65.00	70.00	72.85	80.00
Maximum Diastolic	82.00	90.00	89.83	97.00
Cat. 0 Frequency	0.00	0.00	0.14	0.17
Cat. 1 Frequency	0.05	0.29	0.33	0.50
Cat. 2 Frequency	0.00	0.29	0.33	0.50
Cat. 3 Frequency	0.00	0.00	0.15	0.25
Cat. 4 Frequency	0.00	0.00	0.05	0.00
Score	1.00	1.63	1.65	2.18
Laboratory Values				
Oxygen Saturation (%)	97.00	98.00	97.91	99.00
Cholesterol Serum (mg/dL)	168.00	184.87	185.82	199.07
Cholesterol HDL (mg/dL)	41.00	47.66	48.68	53.69
Cholesterol LDL (mg/dL)	97.13	108.21	108.59	118.23
Triglycerides (mg/dL)	95.94	133.00	143.43	162.93
Hemoglobin, blood (g/dL)	11.48	12.60	12.42	13.48
Hemoglobin MCH (pg)	28.07	29.60	29.44	31.00
Hemoglobin MCHC (g/dL)	32.90	33.51	33.42	34.10
Hemoglobin A1c (%)	6.00	6.82	6.89	7.04
Albumin Serum (g/dL)	3.60	3.69	3.64	3.80
Creatinine (μmol/L)	111.19	166.06	170.01	178.31
Urine pH	5.75	6.02	6.07	6.31
Urine Specific Gravity	1.01	1.02	1.02	1.02
Urobilinogen Semiquantitative	0.22	0.24	0.28	0.30
Protein Semiquantitative	824.95	829.28	915.08	1001.97
Hematocrit (%)	34.50	37.65	37.17	40.17
Chloride Serum (mEq/L)	102.00	103.98	103.78	105.50

Calcium Serum (mg/dL)	9.10	9.40	9.36	9.69
Ferritin Serum (mg/L)	101.90	210.76	232.16	230.34
Iron Serum (μg/dL)	58.26	64.50	66.19	71.17
Iron-binding Capacity (μg/dL)	291.73	302.02	301.42	311.49
Bilirubin (mg/dL)	0.35	0.50	0.54	0.60
Demographics & Physical				
Age	45.25	54.61	55.13	64.60
Height (cm)	160.80	165.19	165.77	170.00
Weight (kg)	71.30	83.60	87.01	98.80
BMI	26.07	30.25	31.25	35.00
Other				
Prescription Duration (days)	64.00	121.00	131.81	170.77
Visit Count	7.00	10.00	13.36	15.00

Table 3. Summary of categorical patient features.

Variable	Percentage of Population
Gender	
Female	58.11%
Male	41.89%
Ethnicity	
Asian	1.71%
Black	57.20%
Caucasian	21.31%
Hispanic	12.38%
Other	7.41%
Language	
English	76.73%
Chinese	2.15%
Creole	9.76%
Spanish	8.46%
Other	2.91%
Religion	
Baptist	17.43%
Catholic	36.34%
Christian	9.68%
Jehovah's Witness	1.88%
Jewish	0.87%
Methodist	1.52%
Muslim	1.62%
Protestant	15.57%
None	4.50%
Other	10.58%
Medical History	
Primary cardiovascular event of myocardial infarction	3.65%
Secondary cardiovascular event of myocardial infarction	3.07%
Urgent cardiovascular event of myocardial infarction	1.58%
Primary cardiovascular event of stroke	7.03%
Secondary cardiovascular event of stroke	5.26%
Urgent cardiovascular event of stroke	5.45%
Primary adverse event of chronic kidney disease	12.21%
Secondary adverse event of chronic kidney disease	15.39%

Urgent adverse event of chronic kidney disease	7.66%
Type II Diabetes Mellitus	45.41%
Previous Medication Dosage	
ACE Inhibitors - Low	21.28%
ACE Inhibitors - Medium	2.12%
ACE Inhibitors - High	8.63%
Angiotensin II Inhibitors - Low	0.36%
Angiotensin II Inhibitors - Medium	0.32%
Angiotensin II Inhibitors - High	0.47%
Blockers - Low	23.64%
Blockers - Medium	20.88%
Blockers - High	15.30%
Diuretics - Low	38.37%
Diuretics - Medium	5.73%
Diuretics - High	3.14%
Others - Low	0.20%
Others - Medium	0.20%
Others - High	0.40%
None	9.82%

Prior to separation of the dataset into cohorts based on treatment type, current treatments were encoded as binary variables and included the following types: ACE Inhibitors, Blockers (Alpha, Beta, and/or Calcium Channel), and Diuretics. The current observation for a patient had to adhere to one of the following characteristics:

- **Monotherapy:** Currently taking *one* of the three current treatment options;
- **Two Treatment Combination Therapy**: Currently taking a *two treatment combination* of the three current treatment options.

The intention of this inclusion criteria was to effectively capture a large portion of the patient population while also limiting the number of potential treatment options for comparative analysis. Starting with the three-month aggregated observations for which one or more medications were taken (N = 82,736), over 53% of the population took either one or two medications.

Outcome of interest

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.

Treatment options

In order to compare multiple treatments, separate models were trained for each treatment option. Thus, the final cohort of 19,926 observations was divided into six mutually exclusive subsets based on the current treatment regimen. Percentages of the final cohort belonging to each type of treatment are listed in Table 4.

Debiasing approach

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

Table 4. Percentage of final cohort belonging to each treatment option.

Treatment	Description	Cohort Percentage
ACE Inhibitors	ACE Inhibitors help relax blood vessels by preventing the formation of a hormone called angiotensin, a substance in the body that narrows blood vessels.	9.03%
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.	21.88%
Diuretics	Diuretics remove excess water and sodium from the body, which decreases the amount of fluid flowing through the blood vessels.	13.12%
Blockers & ACE Inhibitors	Any combination of the drugs classified as either Blockers or ACE Inhibitors.	15.48%
Blockers & Diuretics	Any combination of the drugs classified as either Blockers or Diuretics.	30.88%
Diuretics & ACE Inhibitors	Any combination of the drugs classified as either Diuretics or ACE Inhibitors.	9.61%

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 [31]. 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.

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

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 [31]. 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 [32]. 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 [33]. In this study, we utilized generalized propensity scores, which were developed as an extension of the propensity score measure for binary treatment [34–36], 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 [34]. We verified that common support existed through the use of generalized propensity score (GPS) [35]. 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 multinomial logistic regression. 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 [32]. For example, a patient's age might dictate which treatment options he or she is 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 [32]. 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.

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 [37–39].

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 [40]. In this study, we used a form of optimal matching known as cardinality matching [41]. 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. [42] and Bennett et al. [43] 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` [43]. 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 [44, 45]. Post-matching, we assessed the balance between the treatment groups using pairwise standardized absolute mean differences, as suggested by [47]. Typically, different groups are considered balanced if the standardized absolute mean differences between the groups are less than 0.25 [47, 48]. Once this level of balance is achieved, outcome analysis can be performed.

Predictive models

In order to arrive at a medication recommendation, we had to infer the patient's response to each of the treatment options. For this task we utilized a separate model approach. Thus, rather than adding an indicator of treatment type as a feature, which would risk the learned function not taking the treatment assignment variable into account, we created a suite of models for each treatment separately.

We developed predictive models for both the unmatched version of the dataset and the matched version of the dataset, and we present results for both. In both cases, the final cohort for each treatment option was used to train models to predict the next blood pressure score for each observation. 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 type, consisting of linear and non-linear methods ranging from highly interpretable to black-box, 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), support vector regression (SVR), classification and regression trees (CART), random forest (RF), gradient boosting machine (GBM), optimal regression trees (ORT), and optimal prescriptive trees (OPT) (see S1 Table). 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 [23]; thus, their application is highly relevant to this problem.

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, μ, 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. 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.

Prescriptive component

After using models trained on each separate treatment cohort to predict the counterfactual estimations of the next blood pressure score, we obtained a matrix of model-treatment combinations for each patient in the test set. An example of such a matrix is shown in Table 5. Our algorithmic framework for treatment prescription is similar to that described by Bertsimas et al. [24], except that we only change the treatment regimen if the majority of the models agree that an alternate treatment will result in a better outcome than the current treatment. Our methodology is as follows:

1. Using the matrix of counterfactual estimates, we derive which treatment each regression model selects as the best treatment, based on the lowest expected outcome (next blood pressure score).

2. We also ascertain which of the treatment options is most frequently chosen as the best treatment.

3. If the majority of the regression models agree that a certain treatment 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.

4. If there is not majority agreement among the regression models, the algorithm defers to the physician to determine if the patient should remain on their current treatment regimen or if current treatment should be refined.

Table 5. Example predictions of outcome for a single patient.

	Ace Inhibitors	Blockers	Blockers & ACE Inhibitors	Blockers & Diuretics	Diuretics	Diuretics & ACE Inhibitors
CART	**1.761**	2.206	1.858	1.800	1.911	2.089
GBM	**1.788**	2.053	2.007	2.122	2.012	1.993
LASSO	1.929	2.072	2.176	2.083	1.993	**1.842**
OPT	**1.200**	2.019	2.313	2.108	2.099	2.199
ORT	**1.551**	2.116	2.197	1.775	1.865	1.961
RF	**1.626**	1.941	1.941	1.936	1.957	1.961
SVM	**1.524**	1.669	2.129	1.931	1.713	1.780

Utilizing the patient output displayed 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 ACE Inhibitors is the treatment that will result in the lowest blood pressure score for the next three-month period, while one model (LASSO) predicts a combination of Diuretics and ACE Inhibitors to be the best treatment option. Given that model agreement is 85.7% (6 of 7 models agree), ACE Inhibitors 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. 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.

Results

In this section, we present prediction evaluation metrics for each model-treatment 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.

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 75% of the original final dataset (resulting in $N_{matched} = 14{,}998$) while achieving pairwise balance across all treatments below 0.25 for all but one of the 34 pre-treatment covariates, as shown in Fig 1. For the variable not meeting balance criteria, the pairwise difference was still reduced considerably in comparison with the original, unmatched data. Furthermore, 82% of the pre-treatment variables after matching had a standardized absolute mean difference below 0.10. In Supporting Information S2 Table and S3 Table, we summarize the pre-treatment variables stratified by treatment before and after the matching procedure.

Fig 1. Pre-Treatment covariate balance after matching.

Predictive regression modeling results

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. The R^2 values ranged from 0.15 to 0.50, depending on the treatment subset and model type. MAE ranged from 0.44 to 0.59. The out-of-sample performance for the Diuretics models were superior to all other treatment types. In terms of predictive accuracy, LASSO, RF, and GBM models outperformed the others. Evaluation metrics for the full, unmatched sample are presented in Table 6 and S4 Table. 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 0.22 to 0.49, and the MAE values ranged from 0.43 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 LASSO, RF, and GBM models display the highest predictive accuracy in general. Evaluation metrics for the matched sample are presented in Table 7 and S5 Table.

Prescription algorithm results

In [24], 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

Table 6. R^2 metrics (mean and 95% CI) for full sample.

	LASSO	SVM	CART	Random Forest
ACE Inhibitors	0.43 (0.40, 0.45)	0.37 (0.34, 0.41)	0.36 (0.32, 0.40)	0.43 (0.40, 0.46)
Blockers	0.43 (0.42, 0.45)	0.40 (0.38, 0.42)	0.38 (0.35, 0.41)	0.43 (0.42, 0.45)
Diuretics	0.50 (0.49, 0.51)	0.46 (0.45, 0.46)	0.45 (0.44, 0.46)	0.49 (0.49, 0.50)
Blockers & ACE Inhibitors	0.36 (0.31, 0.41)	0.33 (0.28, 0.37)	0.30 (0.24, 0.35)	0.35 (0.30, 0.40)
Blockers & Diuretics	0.39 (0.37, 0.41)	0.37 (0.35, 0.39)	0.33 (0.31, 0.35)	0.38 (0.37, 0.40)
Diuretics & ACE Inhibitors	0.37 (0.31, 0.44)	0.34 (0.28, 0.40)	0.29 (0.22, 0.36)	0.37 (0.30, 0.44)

	Boosted Trees	ORT	OPT ($\mu = 0.5$)
ACE Inhibitors	0.44 (0.41, 0.46)	0.37 (0.34, 0.41)	0.17 (0.11, 0.23)
Blockers	0.43 (0.42, 0.44)	0.37 (0.36, 0.39)	0.32 (0.28, 0.36)
Diuretics	0.50 (0.49, 0.51)	0.43 (0.40, 0.47)	0.29 (0.25, 0.33)
Blockers & ACE Inhibitors	0.35 (0.30, 0.40)	0.29 (0.25, 0.34)	0.21 (0.16, 0.26)
Blockers & Diuretics	0.38 (0.37, 0.40)	0.33 (0.32, 0.34)	0.27 (0.22, 0.32)
Diuretics & ACE Inhibitors	0.37 (0.30, 0.45)	0.26 (0.19, 0.34)	0.15 (0.06, 0.24)

Table 7. R^2 metrics (mean and 95% CI) for matched sample.

	LASSO	SVM	CART	Random Forest
ACE Inhibitors	0.42 (0.39, 0.45)	0.37 (0.32, 0.41)	0.36 (0.32, 0.40)	0.42 (0.39, 0.44)
Blockers	0.45 (0.41, 0.48)	0.41 (0.38, 0.44)	0.39 (0.35, 0.42)	0.45 (0.42, 0.47)
Diuretics	0.49 (0.44, 0.54)	0.44 (0.39, 0.48)	0.43 (0.39, 0.48)	0.48 (0.43, 0.53)
Blockers & ACE Inhibitors	0.35 (0.30, 0.40)	0.31 (0.27, 0.35)	0.30 (0.24, 0.35)	0.35 (0.30, 0.39)
Blockers & Diuretics	0.39 (0.37, 0.41)	0.35 (0.33, 0.37)	0.33 (0.31, 0.35)	0.38 (0.37, 0.39)
Diuretics & ACE Inhibitors	0.38 (0.35, 0.41)	0.33 (0.28, 0.38)	0.30 (0.26, 0.35)	0.38 (0.34, 0.41)

	Boosted Trees	ORT	OPT ($\mu = 0.5$)
ACE Inhibitors	0.43 (0.40, 0.46)	0.36 (0.32, 0.41)	0.26 (0.19, 0.32)
Blockers	0.44 (0.41, 0.48)	0.39 (0.35, 0.43)	0.34 (0.31, 0.37)
Diuretics	0.49 (0.43, 0.54)	0.42 (0.36, 0.47)	0.35 (0.24, 0.46)
Blockers & ACE Inhibitors	0.35 (0.30, 0.39)	0.29 (0.23, 0.34)	0.22 (0.12, 0.32)
Blockers & Diuretics	0.38 (0.36, 0.41)	0.31 (0.29, 0.34)	0.27 (0.26, 0.28)
Diuretics & ACE Inhibitors	0.37 (0.34, 0.41)	0.29 (0.23, 0.34)	0.22 (0.18, 0.26)

was used because we can only ever realize one factual outcome for each observation, and the other five 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 0.60 [95% CI, 0.56-0.64] using the unmatched dataset and 0.55 [95% CI, 0.52-0.59] using the matched dataset.

We also adopted the Prescription Effectiveness (PE) and Prescription Robustness (PR) metrics introduced by [24]. 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. The PR metric is determined by generating alternative ground truths assuming that each regression model knew the outcome reality 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 for our outcome of

interest, we transform the raw outcome (which corresponds to the decrease in 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 8, while the raw outcome PE and PR metrics are summarized in Supporting Information S6 Table.

Table 8. PE and PR metrics for all models and ground truths considered, converted to percentage decrease in next blood pressure score.

	Ground Truth							
Estimation Model	Baseline (PE)	LASSO	SVM	CART	RF	GBM	ORT	OPT
LASSO	4.39	5.55	3.42	5.45	5.70	5.53	5.41	5.13
SVM	7.80	8.92	6.86	8.81	9.06	8.89	8.78	8.51
CART	10.67	11.75	9.76	11.66	11.89	11.73	11.63	11.36
RF	6.30	7.44	5.34	7.33	7.58	7.41	7.30	7.02
GBM	7.75	8.87	6.80	8.76	9.01	8.84	8.73	8.46
ORT	11.16	12.24	10.26	12.15	12.38	12.22	12.12	11.85
OPT	21.41	22.37	20.62	22.28	22.49	22.35	22.25	22.02
Prescription Algorithm	15.02	16.05	14.15	15.96	16.18	16.03	15.93	15.67

Table 8 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 estimatation models as the ground truth. Relative to the current allocation plan, our prescription algorithm allocation plan represents a 15.02% expected decrease in next blood pressure score. By observing the first column of Table 8, we find that OPT is the most optimistic model, estimating a 21.41% decrease over the baseline ground truth. LASSO, on the other hand, is the most conservative, estimating an approximate 4.39% decrease in blood pressure score relative to the baseline. The remaining regression models estimate fairly similar decreases, between 6.30% and 11.16%. 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. As was the case for the PE metric, OPT is the most optimistic model across all ground truths and LASSO is the least optimistic. Although we observe that OPT is more optimistic than our prescription algorithm, we find that the final R^2 values obtained from comparing cases where the treatment that OPT determines as optimal matches the actual treatment is 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.

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 and 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 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 as well as on the Diuretics & ACE Inhibitors subset. Additional lab measurements that were frequently identified as significant included creatinine, iron, and triglycerides. Age was identified as an important factor for the Blockers, Blockers & Diuretics, and Diuretics & ACE Inhibitors models. Furthermore, duration of drug prescription was important for the ACE Inhibitors, Blockers, and Diuretics & ACE Inhibitors models. As demonstrated by these examples, factors identified as important for a particular monotherapy were typically also identified by the combination therapies for which the monotherapy was a component. Feature importance was consistent between the unmatched and matches samples for each treatment type.

Model agreement

For the prescription algorithm, we recorded the level of agreement among the regression models. In Table 9, we report the percentage of agreement of the machine learning models by treatment type and overall. Of the testing set observations, majority agreement among the seven models is achieved in approximately 46% 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.

Table 9. Percentage of agreement of machine learning models.

Number of Models in Agreement	Ace Inhibitors	Blockers	Diuretics	Blockers & ACE Inhibitors	Blockers & Diuretics	Diuretics & ACE Inhibitors	Overall
2 of 7	10.68%	30.36%	10.23%	39.58%	27.57%	12.66%	14.02%
3 of 7	42.29%	41.45%	36.14%	42.88%	46.41%	42.22%	39.89%
4 of 7	29.83%	19.00%	30.62%	14.52%	20.56%	29.54%	28.41%
5 of 7	13.41%	7.41%	16.76%	2.36%	4.97%	12.47%	13.37%
6 of 7	3.30%	1.65%	5.56%	0.65%	0.49%	2.85%	3.83%
7 of 7	0.49%	0.13%	0.69%	0.00%	0.00%	0.26%	0.48%

Treatment recommendation distributions

In Table 10, 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, we assume for these distributions that the patient will remain on his or her current line of treatment. By looking at the reallocation percentages, the results suggest that monotherapies are preferable to combination therapies. Furthermore, Diuretics and ACE Inhibitors appear to be the most frequently chosen treatment options. Blockers, on the other hand, is the least frequent monotherapy option. Among the two-treatment combination therapy options, Diuretics & ACE Inhibitors is the most frequently prescribed option.

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 Beta Blockers for initial treatment due to complications associated with cardiovascular death, myocardial infarction, or stroke [4].

Improvement over standard of care

In cases where the majority of models agree on a best treatment, our algorithm agrees with the actual treatment the patient received between 14% and 18% of the time, depending on the split of the data. After obtaining final prediction outcomes for patients

Table 10. Distribution of predicted best treatment.

Actual \ Predicted	ACE Inhibitors	Blockers	Blockers & ACE Inhibitors	Blockers & Diuretics	Diuretics	Diuretics & ACE Inhibitors
ACE Inhibitors	65.75%	2.49%	0.64%	1.33%	20.94%	8.85%
Blockers	12.97%	55.92%	0.44%	2.21%	19.22%	9.23%
Blockers & ACE Inhibitors	14.78%	2.40%	54.75%	2.34%	19.52%	6.22%
Blockers & Diuretics	13.33%	1.33%	0.20%	55.94%	18.05%	11.14%
Diuretics	8.18%	1.36%	0.39%	1.65%	75.96%	12.48%
Diuretics & ACE Inhibitors	10.08%	1.66%	0.63%	1.51%	21.84%	64.29%

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 predictions.

Table 11 displays the mean actual outcome, mean predicted best outcome, and the corresponding percentage decrease in patients' next blood pressure scores, where 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 Blockers, Blockers & ACE Inhibitors, and Blockers & Diuretics. For these patients, the best potential outcome represents a decrease in score of approximately 17%.

Table 11. Potential outcome improvement over standard of care.

Actual Treatment	Average of Actual Next Score	Average of Best Next Score	% Decrease in Next Score
Ace Inhibitors	1.485	1.281	13.78%
Blockers	1.531	1.268	17.21%
Blockers & Ace Inhibitors	1.644	1.337	18.71%
Blockers & Diuretics	1.568	1.316	16.09%
Diuretics	1.550	1.357	12.41%
Diuretics & Ace Inhibitors	1.387	1.203	13.22%

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):

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

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:

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

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 over current standard of care. Improvement among specific sub-groups is detailed in the subgroup analysis section.

Subgroup analysis

We expand the results of our study through subgroup analysis, whereby we investigate the model agreement, final out-of-sample R^2, and potential decrease in next blood pressure score for each subgroup shown in Table 12.

Table 12. Subgroup analysis by ethnicity, age, and gender.

Subgroup	% of Models with Majority Agreement	Out-of-Sample R^2	Average of Actual Next Score	Average of Best Next Score	% Decrease in Next Score
Ethnicity					
Black	46.20%	0.38 (0.31, 0.46)	1.620	1.347	16.87%
Caucasian	46.67%	0.74 (0.66, 0.81)	1.339	1.152	13.95%
Hispanic	46.42%	0.50 (0.33, 0.66)	1.528	1.298	15.10%
Other	43.58%	0.55 (0.48, 0.63)	1.648	1.372	16.75%
AgeBucket					
[18-40)	47.03%	0.39 (0.16, 0.63)	1.570	1.310	16.55%
[40-60)	46.68%	0.62 (0.56, 0.69)	1.530	1.279	16.38%
[60-80)	45.28%	0.48 (0.37, 0.58)	1.549	1.316	15.04%
[80-110)	44.44%	0.54 (0.35, 0.73)	1.592	1.303	18.15%
Gender					
Female	46.87%	0.56 (0.51, 0.60)	1.549	1.292	16.61%
Male	44.95%	0.56 (0.51, 0.60)	1.539	1.305	15.22%

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 74.0% was achieved when comparing patients for whom recommended treatment matched actual treatment. Additional insights from this analysis are that patients predicted to experience the greatest decrease in their next blood pressure score are those whose ethnicity is Black or Other.

Age

We also grouped patients into 4 different age buckets to investigate model performance on different age populations. We found that out-of-sample performance was highest for patients whose age is between 40 and 60. We also found that the patients in the eldest [80-110) age group had the greatest potential benefit in terms of decrease in blood pressure score.

Gender

Outcomes in terms of model agreement, out-of-sample performance, and potential blood pressure score decrease are similar. This finding is not surprising, given that gender was not indicated as an important factor by the regression models.

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 Fig 2.

Fig 2. Online application prototype.

Discussion

In this study we leveraged longitudinal EHR data from an academic medical center and multiple machine learning techniques to 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 combined with state-of-the-art machine learning algorithms, we demonstrate the potential of personalized treatments to improve medical outcomes.

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. Our results also indicate 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. With respect to potential improvement over the standard of care, we find that patients aged 80 to 110 have the greatest potential benefit based on our recommendations.

The strength of our approach is derived from our ability to combine predictions from independent machine learning models to increase trust in our final predictions. Furthermore, our approach 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 [49]. Thus, an important learning outcome is the need to take multiple measurements into account when making continuous treatment regimen decisions for hypertensive patients.

The goal of this work was to provide a framework for better management of hypertension by focusing on treatment at an individual level. Our work demonstrates the heterogeneity in treatment responses, resulting in a wide variety of outcomes across the patient population. We recognize that physicians have domain knowledge that cannot be replaced by algorithms, and thus our intention is to provide a method that supports physicians in the process they currently use to determine treatments, and to increase the personalization of the treatment process.

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 population. 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. Furthermore, it is important to note that our matching methodology reduces bias resulting from observed confounding variables. 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. Finally, 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.

Conclusion

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.

References

1. Hypertension [Internet]. World Health Organization. World Health Organization; [cited 2020Apr24]. Available from: https://www.who.int/news-room/fact-sheets/detail/hypertension
2. World Hypertension Day 2019 [Internet]. World Health Organization. World Health Organization; 2019 [cited 2020Apr24]. Available from: https://www.who.int/cardiovascular_diseases/world-hypertension-day-2019/en/
3. Guan WJ, Liang WH, Zhao Y, Liang HR, Chen ZS, Li YM, et al. Comorbidity and its impact on 1,590 patients with COVID-19 in China: A Nationwide Analysis [Internet]. medRxiv. Cold Spring Harbor Laboratory Press; 2020 [cited 2020Apr24]. Available from: https://www.medrxiv.org/content/10.1101/2020.02.25.20027664v1
4. Technical package for cardiovascular disease management in primary health care: healthy-lifestyle counselling [Internet]. World Health Organization. World Health

Organization; 1970 [cited 2020Apr24]. Available from: https://apps.who.int/iris/handle/10665/260422

5. James PA, Oparil S, Carter BL, Cushman WC, Dennison-Himmelfarb C, Handler J, et al. 2014 Evidence-Based Guideline for the Management of High Blood Pressure in Adults. Jama. 2014;311(5):507-520.

6. Byrd JB. Personalized medicine and treatment approaches in hypertension: current perspectives. Integrated Blood Pressure Control. 2016;9:59–67.

7. Burnier M, Egan MB. Adherence in Hypertension. Circulation Research. 2019;124(7):1124–40.

8. Chadachan VM, Ye MT, Tay JC, Subramaniam K, Setia S. Understanding short-term blood-pressure-variability phenotypes: from concept to clinical practice. International Journal of General Medicine. 2018;11:241–54.

9. Turner ST, Schwartz GL, Boerwinkle E. Personalized Medicine for High Blood Pressure. Hypertension. 2007;50(1):1–5.

10. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974;66(5):688–701.

11. Rubin DB. [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. Statistical Science. 1990;5(4):472–80.

12. Angrist J, Imbens G, Rubin DB. Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association. 1996;91(434):444–55.

13. Stoehlmacher J, Park DJ, Zhang W, Yang D, Groshen S, Zahedy S, et al. A multivariate analysis of genomic polymorphisms: prediction of clinical outcome to 5-FU/oxaliplatin combination chemotherapy in refractory colorectal cancer. British Journal of Cancer. 2004;91(2):344–54.

14. Feldstein ML, Savlov ED, Hilf R. A Statistical Model for Predicting Response of Breast Cancer Patients to Cytotoxic Chemotherapy. American Association for Cancer Research. 1978;38(8):2544–8.

15. Qian M, Murphy SA. Performance guarantees for individualized treatment rules. The Annals of Statistics. 2011;39(2):1180–210.

16. Imbens GW, Rubin DB. Causal inference: for statistics, social, and biomedical sciences: an introduction. New York, New York: Cambridge Univ. Press; 2019.

17. Athey S, Imbens GW, Guido, Ramachandra V. Machine Learning Methods for Estimating Heterogeneous Causal Effects. 2015.

18. Hassanpour N, Greiner R. CounterFactual Regression with Importance Sampling Weights. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. 2019.

19. Bertsimas D, Kallus N, Weinstein AM, Zhuo YD. Personalized Diabetes Management Using Electronic Medical Records. Diabetes Care. 2016;40(2):210–7.

20. Kallus N. Recursive Partitioning for Personalization using Observational Data. Proceedings of the 34th International Conference on Machine Learning. 2017;70:1789–98.

21. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences. 2016;113(27):7353–60.

22. Wager S, Athey S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. Journal of the American Statistical Association. 2018;113(523):1228–42.

23. Bertsimas D, Dunn J, Mundru N. Optimal Prescriptive Trees. INFORMS Journal on Optimization. 2019;1(2):164–83.

24. Bertsimas D, Orfanoudaki A, Weiner RB. Personalized Treatment for Coronary Artery Disease Patients: A Machine Learning Approach [Internet]. arXiv.org. 2019 [cited 2020Apr25]. Available from: https://arxiv.org/abs/1910.08483

25. Savoia C, Volpe M, Grassi G, Borghi C, Rosei EA, Touyz RM. Personalized medicine—a modern approach for the diagnosis and management of hypertension. Clinical Science. 2017;131(22):2671–85.

26. Mancia G, Grassi G. Individualization of Antihypertensive Drug Treatment. Diabetes Care. 2013;36(Supplement_2).

27. Sheppard J, Tucker K, Stevens R, Bosworth H, Kantola I, Kerry S, et al. Self-monitoring of blood pressure in hypertension: A systematic review and individual patient data meta-analysis. PLoS Medicine. 2017;14(9):1–29.

28. Duan T, Rajpurkar P, Laird D, Ng AY, Basu S. Clinical Value of Predicting Individual Treatment Effects for Intensive Blood Pressure Therapy. Circulation: Cardiovascular Quality and Outcomes. 2019;12(3).

29. Bertsimas D, Orfanoudaki A, Pawlowski C. Imputation of Clinical Covariates in Time Series [Internet]. arXiv.org. 2018 [cited 2020Apr23]. Available from: https://arxiv.org/abs/1812.00418

30. Understanding Blood Pressure Readings [Internet]. www.heart.org. [cited 2020Apr23]. Available from: https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

31. Pearl J. Causality. Cambridge: Cambridge University Press; 2009.

32. Gelman AB, Hill J. Data analysis using regression and multilevel/hierarchical models. Cambridge: Cambridge University Press; 2018.

33. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research. 2011;46(3):399–424.

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

35. Imbens G. The role of the propensity score in estimating dose-response functions. Biometrika. 2000;87(3):706–10.

36. Imai K, Dyk DAV. Causal Inference With General Treatment Regimes. Journal of the American Statistical Association. 2004;99(467):854–66.

37. Cochran WG, Rubin DB. Controlling Bias in Observational Studies: A Review. Sankhya: The Indian Journal of Statistics, Series A. 1973;35(4):417–46.

38. Rubin DB. Matching to Remove Bias in Observational Studies. Biometrics. 1973;29(1):159–84.

39. Rubin DB. The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. Biometrics. 1973;29(1):185–203.

40. Harder VS, Stuart EA, Anthony JC. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychological Methods. 2010;15(3):234–49.

41. Zubizarreta JR, Paredes RD, Rosenbaum PR. Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. The Annals of Applied Statistics. 2014;8(1):204–31.

42. Silber JH, Rosenbaum PR, Ross RN, Ludwig JM, Wang W, Niknam BA, et al. Template Matching for Auditing Hospital Cost and Quality. Health Services Research. 2014;49(5):1446–74.

43. Bennett M, Vielma JP, Zubizarreta JR. Building Representative Matched Samples with Multi-valued Treatments in Large Observational Studies. Journal of Computational and Graphical Statistics. 2020;:1–42.

44. Cochran WG. Analysis of Covariance: Its Nature and Uses. Biometrics. 1957;13(3):261–81.

45. Cochran WG, Chambers SP. The Planning of Observational Studies of Human Populations. Journal of the Royal Statistical Society Series A (General). 1965;128(2):234–65.

46. Rosenbaum PR, Rubin DB. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. The American Statistician. 1985;39(1):33–8.

47. Rubin DB. Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. Health Services & Outcomes Research Methodology. 2001;2:169–188.

48. Cohen J. Statistical power analysis for the behavioral sciences. New York, New York: Psychology Press, Taylor & Francis Group; 2009.

49. Frattola A, Parati G, Cuspidi C, Albini F, Mancia G. Prognostic value of 24-hour blood pressure variability. Journal of Hypertension. 1993;11(10):1133–7.

Supporting information captions

S1 Table. Description of machine learning models used for regression task.

S2 Table. Summary of pre-treatment variables before matching.

S3 Table. Summary of pre-treatment variables after matching.

S4 Table. MAE **metrics (mean and 95% CI) for full sample.**

S5 Table. MAE **metrics (mean and 95% CI) for matched sample.**

S6 Table. PE and PR metrics for all models and ground truths considered.

S7 Table. Description of abbreviated terms.