

Personalized Treatment for Coronary Artery Disease Patients: A Machine Learning Approach

Cite this article as: Dimitris Bertsimas, Agni Orfanoudaki and Rory B. Weiner, Personalized Treatment for Coronary Artery Disease Patients: A Machine Learning Approach, *Health Care Management Science* doi: [10.1007/s10729-020-09522-4](https://doi.org/10.1007/s10729-020-09522-4)

This Author Accepted Manuscript is a PDF file of a an unedited peer-reviewed manuscript that has been accepted for publication but has not been copyedited or corrected. The official version of record that is published in the journal is kept up to date and so may therefore differ from this version.

Terms of use and reuse: academic research for non-commercial purposes, see here for full terms. <http://www.springer.com/gb/open-access/authors-rights/aam-terms-v1>

Personalized Treatment for Coronary Artery Disease Patients: A Machine Learning Approach

Received: date / Accepted: date

Abstract Current clinical practice guidelines for managing Coronary Artery Disease (CAD) account for general cardiovascular risk factors. However, they do not present a framework that considers personalized patient-specific characteristics. Using the electronic health records of 21,460 patients, we created data-driven models for personalized CAD management that significantly improve health outcomes relative to the standard of care. We develop binary classifiers to detect whether a patient will experience an adverse event due to CAD within a 10-year time frame. Combining the patients' medical history and clinical examination results, we achieve 81.5% AUC. For each treatment, we also create a series of regression models that are based on different supervised machine learning algorithms. We are able to estimate with average $R^2 = 0.801$ the outcome of interest; the time from diagnosis to a potential adverse event (TAE). Leveraging combinations of these models, we present ML4CAD, a novel personalized prescriptive algorithm. Considering the recommendations of multiple predictive models at once, the goal of ML4CAD is to identify for every patient the therapy with the best expected TAE using a voting mechanism. We evaluate its performance by measuring the prescription effectiveness and robustness under alternative ground truths. We show that our methodology improves the expected TAE upon the current baseline by 24.11%, increasing it from 4.56 to 5.66 years. The algorithm performs particularly well for the male (24.3% improvement) and Hispanic (58.41% improvement) subpopulations. Finally, we create an interactive interface, providing

physicians with an intuitive, accurate, readily implementable, and effective tool.

Keywords Precision Medicine · Personalization · Coronary Artery Disease · Machine Learning · Prescriptions

Highlights

- We present the first prescriptive methodology that utilizes electronic medical records and machine learning to provide personalized treatment recommendations for the management of coronary artery disease patients.
- We introduce a new quantitative framework to evaluate the performance of prescriptive algorithms.
- We show that our data-driven approach can substantially improve patient outcomes, increasing the average time to an adverse event by 13 months for the overall population.
- We provide an online user-friendly application that is available to physicians where the algorithm suggestions can be tested in real time.

1 Introduction

The clinical condition of Coronary Artery Disease (CAD) also referred to as ischemic heart disease, is present when a patient presents one or more symptoms or complications from an inadequate blood supply to the myocardium (Fuster et al., 1992). This is most commonly attributed to the obstruction of the epicardial coronary arteries due to atherosclerosis (Ross, 1999). CAD remains the number one cause of death in the United States, accounting for over 360,000 annual casualties (AHA, 2017). CAD is mostly prevalent in older patients (above the age of 50 years) in the form of a chronic condition which requires a principal intervention and subsequent systematic medical therapy and monitoring (Fuster et al., 1992). The primary care of patients with CAD includes ascertainment of the diagnosis and its severity (with non-invasive and/or invasive imaging), control of symptoms, and therapies to improve survival (Hansson, 2005). The mainstay of treatment is medical therapy. The latter may or may not be combined with coronary revascularization (either Coronary Artery Bypass Graft (CABG) surgery or Percutaneous Coronary Intervention (PCI)) in an effort to slow the progress of the disease and relieve its symptoms. Considering the magnitude and the repercussions of CAD, the importance of medical therapy to reduce its symptoms and prolong life expectancy is being increasingly recognized (Sedlis et al., 2015).

There has been growing interest in using clinical evidence to understand the effects of treatments in patients with CAD. Nowadays, there are numerous evidence-based clinical guidelines for CAD management (Fihn et al., 2015, 2014) and angiographic tools for grading its complexity, such as the SYNTAX Score (Serruys et al., 2009; Sianos et al., 2005). However, it is not clear how to choose among different types of available therapies (pharmacological, percutaneous intervention, and surgery) to maximize effectiveness at an individual level. This is likely due to the multitude of parameters that define the form of the disease for each patient and the uncertainty that lies behind an individual patient’s response to a particular treatment (Warnes, 2017). One of the greatest challenges in developing evidence-based guidelines applicable to large populations is paucity of information about special subpopulations with unique characteristics. This is attributed to the absence of specialized clinical trials (Fihn et al., 2014).

Considering the challenges and the significance of CAD, a personalization approach may greatly impact the effective management of the disease. Personalization is the problem of identifying the best treatment option for a given instance, i.e., a display add (Zhou et al., 2008) or medical therapy (Lesko, 2007). There are two main challenges for designing personalized prescriptions for a patient as a function of the features recorded in the data:

1. While the outcome of the administered treatment for each patient is observed, the counterfactual outcomes are unknown. That is, the outcomes that would have occurred had another treatment been administered. Note that if this information were known, the prescription problem would reduce to a multi-class classification problem. Thus, the counterfactual outcomes need to be inferred.
2. In the data, there is an inherent bias that needs to be taken into account. The nature of data from Electronic Medical Records (EMR) is observational as opposed to data from randomized trials. In a randomized trial setting, patients are randomly assigned different treatments, while in an observational setting, the assignment of treatments potentially depends on features of the population.

1.1 Literature Review

Our objective is to solve the problem of prescribing the best option among a set of predefined treatments to a given patient as a function of the samples’ features. We are provided with observational data of the form $\{(\mathbf{x}_i, y_i, z_i)\}_{i=1}^n$, comprising n observations. Each data point $\{(\mathbf{x}_i, y_i, z_i)\}$ is characterized by features $\mathbf{x}_i \in \mathbb{R}^p$, the prescribed treatment $z_i \in [T] = \{1, \dots, T\}$, and the corresponding outcome $y_i \in \mathbb{R}$. We denote $y(1), \dots, y(T)$ the T “possible outcomes” resulting from assigning each of the T treatments respectively.

A similar question has been studied in the causal inference literature. In this setting, the main focus lies on observational studies to identify causal relationships between an intervention and outcomes in a particular population (Pearl et al., 2009). Introduced by Neyman and popularized by Rubin, the Potential Outcomes Framework uses a probabilistic assignment mechanism to mathematically describe how treatments are given to patients. It also accounts for a potential dependence on background variables and the potential outcomes themselves (Rubin, 1990; Angrist et al., 1996). More

specifically, it focuses on the case where $S = \{C, T\}$ (treatment and control). For each patient i , the potential outcome $y_i(T)$ is the experienced outcome if exposed to treatment T . The causal effect of T compared to C is then computed as $\delta_i := y_i(T) - y_i(C)$. Thus, causal effects are solely defined for one treatment relative to another and only if the individual could have been reasonably exposed to both. The fundamental problem of causal inference is that $(y_i(T), y_i(C))$ are not jointly observable. That is, only one observed response is present depending on the treatment assignment. As a result, Rosenbaum and Rubin (1983) focus on the average treatment effect for a completely randomized experiment. This scenario considers the difference of the sample means for the units receiving the treatment and control.

$$\text{ATE} = \frac{1}{n_T} \sum_{j: z_j=T} y_j(T) - \frac{1}{n_C} \sum_{j: z_j=C} y_j(C). \quad (1)$$

However, in observational studies, treatment assignment is not independent of the potential outcomes. Thus, further analysis is required to account for latent differences between the treated and control groups on the basis of observed covariates X (inverse probability weighting, propensity score matching, nonparametric regression, etc.) (Rosenbaum, 2010).

Causal effect approaches do not provide personalized estimations of the treatment effect for each unit since they focus on the aggregate population level. A personalized prescription methodology would require a quantification of the impact of each regimen for every individual in isolation. This is the essence of the personalized medicine field (Hamburg and Collins, 2010): identifying the optimal therapy for a particular set of phenotypic and genetic patient characteristics. Machine learning (ML) algorithms are expected to enable the utilization of rich datasets. They could provide improved solutions for patients by learning the outcome function for each treatment. They will particularly impact those that belong to very specific subgroups and respond in unusual ways to the available treatments (Frohlich et al., 2018).

A common approach in the literature to leverage these algorithms is called “Regress and Compare”. It identifies the expected effect $y_i(z_i)$ of treatment $z_i \in [T]$ for each patient i based on the covariates \mathbf{x}_i and consequently prescribes the regimen with the best potential impact;

$$\max_{z_i \in [T]} y_i(z_i | \mathbf{x}_i) \quad \forall i \in [n],$$

where $[n]$ is the set of patients in the sample. The “Regress and Compare” methodology follows this paradigm, choosing a treatment by maximizing among T regression functions. A different regression model is fitted to the subset of the data that received each treatment. It subsequently uses them to predict outcomes and pick the one with the more optimistic prediction (Stoehlmacher et al., 2004). This approach has been historically followed by several authors in clinical research (Feldstein et al., 1978), and more recently by researchers in statistics (Qian and Murphy, 2011) and operations research (Bertsimas et al., 2017). The online version of this problem, called the contextual bandit problem, has been studied by several authors (Li et al., 2010; Goldenshluger and Zeevi, 2013) in the multi-armed bandit literature (Gittins et al., 1989). Even though it is intuitive, this methodology is subject to prediction errors and potential biases of a single method.

In the field of precision medicine, Bertsimas et al. (2017), first, introduced a personalized prescriptive algorithm for diabetes management that harnesses the power of EMR. It was based on a “Regress and Compare” k nearest neighbors (k -NN) approach. This methodology yielded substantial improvements in patient outcomes relative to the standard of care. Moreover, it provided physicians with a prototyped dashboard visualizing the algorithm’s recommendations. Their work showed that tailored approaches to particular diseases coupled with medical expertise provide the medical community with highly accurate and effective tools that will ameliorate patient treatment. Even though this effort provided promising results, the k -NN approach is not applicable to diseases where the effects of a treatment are not promptly observable. The same individual was tracked via multiple visits in the hospital system. Thus, the algorithm suggested alterations in the medication only when there was significant reduction on the expected Hemoglobin A1c measurement. The physician could measure the effectiveness of a treatment by ordering a blood test in the near future. On the contrary, at the CAD setting the adverse effects of the disease are observed in the span of ten years from the time of diagnosis.

Focusing mostly on the personalization and not the prediction objective, Kallus (2017) proposes a recursive partitioning methodology for personalization using observational data. This new algorithm is tailored to optimize a personalization impurity measure. As a result, it hardly places

any emphasis on the predictive task. Therefore, it raises questions regarding the accuracy of the suggested treatment effect. Bertsimas et al. (2019) modify the latter’s objective to account for the prediction error, and use the methodology of Bertsimas and Dunn (2019, 2017) to design near optimal trees, improving performance substantially. Continuing on tree based approaches, Athey and Imbens (2016), and Wager and Athey (2018) also use a recursive splitting procedure of the feature space to construct causal trees and causal forests respectively. They estimate the causal effect of a treatment for a given sample, or construct confidence intervals for the treatment effects. However, they do not infer explicit prescriptions or recommendations. Also, causal trees (or forests) are designed exclusively for studies comparing binary treatments.

In the cardiovascular field, the benefit of ML based personalization methods has been recognized and is expected to play a significant role in facilitating precision cardiovascular medicine (Krittanawong et al., 2017). Nevertheless, in the case of CAD, personalization approaches have been primarily focused on utilizing genomic information (Beitelshees, 2012), and not on employing EMR and ML. Since 2014, the US mandated all public and private healthcare providers to adopt and demonstrate “meaningful use” of EMR to maintain their existing Medicaid and Medicare reimbursement levels. This decision contributed to the creation of clinical databases that contain in-depth information for many patients. These data can be leveraged using ML to construct models and algorithms that can learn from and make predictions on data (Ron Kohavi, 1998).

One of the greatest challenges of EMR is the presence of right censored patients (Lagakos, 1979; Imbens and Rubin, 2015), which arises when a patient disappears from the database after diagnosis and treatment of the disease. Traditional approaches to address right censoring, including the Cox proportional hazards model (Cox, 1972) or the Weibull Regression (Ibrahim et al., 2014), do not allow for time-varying effects of covariates. Their weaknesses are especially relevant to datasets that span over long periods of time, providing results that are not validated by the medical literature (e.g. positive correlation between a patient’s BMI and his/her expected time to adverse event).

Our work addresses most of the challenges encountered in the personalized prescription setting

that uses EMR, including counterfactual estimation and censoring.

1.2 Contributions

In this paper, our objective is to find the best primary treatment for a CAD patient to maximize the time from diagnosis to a potential adverse event (TAE) (myocardial infarction or stroke). We consider the latter as the primary endpoint of our models. Our dataset includes CAD patients who were administered treatment through the Boston Medical Center (BMC), a private, not-for-profit, 487-bed, academic medical center located in Boston, MA, USA. We retrieved each patient’s medical history, the primary treatment followed after diagnosis, and the most recent clinical examination results to the time of diagnosis. We considered five primary prescription approaches available for each patient. We developed predictive and prescriptive algorithms that provide personalized treatment recommendations. We propose a new prescription algorithm to assign the regimen with the best predicted outcome leveraging simultaneously multiple regression models. The effect of the prescriptive algorithm was evaluated by comparing the expected TAE under our recommended therapy with the observed outcome prescribed by physicians at the medical center. Successful treatment recommendations increase the TAE. On the contrary, ineffective prescriptions negatively impact the patient, decreasing the time from diagnosis to a myocardial infarction or stroke. We tested the robustness and effectiveness of our methodology. We considered different ground truths regarding the treatment effect of a given therapy to a patient. The ground truths comprise the standard of care as well as combinations or individual predictions from ML models. The main contributions of this paper are:

1. A new methodology to treat right censored patients that utilizes a k -NN approach to estimate the true survival time from real-world data.
2. Interpretable and accurate binary classification and regression models that predict the risk and timing of a potential adverse event for CAD patients. We selected a diverse set of well-established supervised machine learning algorithms for these tasks.
3. The first prescriptive methodology that utilizes EMR to provide treatment recommendations for CAD. Our algorithm, ML4CAD (Machine Learning for CAD), combines multiple state-of-the-art ML regression models with clinical exper-

tise at once. In particular, it uses a voting scheme to suggest personalized treatments based on individual data.

4. A novel evaluation framework to measure the out-of-sample performance of prescriptive algorithms. It compares counterfactual outcomes for multiple treatments under various ground truths. Thus, we assess both the accuracy, effectiveness, and robustness of our prescriptive methodology. Using this evaluation mechanism, we demonstrate that ML4CAD improves upon the standard of care. Its expected benefit was validated by all considered ground truths and TAE estimation models.
5. An online application where physicians can test the performance of the algorithm in real time bridging the gap with the clinical practice.

The structure of the paper is as follows. In Section 2, we describe the data used to train and validate our methods. In Section 3, we outline the method used to handle the challenge of censoring. Section 4 describes the methods and results of the binary classification models, and similarly Section 5 refers to regression. In Section 6, we present the personalized prescription algorithm and its evaluation framework. Results under different ground truths and recommendation policies are compared in Section 7. We conclude our work in Section 8. We provide a list of all the abbreviations definitions in alphabetical order in the Appendix.

2 Data

In this section, we provide detailed information about the dataset under consideration. We outline the patient inclusion criteria as well as a description of the covariates included in the ML models. Subsequently, we refer to the treatments identified from the EMR and their aggregation as features for our algorithms. We also present the missing data imputation procedure that was followed.

Sample Population Description

Through a partnership with the BMC we obtained EMR for 1.1 million patients from 1982 to 2016. In this dataset, 21,460 patients met, at least, one of the following inclusion criteria:

- **Population 1:** Patients associated with CAD risk of at least 10% based on the Framingham Heart Study formula (Wilson et al., 1998)

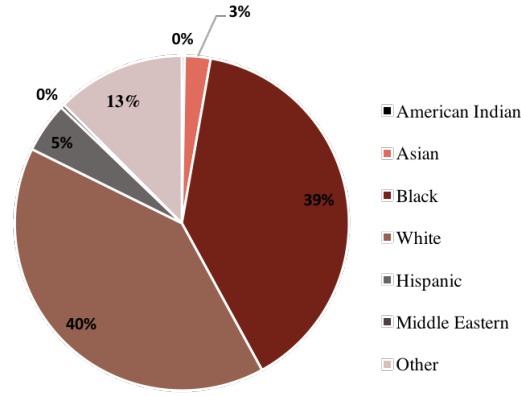
who were prescribed antihypertensive medication as primary treatment. The 10% threshold was selected since it is considered one of the primary indications for physicians to prescribe CAD treatment to their patients (Wilson, 2017);

- **Population 2:** Patients who were administered at least one CABG surgery or, at least, one PCI and were prescribed antihypertensive medication;

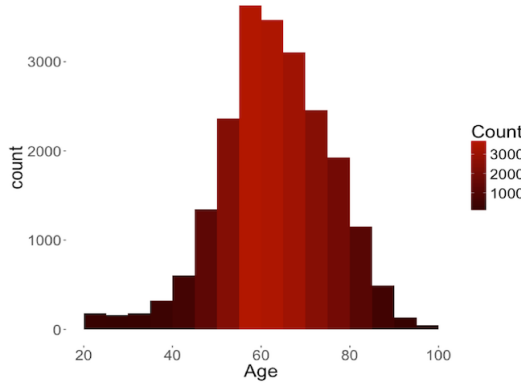
We used the conditions outlined above due to the absence of a systematic CAD diagnosis code in the system (Strom, 2001). Note that the two inclusion criteria are mutually exclusive as a primary CAD prescription could either involve exclusively pharmacological treatment or a drug combination with a CABG surgery or a PCI. All patient EMR were processed to identify the time t_0 that corresponds to the point of initial diagnosis prior to any coronary revascularization. We reverted to the record that corresponds to this time to create the patient features X . Thus, we avoided the inclusion of two populations whose conditions are fundamentally dissimilar. Our sample comprised recently diagnosed CAD patients, similar to the ones physicians encounter in practice. We identified, using the totality of the EMR after the time t_0 , the main therapy prescribed to each patient while being in the system. Notice that every member of the sample population was medicated with antihypertensive drugs. If in addition to the pharmacological therapy they were administered surgical or percutaneous interventions, we set the latter as the main treatment administered by the hospital.

BMC patients come predominantly from underprivileged socioeconomic backgrounds. As a result, in most cases they do not have the financial capability to support alternative health providers. They need to appeal to the BMC for healthcare services for the majority of their medical needs. Thus, most of their EMR are concentrated in the same database, allowing us to follow the trajectory of each patient's health from a single source. The ethnicity and age distributions of the population are depicted in Figures 1a and 1b, respectively.

We excluded all patients whose diagnosis date was identical to their last observation in the healthcare system. Moreover, we removed from the data those whose cause of death was observed but not related to heart disease (e.g., cancer non-survivors). We retrieved for each patient a set of values that describe their demographics, medical therapy, and clinical characteristics at the time of diagnosis t_0



(a) Ethnicity distribution.



(b) Age distribution.

Fig. 1: Demographic Characteristics of the population

(Table 1). We used ICD-9, CPT, and hospital specific codes to identify the corresponding records as well as lab test results for particular measurements (i.e., low-density lipoprotein (LDL) or high-density lipoprotein (HDL) levels). Along with demographic information, we included features that are considered risk factors for heart disease, according to the medical literature. We excluded all covariates whose values were not known for at least 50% of the patients in the dataset. Further information regarding the characteristics of the overall population, as well as split by training, validation, and testing set are available in the Supplemental Material. We identified an adverse event (myocardial infarction or stroke) attributable to CAD and recorded the date of occurrence. This way, we define the time between a diagnosis and an adverse event. In case the patient disappeared from the EMR before the lapse of 10 years after diagnosis, we recorded that the patient was right censored. We did not take into account the severity of the adverse event in our evaluation.

Category	Variable Name	% NA
Demographics	Age	0.0%
	Gender	0.0%
	Ethnicity	0.0%
	Language	0.0%
	Marital Status	15.3%
	Ethnicity	0.0%
Treatment	ACE inhibitors	0.0%
	Adrenergic Receptors	0.0%
	Angiotensin Agonists	0.0%
	Antiarrhythmics	0.0%
	Blockers	0.0%
	(beta, alpha, etc.)	0.0%
	CABG	0.0%
	Cardiac Glycosides	0.0%
	Diuretics	0.0%
	Lipid Lowering medication	0.0%
	Muscle relaxants	0.0%
	Nitrates	0.0%
	Other antihypertensive	0.0%
	PCI	0.0%
	Phosphodiesterase inhibitors	0.0%
	Statins	0.0%
Family History	Diabetes	26.8%
	Hypertension	23.9%
Medical Records	Body Mass Index (BMI)	16.6%
	LDL Cholesterol	21.4%
	HDL Cholesterol	21.3%
	Diastolic Blood Pressure	7.1%
	Systolic Blood Pressure	7.1%
	Diabetes	0.5%
Observed Behavior	Smoking	23.6%
	Time observed in the EMR database	0.0%

Table 1: Patient characteristics considered. The column “% NA” indicates the percent of missing data that was present in the original dataset

Treatment Options

We considered five primary options for each patient, shown in Table 2. These options are mutually exclusive and thus each patient received only one of them as primary treatment. CAD is a chronic disease whose management may differ across time. However, we noticed that a certain pattern was followed for the vast majority of the patients throughout their presence in the academic medical center. Coronary revascularization is a major operation and thus we distinguish CABG and PCI as separate treatment categories. In agreement with the general guidelines of the American Heart Association for the management of Stable Ischemic Heart Disease (Fihn et al., 2014), most of the patients are prescribed blocking medication to treat hypertension and statins as a lipid lowering treatment. Therefore, we chose combinations of those

two lines of therapy as primary prescription options. Nevertheless, the pharmaceutical treatment for a CAD patient may include not only blockers, but also a more complicated combination of drugs, depicted in Table 3 under “Treatment”. As the set of all those combinations is too wide, we considered only the most common prescription options. We did not account for aspirin (ASA) since all patients were prescribed this line of therapy.

Note that we did not consider ACE inhibitors as a prescription option because they usually accompany another type of antihypertensive medication for CAD patients (Rejnmark et al., 2006). They are prescribed in combination to blockers or as a substitute of the latter in cases where a patient has some prohibitive medical condition to the former. Thus, the majority of the population that belongs in the “Drugs 2 and 3” categories are effectively under ACE inhibitors. The latter drug class was administered in less than 50% of the sample population. As a result, a separate pharmacological treatment option would thin the training sets presented in the following sections significantly.

Option	Description	Num. of patients	%
CABG	Coronary Artery Bypass Graft Surgery with pharmaceutical treatment	1854	8.64%
PCI	Percutaneous Coronary Intervention with pharmaceutical treatment	4042	18.85%
Drugs 1	Pharmaceutical treatment including blockers and statins	6833	31.86%
Drugs 2	Pharmaceutical treatment including blockers and excluding statins	3767	17.56%
Drugs 3	Pharmaceutical treatment excluding blockers (potentially including statins)	4964	23.09%

Table 2: The Prescription Options.

Handling of missing values

We collected each patient’s medical records (lab test results and clinical measurements) associated with the most recent clinical examination before or at the time of diagnosis. We omitted from our analysis any risk factors whose missing values proportion was higher than 50% (i.e., ejection fraction, ECG measurements). Table 1 shows the percent

Treatment Name	Proportion
ACE inhibitors	46.12%
Adrenergic Receptors	6.38%
Angiotensin Agonists	13.62%
Antiarrhythmics	13.65%
Blockers (beta, alpha, etc.)	68.03%
CABG	7.01%
Cardiac Glycosides	2.45%
Diuretics	47.90%
Lipid Lowering medication	5.29%
Muscle relaxants	4.81%
Nitrates	77.02%
Other antihypertensive	11.37%
PCI	19.60%
Phosphodiesterase inhibitors	3.59%
Statins	58.78%

Table 3: The percentage of the overall population that received each treatment based on the sample population. Note that the same patient may have been prescribed multiple treatments.

of missing data that was present in the original dataset. Note that all demographic variables other than Marital Status were consistently recorded for all patients. A treatment was considered to be present if there was an active prescription for the patient in the EHR. If there was no record of a treatment, we assumed that the patient was not administered the specific medication. Thus, the missing percentage for all treatments is 0.0%. Family history and smoking habits were available in the database for only a portion of the patients. Continuous features, such as cholesterol and blood pressure levels, were extracted from the vitals and lab tests records.

We imputed missing values using `opt.cv`, the state-of-the-art ML algorithm proposed by Bertsimas et al. (2018). Given that the underlying pattern of missing data was not known, we opted for a method whose performance remained consistent across different types of “missingness”. In (Bertsimas et al., 2018), the authors demonstrated on 84 data sets that the accuracy of their algorithm relative to benchmark ones does not appear to differ drastically between the missing completely at random (MCAR) and not missing at random (NMAR) patterns. The latter constitutes the most common type of missing data in health care applications, as values are not usually randomly incomplete for reasons such as missed study visits, patients lost to follow-up, missing information in source documents, and lack of availability among others. We created artificial missing data under the NMAR mechanism and compared `opt.cv` with other well-established missing data imputation techniques in

our dataset. We evaluated the resulting imputation error and the effect on downstream predictive performance for the binary classification task. Our results showed that `opt.cv` provided an edge across all metrics considered. Thus, it was selected as the imputation algorithm for the independent covariates of both the binary classification and regression models (see Table 3 of the Supplemental Material).

3 Estimating time to adverse event for right censored patients

In censored datasets the outcome of interest is generally the time until an event (onset of disease, death, etc.), but the exact time of the event is unknown (censored) for some individuals. When a lower bound for these missing values is known (for example, a patient is known to be alive until at least time t) the data is said to be right censored. In our dataset, we considered the time of censoring to be the last event-free visit of the patient to the academic medical center. Thus, for each patient i where $t_i < 10$ (years) and no adverse event (stroke/heart attack) has been recorded, we set the censoring time $c_i = t_i$, the last time observed in the EMR. Our sample was comprised of 13,498 censored observations (62.9% of the overall population).

Methods from the survival analysis literature are usually employed in the presence of censored populations. A common survival analysis technique is the Cox proportional hazards regression (Cox, 1972) which models the hazard rate for an event as a linear combination of covariate effects. Although this model is widely used and easily interpreted, its parametric nature makes it unable to identify non-linear effects or interactions between covariates (Bou-Hamad et al., 2011).

We propose a data-driven methodology that utilizes a k -NN approach to identify patients with similar outcomes and known trajectories based on their covariates. We consider the set A (B) of patients that had (did not have) an adverse event within 10 years. Note that within set B the EMR indicate that no adverse event occurred within the defined time frame. Let C be the set of censored patients that did not have an adverse event within a time t_c (less than 10 years) and they disappear from the EMR after t_c . It is not known whether they experienced an adverse event within 10 years or not. In order to estimate the TAE for patient

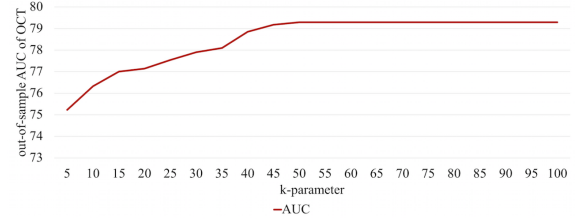


Fig. 2: Graph of a Cross-validation results for the selection of the k parameter for the k -NN model.

X in the set C , we consider patients within $A \cup B$ such that:

1. They have the same gender as X . It has been recognized that women form a distinct subpopulation within patients with CAD (Roeters van Lennep et al., 2002).
2. They belong to the same age group as X . Age at time of diagnosis plays a major role in the development and the effects of CAD (Wilson et al., 1998).
3. Their ground truth outcome metric is greater or equal to the censoring time of X . The patient will potentially experience an adverse event after the censoring time t_c .

Based on the Euclidean distance across the patient specific factors depicted in Table 1 (factors with continuous values were normalized to have zero mean and standard deviation of one), we find the k -nearest neighbors of X within the cohort outlined. We assign to the censored patient X the average time to adverse event of their k -nearest neighbors. We used cross-validation to set the parameter $k = 50$. The outcome of interest was the area-under-the-curve (AUC) performance of the binary classification model presented in Section 4 (Figure 2). We selected the value of the unsupervised learning model parameter according to the performance of the binary classification model on the 10-year risk task. Our method allows us to build for every censored patient a unique cluster of k -NN, introducing a personalization aspect in the estimation of TAE.

Our k -NN algorithm's performance is $R^2 = 0.81$ according to the following process:

1. Select a sample of the population which was not censored (the TAE t_i is known).
2. Artificially generate a censoring time t_i^c , sampled uniformly across the interval $[1, t_i]$ corresponding to a day in the 10 year time frame.
3. Apply the k -NN algorithm to estimate the TAE and compare the results with the ground truth that is known.

We impute the outcomes of 13,679 censored observations, following this approach. We create a complete dataset that is further used for the creation and validation of the predictive and prescriptive models. The inclusion of the censored patients permitted a higher sample size for the binary classification and regression models that led to more accurate and stable results (see Tables 4 and 5 of the Supplemental Material). The exclusion of such cases would restrict the overall population to only 7,962 observations, limiting the downstream predictive performance of the models.

4 The Binary Classifications Models

The first problem we addressed is the creation of personalized risk prediction models for CAD patients. Our binary outcome of interest is the occurrence of an adverse event (stroke or heart attack) within a 10-year time period. This time frame is in accordance with the vast majority of established CAD risk calculators (Goff et al., 2014; D’agostino et al., 2008; Ridker et al., 2007). The medical community recognizes the chronic nature of the disease and as a result it focuses on evaluating its impact on the health of the patient over a long-term horizon. Both the American Heart Association and the American College of Cardiology annually update their guidelines on the primary prevention of cardiovascular disease releasing new versions of 10-year CAD risk scores (Arnett et al., 2019). Although this time frame is challenging and the health condition can significantly change over years, we decided to follow the paradigm of the existing literature. Moreover, we present corresponding results for two and five year horizons in Table 6 of the Supplemental Material. Thus, a comparison of different time windows is available to the reader for comparison.

We apply state-of-the-art ML algorithms to the data and compare their out-of-sample performance on the testing set. Table 4 provides a summary of the results for Logistic Regression, Random Forest (Breiman, 2001), Boosted Trees (Chen and Guestrin, 2016), CART (Breiman et al., 1984), and Optimal Classification Trees (OCT) (Bertsimas and Dunn, 2017, 2019).

We split the $n = 21,460$ patients in 75% for Training and Validation and 25% for Testing, using $p = 31$ patient characteristics (Table 1). Our sample includes all censored observations whose values were imputed using the methodology described in Section 3. These observations were not excluded as

a higher sample size improved the model’s out-of-sample performance. A higher sample size had a significant positive effect on the downstream performance of the binary classification models. We evaluated the predictive power of the algorithm under additional random splittings of the data. Thus, we ensured that the evaluation of the global algorithm was not sensitive to a particular split of the dataset.

L_2 regularization was used for the logistic regression model and 10-fold-cross-validation was employed to set the hyper-parameters of each method. In the case of OCT and CART, we tuned the complexity parameter, the maximum depth, and minimum bucket. Based on cross-validation results, the number of greedy trees used for the Random Forest model was set to 500.

Our objective was to create an accurate model that would have high chances of affecting the medical practice. Even though there has been a steep increase in publications that utilize artificial intelligence and ML in the field of medicine, only a small proportion of those models have been integrated into the healthcare system (Emanuel and Wachter, 2019). Clinicians need actionable insights and guidelines they can explain and understand (Nevin et al., 2018). Algorithms have to satisfy this condition. Otherwise, the final outputs of these methods do not actually impact the patients. The FDA (2017) validated such concerns by mandating the use of interpretable ML models when it comes to medical decision making.

For this reason, we decided to focus on the model of the Optimal Classification Trees (OCT) algorithm, which was proposed by Bertsimas and Dunn (2017), see also (Bertsimas and Dunn, 2019). Its tree structure accounts for non-linear interactions among variables providing an edge compared to Logistic Regression. This new supervised learning method uses modern mixed-integer optimization techniques to form the entire decision tree in a single step, allowing each split to be determined with full knowledge of all other splits. The OCT algorithm creates the final model in a holistic manner yielding better performance than traditional decision tree approaches, such as CART (Table 4). It increases interpretability due to its tree form which allows predictions through a few decision splits on a small number of high-importance variables. Thus, physicians are able to associate a risk profile to each patient that comprises up to seven risk factors even if the entire dataset includes a significantly higher number of features. This property

	Out-of-sample AUC	In-sample AUC	Out-of-sample Accuracy	In-sample Accuracy
OCT	81.54%	81.35%	81.45%	81.36%
CART	73.33%	72.66%	80.23%	80.12%
Random Forest	84.29%	83.29%	81.88%	82.35%
Logistic regression	80.83%	82.21%	80.55%	80.98%
Boosted Trees	81.43%	82.76%	81.03%	81.27%
Baseline			73.51%	73.51%

Table 4: Results of the different ML algorithms used to predict the occurrence of an adverse event within 10 years after diagnosis. We consider as Baseline the simple model that predicts that all patients will experience an adverse event. Accuracy is measured using a probability above 50% as the threshold. The term “Out-of-sample” signifies the performance of the model on the Test set and “In-sample” on the Training set.

is not inherently shared by other well-established non-linear algorithms, such as Random Forest or Boosted Trees. As a result, the users cannot easily attribute changes in the patient’s estimated risk to specific model variables. To address this challenge, complementary frameworks, like the SHapley Additive exPlanations approach (Lundberg and Lee, 2017), are needed to explain the output of these machine learning models.

Random Forest (84.29%) yields better AUC results compared to OCT (81.54%), although quite similar in terms of accuracy for a fixed threshold (81.88%, 81.45% respectively). However, Random Forest grows multiple decision trees and assigns for each observation the class that is indicated by the majority of the decision trees. OCT provides us with a single tree whose branches can be easily explained to physicians. Each path leads to comprehensible clinical decision rules that could positively affect the cardiovascular practice. Its model achieves superior performance in both accuracy and AUC when compared to all other ML methods, including the advanced ensemble algorithm of Boosted Trees. Moreover, Logistic Regression (80.83% AUC) is more accurate compared to CART (73.33% AUC), but slightly under-performing with respect to more sophisticated algorithms (81.43% AUC).

The final OCT model is depicted in Figures 3, 4, 5. Table 5 presents its ten most significant variables. An analysis of the most predictive features follows below:

- **Time in the System** (TimeinSystem): the time that the patient has been observed in the BMC database (from the first record until time of diagnosis t_0). It serves as an indicator of their medical condition and history information depth. TimeinSystem does not incorporate

any patient details after the time t_0 , avoiding the inclusion of survivorship bias in the data. As shown in Figures 3, 4, 5, higher values of the TimeinSystem variable are associated with leaves that predict positive outcomes for the patient. This result indicates that physicians are more effective when they have extensive amount of information available and follow their patients’ trajectories over longer periods of time.

- **Prescription of Medication** (Nitrates/ Beta Blockers/ Statins/ ACE Inhibitors): whether a patient has been systematically treated with one particular type of medication. Depending on the decision path of the tree, the risk of an adverse event might increase or decrease if the medication has been prescribed. There need not be a causality relation for the changes in risk. Only association can be deduced from such a model. However, these results reinforce the argument that personalization in the treatment can indeed affect the survival of the CAD population.
- **CABG/PCI**: whether the patient has performed a revascularization procedure. We notice that positive values in these two variables are associated with leaves that suggest pessimistic patient prognoses. Diagnosed CAD patients with more severe symptoms of atherosclerosis are usually suggested to perform at least one of these interventions (CABG, PCI) (Fihn et al., 2014).
- **Patient Age at Diagnosis**: the age of the patient at the time of diagnosis in the EMR system. Across the model we notice that older populations are associated with higher risk, confirming a wide range of CAD risk calculators published in the medical literature (Conroy et al.,

2003; Polonsky et al., 2010; D’agostino et al., 2008).

- **HDL (mg/dL) levels:** the HDL (mg/dL) levels from a blood test conducted at the time of diagnosis. Depending on the position of the split in the tree, higher levels of HDL may positively or negatively impact the ten year risk of CAD.
- **Median Systolic Blood Pressure:** the median of the systolic blood pressure measurements recorded in the EMR across all visits in a window of three months before t_0 . We consider the median due to the noise frequently encountered in systolic blood pressure measurements (Tucker et al., 2017; Epstein, 2014; Duan et al., 2019).

Feature	Importance
Time in the System	27.40%
Prescription of Nitrates	19.80%
Prescription of Beta Blockers	15.01%
PCI operation	12.96%
Prescription of Statins	10.53%
CABG operation	3.23%
Patient Age at Diagnosis	2.87%
Prescription of ACE inhibitors	1.86%
HDL (mg/dL) levels	1.31%
Median Systolic Blood Pressure	1.06%

Table 5: Demonstration of the independent variable ranking in the OCT binary classification model. The importance of each variable is measured as the total decrease in the loss function as a direct result of each split in a tree that uses this variable. The results are normalized so that they sum to one.

4.1 Analysis of characteristic decision paths

We analyze distinctive risk profiles from the OCT model that provide interesting insights for the management of CAD patients.

- **Paths 1 & 2:** Contain samples whose presence in the EMR was recorded only for two months before the diagnosis. Leaf 1 refers to patients that are administered a PCI operation and leaf 2 to those who perform a CABG surgery. Both paths associate extremely high risk to the corresponding population.
- **Paths 3 & 4:** Refer to individuals who are present in the BMC system at least seven years.

They are not treated with PCI, neither with beta blockers nor statins. Their baseline risk of an adverse event is 7.78%. However, this risk differs depending on the age group they belong. Specifically, those individuals under 68 years old have 1.45% probability of having a stroke or heart attack over the next ten years. On the contrary, older patients have 18.11% chance of experiencing an adverse event.

- **Paths 5 & 6:** Include patients who are present in the BMC system for at least two months and are prescribed PCI but no CABG surgery. They are not treated with beta blockers nor statins and their blood glucose levels are lower than 149 mg/dL. Their baseline risk of an adverse event is 12.53%. This risk differs again depending on the age group they belong. Specifically, those under 57 years old have 95.19% probability of avoiding a stroke or heart attack over the next ten years. On the contrary, patients older than 57 years of age have 14.03% chance of experiencing such an event.

5 The Regression Models

Predicting the risk of an adverse event within a 10-year time frame is an important question that we address in Section 4. However, a personalized prescriptive algorithm requires the creation of accurate regression models that, given the condition of a patient, estimate the exact TAE for each potential treatment. We leveraged various state-of-the-art ML methods, both interpretable and non-interpretable, to generate a set of estimations at an individual level (Breiman et al., 1984; Breiman, 2001; Bertsimas and Dunn, 2017, 2019; Chen and Guestrin, 2016). We trained a separate model for each combination of method and treatment using as sample population patients that exclusively received this regimen. For example, we applied the Random Forest algorithm to generate five predictive models that correspond to CABG, PCI, Drugs 1, 2, and 3. We followed the same process for CART, Linear Regression, Boosted Trees, and Optimal Regression Trees (ORT). As in the classification task, we applied 10-fold-cross-validation to determine the hyper-parameters of each model, including the complexity parameter, the maximum depth, and minimum bucket for ORT and CART. Based on the cross-validation results for the regression task, the number of greedy trees for the Random Forest model was set to 250 in contrast to 500 that were chosen for the binary classification outcome. We

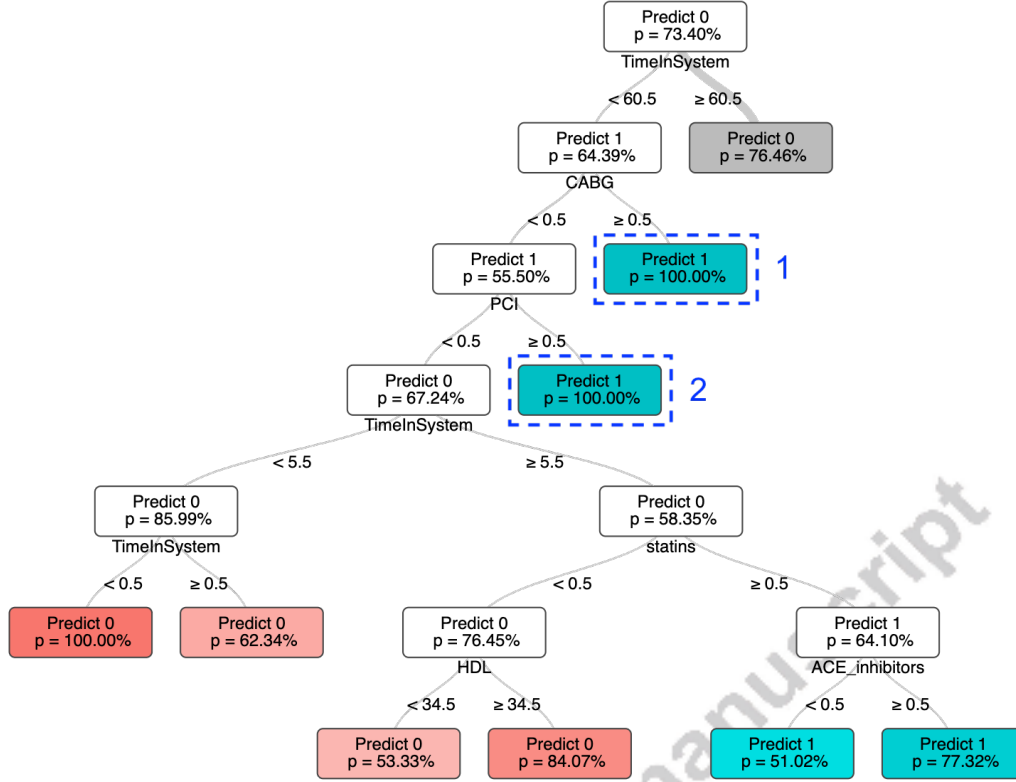


Fig. 3: Visualization of the first part of the OCT model. Paths 1 and 2 are indicated with blue dashed rectangular frames. Shaded nodes include a collapsed subset of the tree model.

used L_2 regularization for the linear regression model. Table 6 provides a summary of each method’s out-of-sample performance for every treatment option in terms of the R^2 metric.

The results from Table 6 indicate that Random Forest outperforms the other methods in all tasks in terms of the R^2 metric. CART, on the other hand, appears as the least performing method across all tasks. ORT have an edge over the greedy tree-based approach, other than in the case of category “Drugs 3”. We observe that Linear Regression and Boosted Trees have comparable performance for all types of treatment. We will leverage all these models as the main component of our prescriptive algorithm, presented in Section 6.

We created separate models for each treatment population to avoid biases in the prediction due to the existing treatment prescription patterns in the EMR (Gianfrancesco et al., 2018). Our goal was to identify, for each patient, what is the therapy that would maximize their TAE. Therefore, a distinction was needed between the different populations that received each treatment option. The existing regimen allocation process could have significantly biased the prescriptive algorithm if included as

an independent feature in the set of covariates X (Schulz et al., 1995). For instance, if physicians in BMC prescribed CABG only to the younger population, the ML model would not have been able to distinguish between the effect of CABG and the age of the patient.

6 ML4CAD: The Prescription Algorithm

The regression models serve as the basis for the prescription algorithm, utilizing the point predictions as counterfactual estimations. The objective of the prescription algorithm is to understand the potential effect of every therapy that each patient would have experienced, had it been prescribed to them. For example, knowing the outcome of patient X who received CABG surgery, we aim to estimate the outcome metric of a PCI intervention and for each of the Drugs options. We present ML4CAD, a personalized prescriptive algorithm that utilizes multiple ML models at once to identify the most effective therapy for CAD patients. Our method is structured as follows:

1. We impute the missing values of the patient characteristics (Table 1) using a state-of-the-

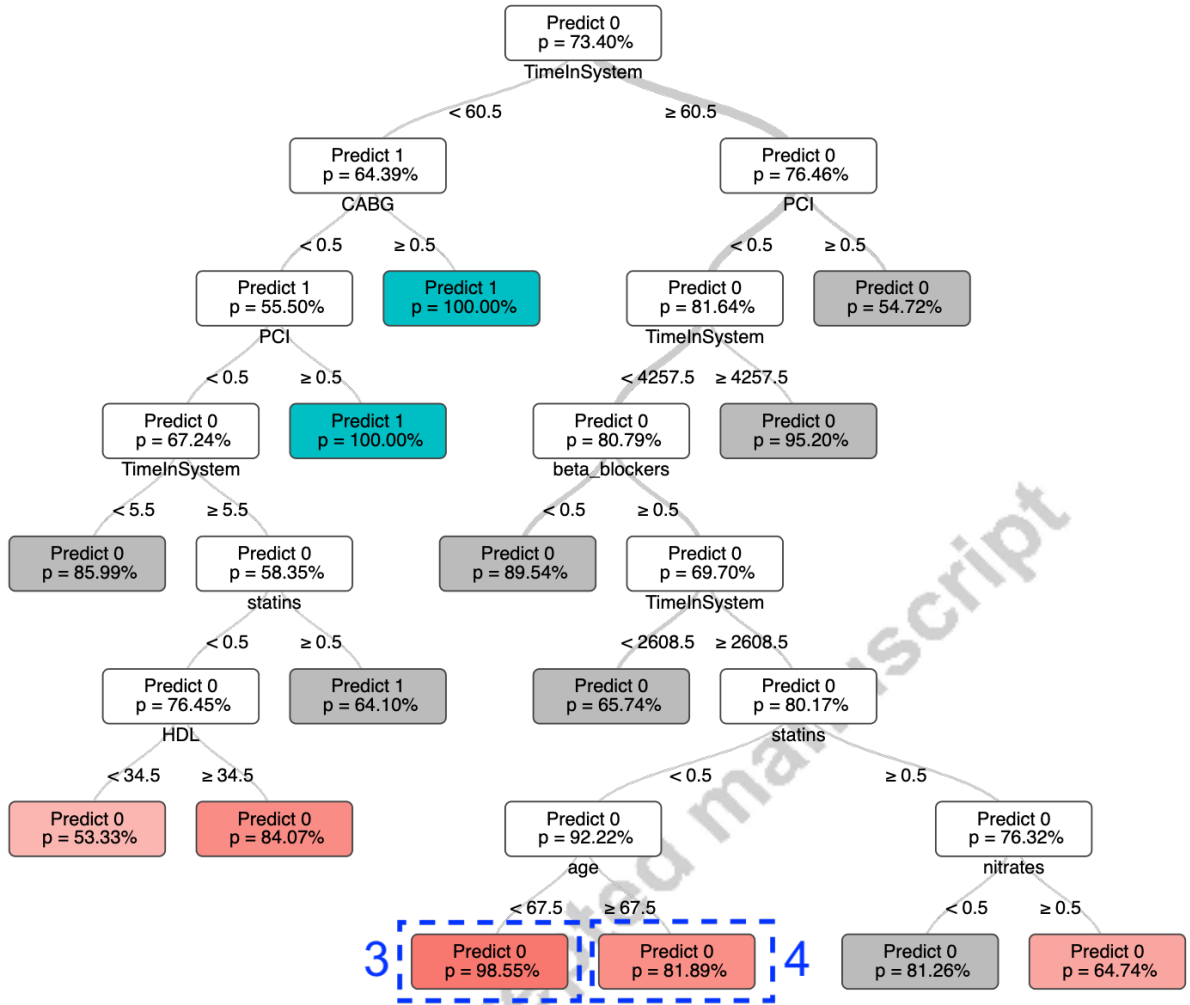


Fig. 4: Visualization of the second part of the OCT model. Paths 3 and 4 are indicated with blue dashed rectangular frames. Shaded nodes include a collapsed subset of the tree model.

	ORT	CART	Random Forest	Linear Regression	Boosted Trees
CABG	73.14%	71.91%	83.00%	80.32%	80.06%
PCI	68.30%	67.73%	74.58%	73.21%	73.21%
Drugs 1	78.64%	75.35%	83.92%	82.94%	82.48%
Drugs 2	73.46%	72.56%	80.02%	79.98%	79.50%
Drugs 3	67.10%	69.03%	77.71%	75.34%	75.29%

Table 6: Results of supervised ML algorithms to predict the TAE since diagnosis. We report the “Out-of-sample” R^2 performance of each model on the Testing set.

- art optimization framework (Bertsimas et al., 2018).
- We compute the TAE for right censored patients.
- We split the population into training and test sets. The training set is used to train the regression models and the test set is utilized to assess the predictive and prescriptive performance of the algorithm.
- We train a separate regression model for each treatment option for all predictive algorithms to estimate the TAE. The set of covariates X' used to create the predictive models does not include any features that refer to the treatment options (see Table 1 for a summary of the in-

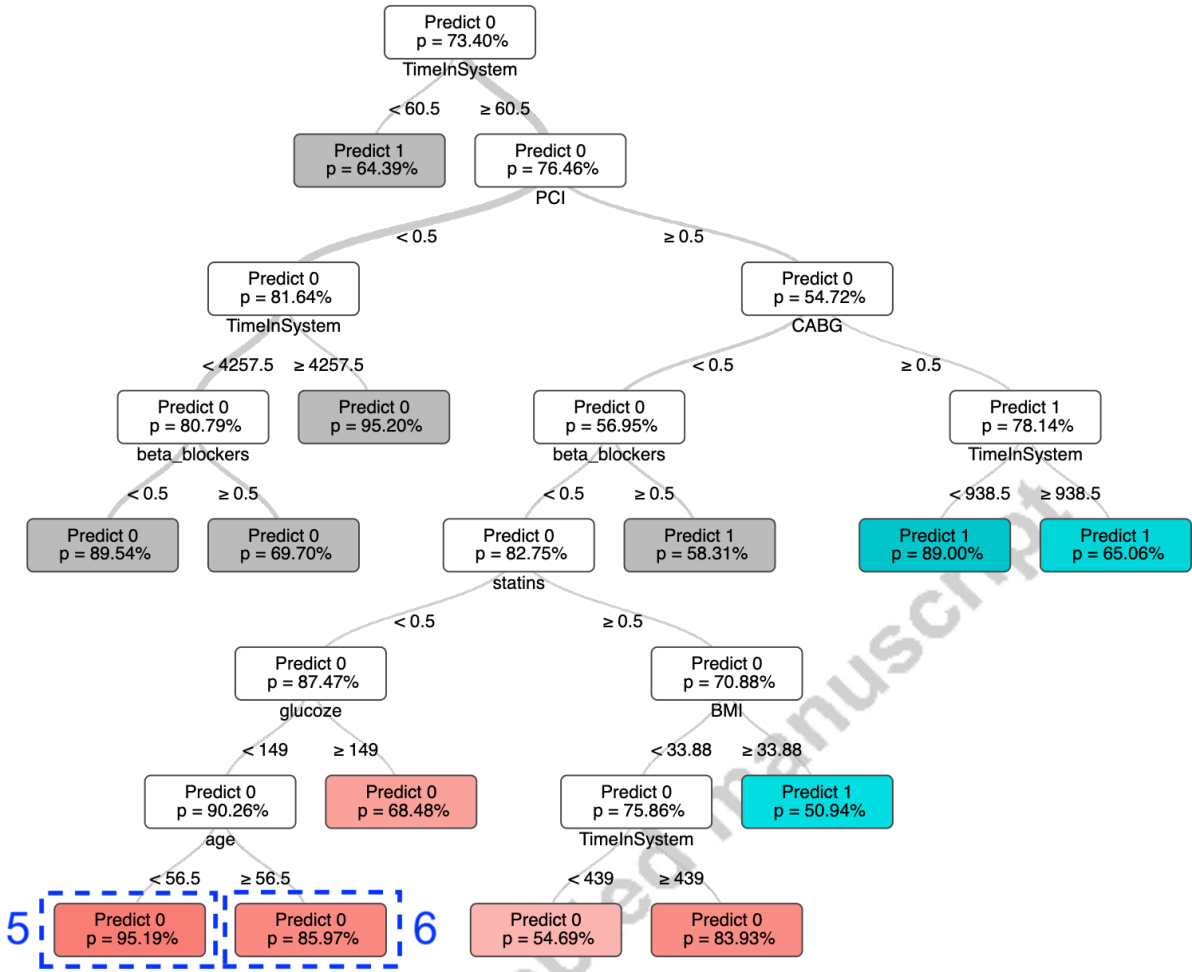


Fig. 5: Visualization of the third part of the OCT model. Paths 3 and 4 are indicated with blue dashed rectangular frames. Shaded nodes include a collapsed subset of the tree model.

dependent features and Table 2 for the list of prescription options).

5. We use all models to get estimations of the TAE for each treatment option and every patient in the test set. Thus, we have at our disposal a table of estimations for any new individual considered. Table 7 provides an illustration of the output for patient X.
6. We select the most effective treatment for the patient according to a voting scheme among the ML methods:
 - (a) If the majority of the regression models votes a single treatment (regimen with the best expected effect), the algorithm recommends this therapy to the physician. In the example of patient X (see Table 7), ML4CAD suggests the prescription of CABG.
 - (b) If there are ties between the different therapies (i.e., two methods suggest Drugs 1 and

two others indicate Drugs 2), then the votes get weighted by the out-of-sample accuracy of the predictive models. For the analysis of this paper, the R^2 metric was used.

7. The final TAE is computed as the average of the ML methods whose suggestion agreed with the algorithm recommendation.

ML4CAD provides a new framework for personalized prescriptions which is structured on the plurality of different ML models. In contrast to the simple Regress and Compare approach, it combines multiple ML models to identify the most beneficial treatment option. The validity of the algorithm's recommendations gets reinforced by an increasing number of underlying ML models that provide accurate estimations of the counterfactuals. In other words, the user gains more confidence in the capability of the algorithm to identify the optimal therapy the more models are available for

ML Method	CABG	PCI	Med. 1	Med. 2	Med. 3
ORT	4.65	4.59	3.89	3.76	3.54
CART	7.13	3.38	6.10	4.16	3.96
Random Forest	5.77	4.93	5.44	4.26	4.49
Linear Regression	5.75	3.53	5.75	4.17	4.44
Boosted Trees	4.08	6.28	5.39	5.31	3.37

Table 7: Estimations of TAE (years) for patient X from the five ML methods considered for each treatment option. We highlight the best treatment option for each ML model. Note that four out of the five models agree on the CABG recommendation.

comparison. This methodology also allows for transparency towards the decision maker. Potential recommendations can be compared at an individual level to be decided what would be the best option for each particular case.

Bridging the gap with practitioners

We created an online ML4CAD application for physicians who would be interested to inform their decision making process using our personalized algorithm. Practitioners can now have access to our website (<https://personalized.shinyapps.io/ML4CAD/>), where they are able to quickly test the recommendations of the algorithm on new patient data. Figure 6 shows an image of the main application dashboard. The platform computes online a table similar to Table 7, demonstrating to the user all the available options and their projected outcomes. The final ML4CAD suggestion is highlighted on the right of the screen. A detailed comparison of the out-of-sample performance of all ML models across the five treatment tasks is also available. Moreover, clinicians can view aggregate results about the treatment allocation mechanism according to different demographic features such as gender, ethnicity, or age group. With this application we aspire to turn the proposed ML-based recommendation system into an actionable framework for the cardiovascular community. The latter can now leverage this tool as an assistance to its decision making process and prolong the life expectancy of its patients.

Prescriptive algorithm evaluation

Assessing the quality of the prescriptive algorithm poses a challenge. We do not have at our disposal data that indicate the TAE for all counterfactual outcomes of each patient. We created appropriate metrics that provide an objective evaluation

framework of the algorithm’s performance. We define the problem as follows, let:

- p be a variable that takes values in the set $[T]$ of all the prescriptive options;
- j be a variable that takes values in the set $[M]$ of all the predictive models;
- z_i be the treatment that patient i followed at the standard of care;
- t_i be the TAE for patient i and treatment z_i ;
- τ_i be the treatment recommendation of ML4CAD for patient i ;
- θ_i^j be the treatment recommendation of machine learning model $j \in [M]$ for patient i using a simple “Regress and Compare approach”;
- $g_i^j(p)$ be the estimated TAE for patient i for treatment p from the regression model j , where $j \in [M]$;
- $y_i(p)$ to be the estimated TAE for patient i when ML4CAD recommends treatment p ;
- \bar{t}_p average TAE observed in the data for all patients who were prescribed treatment p .

Using the notation above, the expected TAE for patient i is according to ML4CAD:

$$y_i(\tau_i) = \frac{1}{K} \sum_{j: \arg \max_p g_i^j(p) = \tau_i} g_i^j(\tau_i), \quad (2)$$

$$K = |j : \arg \max_p g_i^j(p) = \tau_i|, \quad i \in [n].$$

We evaluate the quality of the algorithm’s personalized recommendations based on the following metrics:

1. Prescription Effectiveness and Robustness:

The goal of the first metrics is to compare the performance of the ML4CAD recommendations with the regimens prescribed at the standard of care. Due to the uncertainty in counterfactual estimation, we consider different predictions of the TAE and a multitude of ground truths. Our baseline ground truth refers to realizations of TAE that we observe in the BMC database. This ground truth provides us with

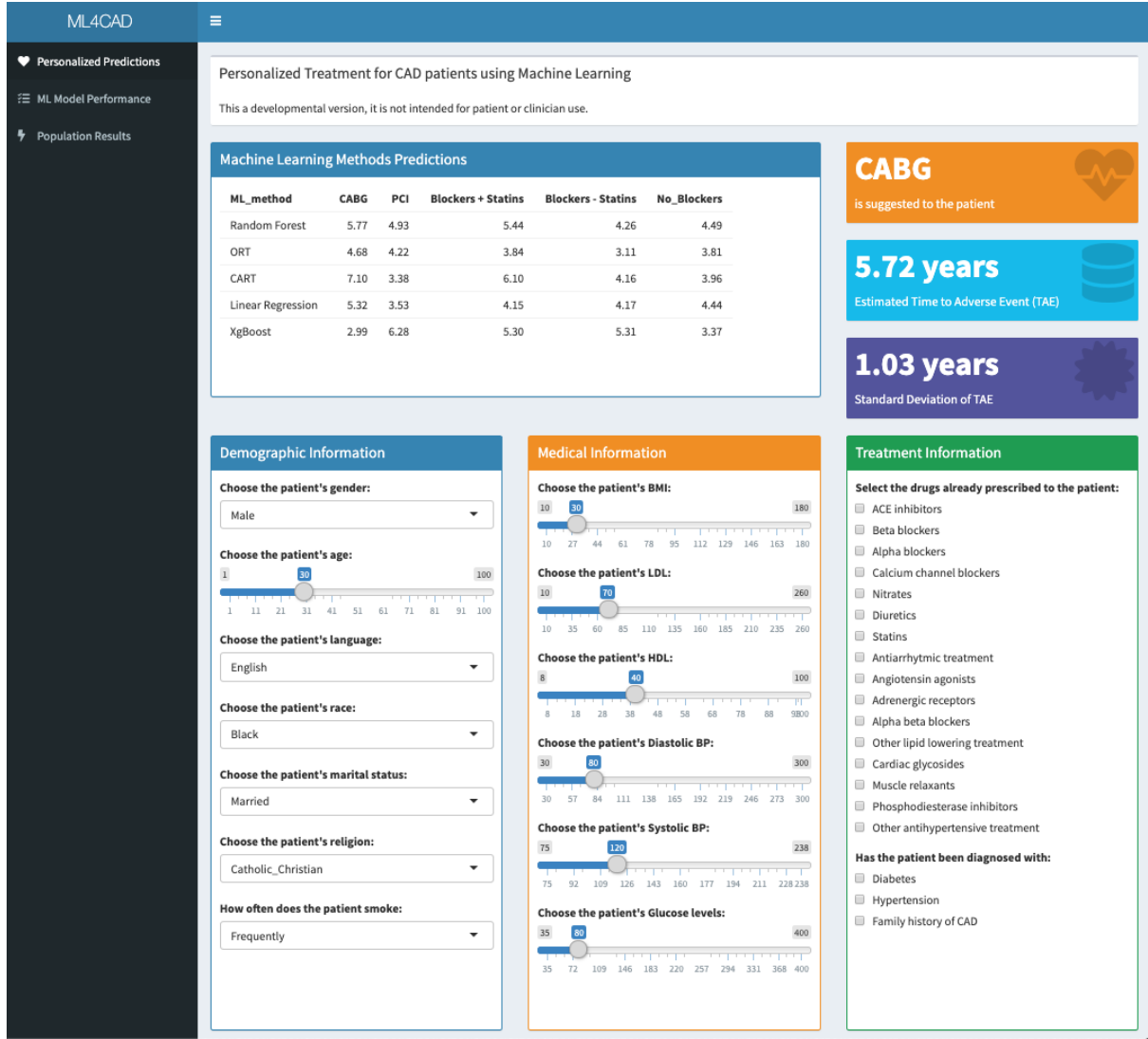


Fig. 6: Treatment Allocation patterns between different ML methods.

the exact TAE associated to the treatment regimen that was prescribed by the physicians at the hospital. Alternative ground truths refer to estimations of the TAE by treatment-based regression models.

– **Prescription Effectiveness (PE)**

We fix, for each patient $i \in [n]$, the treatment suggestion τ_i from the ML4CAD algorithm. We know the outcome t_i for treatment choice z_i (observed in the data - baseline ground truth). Thus, comparing the prescription effectiveness of the ML4CAD versus the standard of care would be equal to:

$$PE(ML4CAD) = \frac{1}{n} \sum_{i=1}^n y_i(\tau_i) - t_i. \quad (3)$$

ML4CAD averages the TAE projected by the regression models that agree on the most beneficial treatment for patient i , namely τ_i . We can evaluate the prescription effectiveness of this recommendation by considering each ML model in isolation. Each regression model j provides for patient i and regimen p an estimation $g_i^j(p)$. Therefore, if we fix $p = \tau_i$, we can get an evaluation of the projected TAE and compare it to the standard of care.

$$PE(ML_j) = \frac{1}{n} \sum_{i=1}^n g_i^j(\tau_i) - t_i, \quad (4)$$

$$\forall j \in \{1, \dots, M\}.$$

Comparing multiple ML estimations for the TAE of the recommendation τ_i renders the results more credible to biases of a specific predictive algorithm.

– **Prescription Robustness (PR)**

The PE metric measures the effect of the ML4CAD recommended therapies against a fixed given ground truth from the EMR of the BMC. Nevertheless, knowing that each patient i was given a treatment t_i , we can generate alternative ground truths. We can, then, evaluate the benefit of the personalization approach against those. Each ground truth corresponds to an estimation of what would happen to patient i if ML model j was an oracle that knew the reality and the effects of treatment z_i .

$$\text{PR}(\text{ML}_{j,k}) = \frac{1}{n} \sum_{i=1}^n (g_i^j(\tau_i) - g_i^k(z_i)), \quad (5)$$

$$\forall j, k \in [M].$$

In this setting, decisions τ_i, z_i are fixed and we evaluate all the combinations between Random Forest, CART, ORT, Boosted Trees, and Linear Regression. We include also the case where ML4CAD is used to estimate the effect of τ_i but not the one of t_i .

$$\text{PR}(\text{ML4CAD}_k) = \frac{1}{n} \sum_{i=1}^n (y_i(\tau_i) - g_i^k(z_i)), \quad (6)$$

$$\forall k \in [M].$$

The goal of this metric is to evaluate the robustness of the treatment effect under different ground truths. In Section 7, we perform an extensive comparison over all methods and ground truths considered (see Table 8). We introduce this approach to avoid biased estimates of performance. The latter could not have been avoided if we were comparing our results only to the baseline ground truth.

2. Prediction accuracy of TAE:

$$\tilde{R}^2(\text{ML4CAD}) = 1 - \frac{\sum_{i \in S} (y_i(z_i) - t_i)^2}{\sum_{i \in S} (\bar{t}_{z_i} - t_i)^2}, \quad (7)$$

$$S = \{i : \tau_i = z_i\}, \quad i \in [n].$$

This metric follows the same structure as the well-known coefficient of determination R^2 . We apply it for each patient $i \in S$, the set of all

samples where there is agreement between the ML4CAD and baseline prescription; $S = \{i : \tau_i = z_i\}$. Similar to the original measure, the known outcome t_i is compared to the estimated treatment effect $y_i(z_i)$ and to a baseline estimation. The latter in our case is \bar{t}_{z_i} , the mean TAE observed in the data for all patients who were prescribed treatment z_i . The adjusted coefficient of determination \tilde{R}^2 helps us evaluate whether the outcome that ML4CAD predicts for the known counterfactuals is accurate or not. It is impossible to evaluate the prescriptive algorithm across all treatment options. Only one out of the five is actually realized in practice. We focused on comparing for each patient the TAE according to the algorithm versus the one present in the data only for the cases where there was agreement between the two. This estimation, even though limited, provides us with a good baseline regarding the accuracy of our recommendations. We can extend the use of this metric to the ‘‘Regress and Compare’’ approach. Thus, we can estimate the $\tilde{R}^2(\text{ML}_j)$ of each predictive model $j \in [M]$.

$$\tilde{R}^2(\text{ML}_j) = 1 - \frac{\sum_{i \in S} (g_i^j(z_i) - t_i)^2}{\sum_{i \in S} (\bar{t}_{z_i} - t_i)^2}, \quad (8)$$

$$S = \{i : \theta_i^j = z_i\}, \quad i \in [n].$$

3. Degree of ML Agreement (DMLA):

This measure refers to the degree of agreement among the ML models (DMLA) with the recommended treatment τ_i . For each patient, we count the number of methods that agree on the ML4CAD suggested treatment τ_i . We report the distribution of this metric across the whole population. Cases where there is high degree of agreement are associated with higher confidence on the suggested prescription. On the contrary, we are less confident in cases where there is misalignment between the ML models regarding the best treatment option.

7 Prescriptive algorithm results

In this Section, we present numerical results with respect to the evaluation metrics introduced in Section 6. We provide insights regarding different sample population subgroups. We also discuss new treatment allocation patterns based on ML4CAD recommendations.

7.1 Prescription Effectiveness (PE) and Robustness (PR)

We summarize our results with respect to the PE and PR metrics in Table 8. The first table column corresponds to PE (baseline ground truth), whereas the rest of the columns refer to PR (ML-based ground truths). Table 8 presents the expected relative gain in TAE of ML4CAD over the baseline. Its values demonstrate the average benefit in years of TAE when comparing the current and ML4CAD treatment allocation plan across different estimation models. Each ground truth (column) refers to alternative estimations of the TAE under the current treatment allocation plan. Thus, if the ground truth is the baseline (BMC Database), the suggested times correspond the TAE observed in the data. When the ground truth is set to be the ORT algorithm, the predicted times $g_i^{ORT}(z_i)$ mirror ORT estimations when the treatment allocation is fixed to the physicians' decisions from the hospital (z_i). Each prediction model (row) provides us with a continuous prediction of a patient's TAE when the treatment allocation plan is set by the ML4CAD algorithm (τ_i). Thus, the values in Table 8 correspond to the metrics defined in Equations 4 (first column) and 5 (subsequent columns).

When compared to the current allocation scheme, our prescription algorithm improves the average TAE by 24.11%, with respect to the PE metric, with an increase from 4.56 to 5.66 years (13 months). Column "Baseline (PE)" of Table 8 summarizes the results with respect to all regression models considered. ML4CAD provides the most optimistic estimations. It suggests a higher TAE versus its counterparts by at least 0.18 years (2 months). Linear Regression appears to be the most pessimistic method with an average benefit over the baseline of 6 months (0.59 years). ORT and Random Forest provide similar estimations of 0.77 and 0.75 years of improvement, respectively.

The comparable performance of the various estimation models presented in Table 8 reinforces the credibility of the prescription algorithm. We show that there is agreement between the potential improvement in the average TAE by an alternative

treatment allocation scheme. Even in cases where we include ML models that did not participate in the ML4CAD recommendation, there is substantial benefit in the patients' life expectancy.

We observe better results across all age and ethnicity patient subgroups and for both genders. The benefit of using the algorithm was 17.09% (0.9 years) for Black patients, 29.03% (1.16 years) for Caucasian patients and 58.41% (1.86 months) for Hispanic patients. We also note 22.5% (0.99 years) improvement for patients 65 – 80 years of age and 46.9% (1.58 years) for patients aged 80 or older. Male patients are expected to increase their time from 4.62 years to 5.73 (24.19% improvement) similar to female patients (from 4.42 years to 5.48). The performance of the prescriptive algorithm for selected patient subgroups compared to the BMC baseline is summarized in Figure 7.

In terms of the PR metric, our results demonstrate a consistent improvement of the patient population TAE across all ground truths and estimation models. Table 8 summarizes the results of our analysis. We note that ML4CAD achieves the highest benefit when compared to all alternative scenarios of outcome realization. This is due to the incorporation of the voting system for the selection of the most effective treatment that accounts for all ML models. We show that even in the case of more pessimistic estimators, such as Boosted Trees or Linear Regression, there is a substantial benefit compared to the standard of care. Our approach does not guarantee optimality for the treatment selection problem. Nevertheless, it is experimentally shown that it can bring about substantial benefit to the CAD population.

We can also identify for each estimation model combinations with ground truths that outperform the rest of the alternatives. All methods demonstrate the highest improvement when associated with the Boosted Trees ground truth. For example, the ORT and CART model increase the average TAE by 0.96 and 1.10 years respectively. The next most optimistic contestant is Linear Regression. This is due to the fact that some methods on average overestimate or underestimate the expected TAE, translating these discrepancies in the PR metric.

7.2 Prediction accuracy of TAE

The "prediction accuracy of TAE" for the proposed prescriptive algorithm is $\tilde{R}^2(\text{ML4CAD}) = 78.7\%$. Table 9 provides a summary of the results for both

*The PE of the algorithm when the estimation model g^j is ML4CAD and the ground truth relates to the patient outcomes observed in the BMC database (See Equation 4).

†The PR of the algorithm when CART is the chosen estimation model g^j for the prescriptions $z_i, i \in [n]$ and the ground truth outcomes are computed according to the Linear Regression model g^k (See Equation 5).

Estimation Model	Ground Truth					
	Baseline	ORT	CART	Random Forest	Linear Regression	Boosted Trees
ML4CAD	1.101*	1.162	1.158	1.140	1.178	1.283
ORT	0.779	0.840	0.835	0.818	0.855	0.961
CART	0.923	0.983	0.979	0.965	0.999	1.105
Random Forest	0.757	0.818	0.813	0.796	0.833 [†]	0.939
Linear Regression	0.485	0.546	0.541	0.524	0.561	0.667
Boosted Trees	0.591	0.652	0.647	0.630	0.667	0.773

Table 8: Comparison of the “Prescription effectiveness” (PE) and “Prescription robustness” (PR) metrics for all estimation models and ground truths considered. The first column (Baseline) presents results with respect to the PE metric and refers to the TAE observed in the BMC database. All subsequent columns refer to the PR measure. Each of them represents a distinct ground truth. All units are shown in years. See Equations 4,3,5.

the suggested method as well as “Regress and Compare” approaches from the baseline ML models. ML4CAD achieves better performance compared to the single prediction model counterparts. Aggregated predictions from different regression models lead to more accurate outcomes. The suggested voting scheme, not only reduces the uncertainty and bias of the estimations (See Section 7.1), but also results in highly accurate predictions.

Method	\tilde{R}^2
ML4CAD	78.70%
ORT	72.68%
CART	70.54%
Random Forest	77.25%
Linear Regression	76.66%
Boosted Trees	76.59%

Table 9: Results summary for the Prediction Accuracy of TAE (\tilde{R}^2) metric.

7.3 Degree of ML agreement (DMLA)

The majority of the ML4CAD recommendations z_i are based on a common suggestion between at least three distinct ML models. Specifically, in 14.53% of the patients all methods suggest the same treatment for each individual. In 26.74% of the cases there is agreement between four models and in 34.48% of the observations three methods participate in the decision. Only in 0.26% of the samples, each regression model suggests a different prescription. In such cases, the ML4CAD recommendation is solely based on the suggestion of the most accurate one.

Table 10 provides detailed results for each treatment option. The last table column summarizes the results as a function of the total population. Each treatment specific column presents the proportional degree of agreement for all patients for which this treatment was suggested. Thus, we notice that CABG as well as Drugs 1 & 2 recommendations are, on average, more confident compared to Drugs 3 or PCI due to the higher degree of agreement. This is particularly true in the case of Drugs 1, where for 85.49% of the patients, three out of the five methods voted for the same regimen.

7.4 Treatment Allocation Patterns

In this section, we present insights regarding the ML4CAD treatment allocation patterns and we perform comparisons with the standard of care at the BMC. Our method agrees with the physicians’ decisions in 28.24% of the cases. The results indicate a shift towards drug therapy and CABG, reducing the overall proportion of PCI (from 18.84% to 6.04%). The prediction model indicates that patients with severe symptoms do not benefit significantly from a PCI versus a CABG surgery due to the eminent need for revascularization. Figure 8 illustrates a significant shift towards “Drugs 1” for both women and men. The algorithm also recognizes that treatment “Drugs 2” is less effective on female patients versus male. The ML4CAD allocation is in agreement with the most recent guidelines published by the American Heart Association (Stout et al., 2018). In the vast majority of cases, a combination of antihypertensive drugs (Blockers) with lipid lowering treatment (statins) is suggested. The overall proportion of the population that is recommended an invasive intervention is

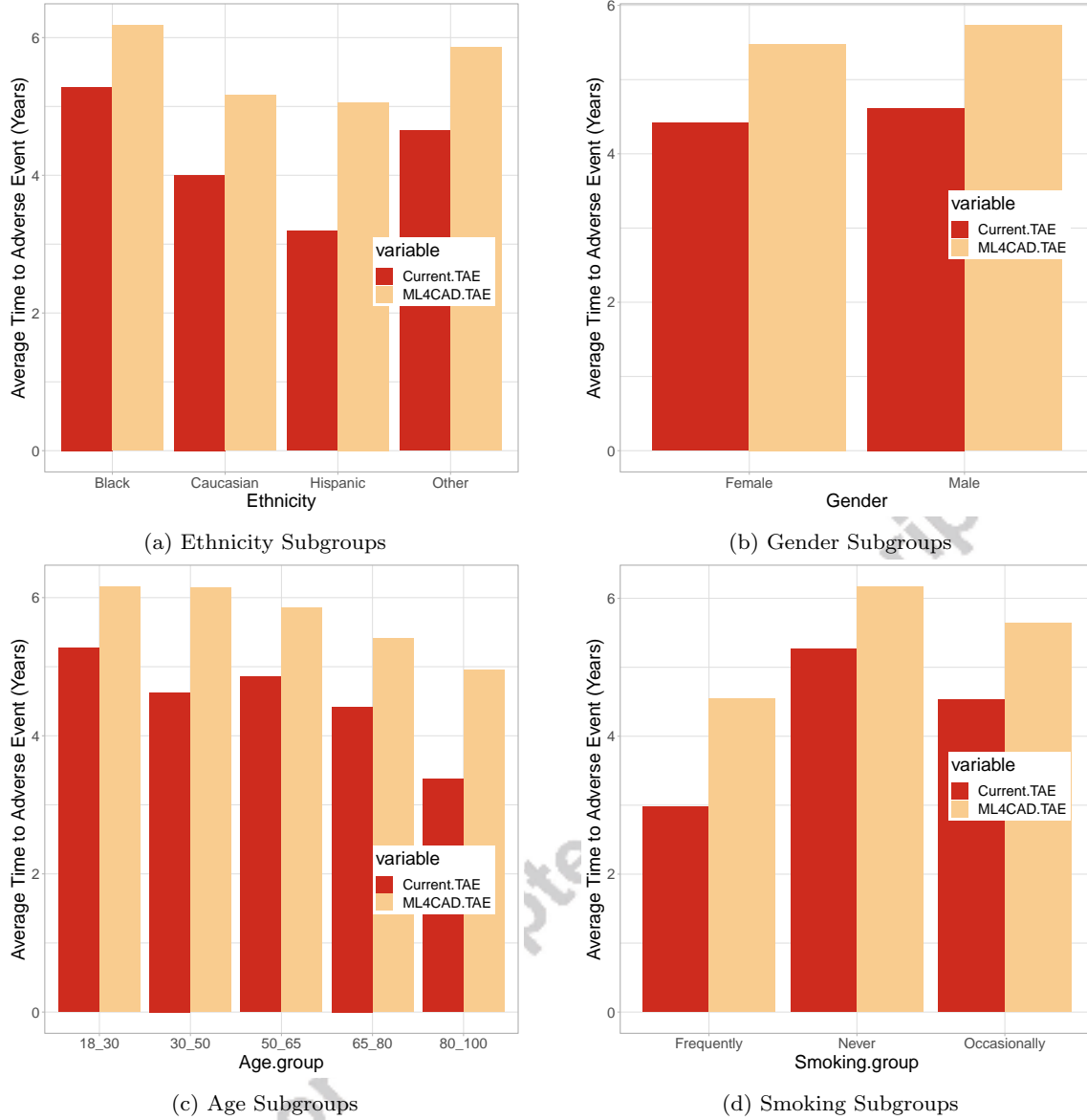


Fig. 7: Comparison of the expected years to adverse event after diagnosis for the age and ethnicity subgroups considered. The difference between the two bars for each sub-population refers to the prescription effectiveness (PE) of the algorithm for each respective patient group. “Current.TAE” refers to the outcomes observed in the EHR of the BMC. “ML4CAD.TAE” represents the expected TAE according to the prescription algorithm.

reduced due to the significant decline of PCI operations.

Figure 9 illustrates a comparison of the treatment allocation patterns between the ML4CAD algorithm, individual “Regress and Compare” models, and the standard of care we observe in the data. The graph demonstrates an agreement across all methods other than CART to increase the proportion of the population under “Drugs 1”. The ML4CAD algorithm is more aligned with the Random Forest policy due to the high predictive per-

formance associated with the latter. We also note the reduction of “Drugs 2 & 3” across all methods. In the case of CABG there is disagreement between the ML models. Boosted Trees and Linear Regression suggest a significant raise in the proportion of CABG surgery at the expense of “Drugs 1”. On the other hand, ORT, Random Forest, and CART identify CABG as the optimal therapy for a lower proportion of the patient population.

Number of ML methods that agree with the recommendation	CABG	Drugs 1	Drugs 2	Drugs 3	PCI	Population Proportion
1	1.13%	0.22%	0.00%	0.00%	0.00%	0.26%
2	20.82%	14.29%	41.54%	59.65%	49.10%	23.99%
3	35.41%	32.30%	43.98%	36.23%	39.07%	34.48%
4	27.34%	33.58%	13.26%	3.64%	10.28%	26.74%
5	15.30%	19.61%	1.22%	0.47%	1.54%	14.53%

Table 10: Degree of ML Agreement between the models analyzed for each treatment option as well as a function of the overall test population.

ML4CAD Allocation						
Current Allocation	<i>Treatment</i>	CABG	Drugs 1	Drugs 2	Drugs 3	PCI
	CABG	1.3%	4.1%	0.9%	1.6%	0.8%
	Drugs 1	2.3%	22.1%	3.7%	2.1%	1.7%
	Drugs 2	2.0%	12.3%	2.0%	0.2%	1.0%
	Drugs 3	3.2%	16.3%	1.0%	1.4%	1.1%
	PCI	2.2%	9.5%	1.3%	4.5%	1.4%

Table 11: Allocation of patients in the treatment options based on the standard of care and ML4CAD.

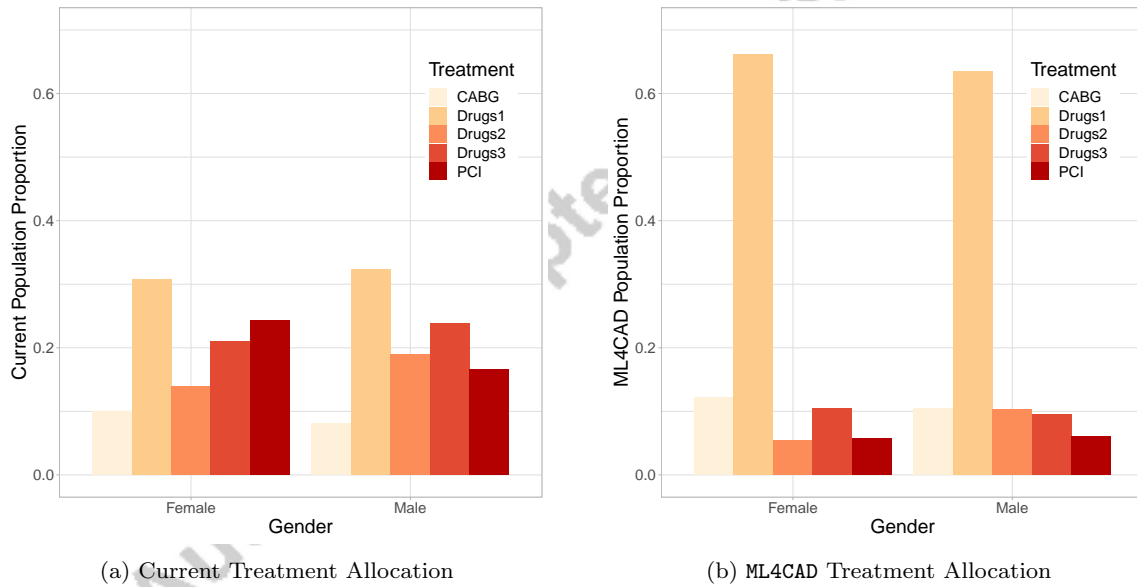


Fig. 8: Population allocation to treatments split by gender.

8 Discussion and Conclusions

Combining historical data from a large EMR database and state-of-the-art ML algorithms resulted in an average TAE benefit of 24.11%% (1.1 years) for patients diagnosed with CAD. Our results show that differing medication regimens and revascularization strategies may produce varying clinical outcomes for patients. The use of ML may facilitate the identification of the optimal treatment strategy. Such efforts could directly address the

primary objectives of the clinical cardiovascular practice, leading to symptoms reduction and an increase in the population life expectancy. Our findings uncover the greatest clinical benefit in medical therapy changes, consistent with themes that have emerged in clinical trials (Boden et al., 2007). The optimal revascularization strategy in patients with multi-vessel CAD is an area of active investigation, with efforts focused on identifying which patient subgroups may benefit from different revas-

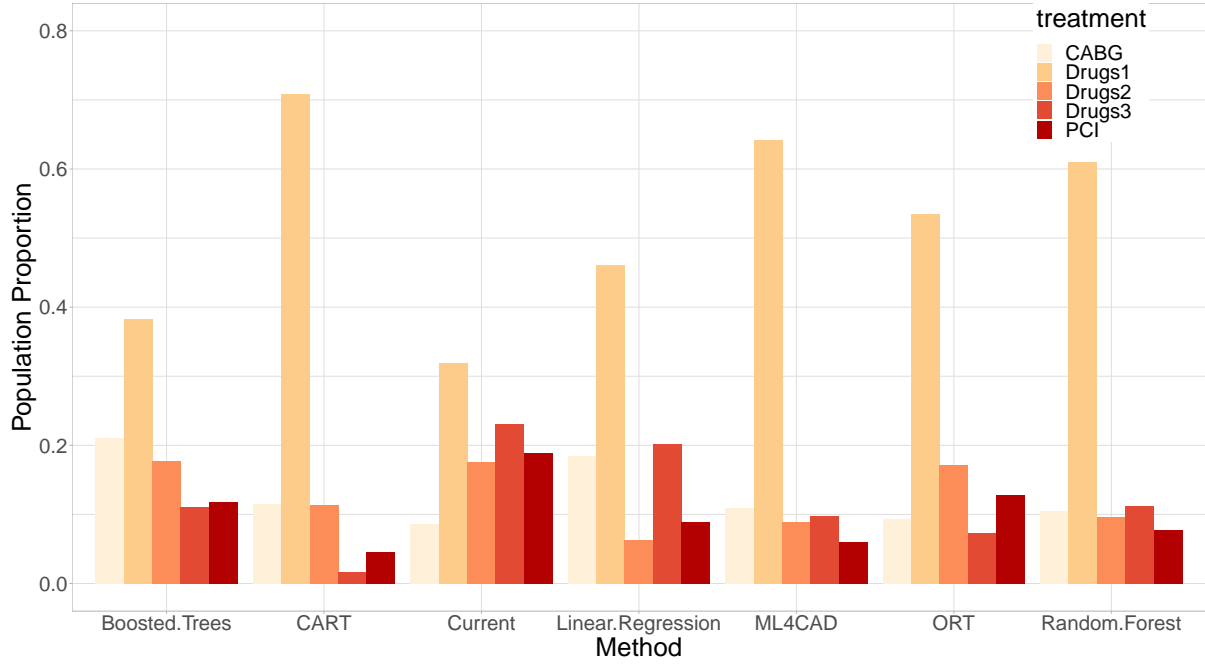


Fig. 9: Treatment Allocation patterns between different ML methods.

cularization procedures (Farkouh et al., 2012). Our technique may add clarity to this clinical challenge.

Our prescriptive approach is accurate, highly interpretable, and flexible for other healthcare applications. The use of multiple ground truths derived from independent ML models renders credibility to the results. In prescriptive problems where counterfactual outcomes cannot be evaluated against a known reference, leveraging multiple ML models can reduce the uncertainty behind suggested recommendations. For this reason, we believe that metrics such as the prescription effectiveness and robustness are key to the validation process.

Moreover, our online application bridges the gap between clinicians and the algorithm. Users can directly and simultaneously interact with multiple ML models from a user-friendly interface. Our method should easily accommodate alternative cardiovascular disease-management approaches within specific disease subpopulations, such as arrhythmia and valvular disease management. A novelty of our approach is in the personalization of the decision-making process. It incorporates patient-specific factors, and provides guidelines for the physician at the time of diagnosis / clinical encounter. We believe this personalization is the primary driver of benefit relative to the standard of care. Similarly, there is emerging data on use of ML techniques to improve cardiac imaging phenotyping of

cardiac disease states, such as heart failure (Omar et al., 2017).

The widespread use of EMR in clinical medicine was initially viewed with much optimism, however more recently it has been met with frustration by clinical providers. Concerns are being raised over the administrative burden to document the EMR and the resultant development of clinician “burn out”. The methodology presented in this paper identifies a mechanism to harness the power of the EMR in an effort to improve patient care and make it more personalized. It is true that the clinical acumen developed over time spent caring for patients cannot be replaced by algorithms. Nevertheless, the prospect of ML to guide clinicians and complement clinical decision making may help improve clinical outcomes for patients with cardiovascular and other diseases (Ebinger et al., 2016).

Our work has several limitations due to the nature of the EMR. A large percentage of the sample was right-censored. Patients were not randomized into treatment groups. Our data does not include socioeconomic factors or patient preferences that may be important in treatment decisions, such as income or fear of invasive treatment strategies. Although our matching methodology controls for several confounding factors that could explain differences in treatment effects, we can only estimate counterfactual outcomes. In addition, the study population of BMC is not representative of the

general U.S. population as we observe a higher representation of non-Caucasian patients. As a result, the ability of ML4CAD to generalize in other institutions needs to be tested. Similarly to other studies, we recommend prospective validation of the models to the new population prior to the application of the algorithm to a different healthcare system (Orfanoudaki et al., 2020). Moreover, we should consider that the accuracy of the prediction model is limited, though significantly better than the baseline model. It leaves room for improvement in that field by including new variables and further risk factors that are associated with CAD. Due to lack of sufficient data, we did not take into account different types of CABG surgery (i.e. arterial versus venous conduits) and PCI (i.e. newer versus older generation drug eluting stents, or bare metal stents versus drug eluting stents). Should more data were available, we could further differentiate the prescription categories beyond the five we include in this analysis, including drug specific recommendations. Moreover, the algorithm does not agree with the standard of care in most cases. This result indicates that new personalization techniques would need further input from clinicians that was not originally recorded in the EMR. Future research could address the issue of right censored patients with different approaches, which incorporate the time varying effects of the explanatory variables using optimization rather than heuristic methodologies. The ultimate validation of our algorithm would be the realization of a clinical trial. There we would be able to test the personalized recommendation to patients directly utilizing their EMR from the hospital system.

Despite these limitations, our approach establishes strong evidence for the benefit of individualizing CAD care. To our knowledge, this work represents the first ML study in treating cardiovascular disease and serves as a proof of concept. Moreover, the success of this data-driven approach invites further testing using datasets from other hospitals and patient populations. That includes care settings that contain more detailed information regarding the patients' condition, such as electrocardiogram findings and exercise and other lifestyle factors. The algorithm could be integrated in practice into existing EMR systems to generate dynamically personalized treatment recommendations. Testing the prescriptive algorithm in a clinical trial setting could provide conclusive evidence of clinical effectiveness. As large-scale genomic data become

more widely available, the algorithm could readily incorporate such data to reach the full potential of personalized medicine in cardiovascular disease care. Our work is a key step toward a fully patient-centered approach to coronary artery disease management and the application of modern analytics in the medical field.

Acknowledgements The authors wish to thank the anonymous reviewers and the associate editor of the journal for their helpful comments on this manuscript. They, also, thank Theofanie Mela MD (Massachusetts General Hospital), and Abeel A. Mangi MD (Yale Medicine Department) for sharing clinical expertise as well as Bill Adams, MD and the Boston Medical Center for the use of its i2b2 database.

Funding

This research was supported by the National Science Foundation grant 6926678 [“SHB: Type II (INT): Collaborative Research: Algorithmic Approaches to Personalized Health Care”].

Conflict of interest

The authors declare that they have no conflict of interest.

Ethics approval

The Massachusetts Institute of Technology and Boston Medical Center Institutional Review Boards approved the study.

Availability of data and material

All datasets that are used in this study come from an academic medical center that applies to the Health Insurance Portability and Accountability Act. Due to the data protection laws, the dataset cannot be directly released to another organization. We invite readers that would like to gain access to the dataset to establish a data use agreement with the BMC.

References

- AHA (2017) Heart disease and stroke statistics 2017. AHA Centers for Health Metrics and Evaluation

- Angrist JD, Imbens GW, Rubin DB (1996) Identification of causal effects using instrumental variables. *Journal of the American statistical Association* 91(434):444–455
- Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, Himmel-farb CD, Khera A, Lloyd-Jones D, McEvoy JW, Michos ED, Miedema MD, Muñoz D, Smith SC, Virani SS, Williams KA, Yeboah J, Ziaeian B (2019) 2019 acc/aha guideline on the primary prevention of cardiovascular disease. *Journal of the American College of Cardiology* 74(10):e177–e232, DOI 10.1016/j.jacc.2019.03.010, URL <https://www.onlinejacc.org/content/74/10/e177>, <https://www.onlinejacc.org/content/74/10/e177.full.pdf>
- Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113(27):7353–7360
- Beitelshees AL (2012) Personalised antiplatelet treatment: a rapidly moving target. *The Lancet* 379(9827):1680 – 1682, DOI [https://doi.org/10.1016/S0140-6736\(12\)60431-0](https://doi.org/10.1016/S0140-6736(12)60431-0), URL <http://www.sciencedirect.com/science/article/pii/S0140673612604310>
- Bertsimas D, Dunn J (2017) Optimal classification trees. *Machine Learning* 106(7):1039–1082
- Bertsimas D, Dunn J (2019) *Machine Learning under a Modern Optimization Lens*. Dynamic Ideas, Belmont
- Bertsimas D, Kallus N, Weinstein AM, Zhuo YD (2017) Personalized diabetes management using electronic medical records. *Diabetes care* 40(2):210–217
- Bertsimas D, Pawlowski C, Zhuo YD (2018) From predictive methods to missing data imputation: an optimization approach. *The Journal of Machine Learning Research* 18(1):7133–7171
- Bertsimas D, Dunn J, Mundru N (2019) Optimal prescriptive trees. *INFORMS Journal on Optimization* 1(2):164–183
- Boden WE, O'Rourke RA, Teo KK, Hartigan PM, Maron DJ, Kostuk WJ, Knudtson M, Dada M, Casperson P, Harris CL, Chaitman BR, Shaw L, Gosselin G, Nawaz S, Title LM, Gau G, Blaustein AS, Booth DC, Bates ER, Spertus JA, Berman DS, Mancini GJ, Weintraub WS (2007) Optimal medical therapy with or without pci for stable coronary disease. *New England Journal of Medicine* 356(15):1503–1516, DOI 10.1056/NEJMoa070829, URL <http://dx.doi.org/10.1056/NEJMoa070829>, PMID: 17387127, <http://dx.doi.org/10.1056/NEJMoa070829>
- Bou-Hamad I, Larocque D, Ben-Ameur H (2011) A review of survival trees. *Statistics Surveys* 5:44–71
- Breiman L (2001) Random forests. *Machine Learning* 45(1):5–32
- Breiman L, Friedman J, Olshen R, Stone C (1984) *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, California
- Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:160302754*
- Conroy R, Pyörälä K, Fitzgerald Ae, Sans S, Menotti A, De Backer G, De Bacquer D, Ducimetiere P, Jousilahti P, Keil U, et al. (2003) Estimation of ten-year risk of fatal cardiovascular disease in europe: the score project. *European heart journal* 24(11):987–1003
- Cox DR (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B (Methodological)* 34(2):187–220, URL <http://links.jstor.org/sici?sici=0035-9246%281972%2934%3A2%3C187%3ARMAL%3E2.0.CO%3B2-6>
- D'agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB (2008) General cardiovascular risk profile for use in primary care. *Circulation* 117(6):743–753
- Duan T, Rajpurkar P, Laird D, Ng AY, Basu S (2019) Clinical value of predicting individual treatment effects for intensive blood pressure therapy: A machine learning experiment to estimate treatment effects from randomized trial data. *Circulation: Cardiovascular Quality and Outcomes* 12(3):e005010
- Ebinger JE, Porten BR, Strauss CE, Garberich RF, Han C, Wahl SK, Sun BC, Abdelhadi RH, Henry TD (2016) Design, challenges, and implications of quality improvement projects using the electronic medical record. *Circulation: Cardiovascular Quality and Outcomes* 9(5):593–599, DOI 10.1161/CIRCOUTCOMES.116.003122, URL <http://circoutcomes.ahajournals.org/content/9/5/593>, <http://circoutcomes.ahajournals.org/content/9/5/593.full.pdf>
- Emanuel EJ, Wachter RM (2019) Artificial Intelligence in Health Care: Will the Value Match the Hype? *Artificial Intelligence in Health Care—Will the Value Match the Hype? Artificial Intelligence in Health Care Will the Value Match the Hype?*

- JAMA DOI 10.1001/jama.2019.4914, URL <https://doi.org/10.1001/jama.2019.4914>, https://jamanetwork.com/journals/jama/articlepdf/2734581/jama_emanuel.2019_vp_190060.pdf
- Epstein CCL (2014) An analytics approach to hypertension treatment. PhD thesis, Massachusetts Institute of Technology
- Farkouh ME, Domanski M, Sleeper LA, Siami FS, Dangas G, Mack M, Yang M, Cohen DJ, Rosenberg Y, Solomon SD, Desai AS, Gersh BJ, Magnuson EA, Lansky A, Boineau R, Weinberger J, Ramanathan K, Sousa JE, Rankin J, Bhargava B, Buse J, Hueb W, Smith CR, Muratov V, Bansilal S, King SI, Bertrand M, Fuster V (2012) Strategies for multivessel revascularization in patients with diabetes. *New England Journal of Medicine* 367(25):2375–2384, DOI 10.1056/NEJMoa1211585, URL <http://dx.doi.org/10.1056/NEJMoa1211585>, pMID: 23121323, <http://dx.doi.org/10.1056/NEJMoa1211585>
- FDA (2017) Clinical and patient decision support software - guidance for industry and food and drug administration staff. Available at <http://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-and-patient-decision-support-software> (2017/05/27)
- Feldstein ML, Savlov ED, Hilf R (1978) A statistical model for predicting response of breast cancer patients to cytotoxic chemotherapy. *Cancer research* 38(8):2544–2548
- Fihn SD, Blankenship JC, Alexander KP, Bittl JA, Byrne JG, Fletcher BJ, Fonarow GC, Lange RA, Levine GN, Maddox TM, Naidu SS, Ohman EM, Smith PK (2014) 2014 acc/aha/aats/pcna/scail/sts focused update of the guideline for the diagnosis and management of patients with stable ischemic heart disease: A report of the american college of cardiology/american heart association task force on practice guidelines, and the american association for thoracic surgery, preventive cardiovascular nurses association, society for cardiovascular angiography and interventions, and society of thoracic surgeons. *Journal of the American College of Cardiology* 64(18):1929 – 1949, DOI <https://doi.org/10.1016/j.jacc.2014.07.017>, URL <http://www.sciencedirect.com/science/article/pii/S0735109714045100>
- Fihn SD, Gardin JM, Abrams J, Berra K, Blankenship JC, Dallas AP, Douglas PS, Foody JM, Gerber TC, Hinderliter AL, King SB, Kligfield PD, Krumholz HM, Kwong RY, Lim MJ, Linderbaum JA, Mack MJ, Munger MA, Prager RL, Sabik JF, Shaw LJ, Sikkema JD, Smith CR, Smith SC, Spertus JA, Williams SV (2015) 2012 accf/aha/acp/aats/pcna/scail/sts guideline for the diagnosis and management of patients with stable ischemic heart disease: A report of the american college of cardiology foundation/american heart association task force on practice guidelines, and the american college of physicians, american association for thoracic surgery, preventive cardiovascular nurses association, society for cardiovascular angiography and interventions, and society of thoracic surgeons. *Circulation* 60(24):e44 – e164
- Frohlich H, Balling R, Beerenwinkel N, Kohlbacher O, Kumar S, Lengauer T, Maathuis MH, Moreau Y, Murphy SA, Przytycka TM, Rebhan M, Rost H, Schuppert A, Schwab M, Spang R, Stekhoven D, Sun J, Weber A, Ziemek D, Zupan B (2018) From hype to reality: data science enabling personalized medicine. *BMC Medicine* 16(1):150, DOI 10.1186/s12916-018-1122-7, URL <https://doi.org/10.1186/s12916-018-1122-7>
- Fuster V, Badimon L, Badimon JJ, Chesebro JH (1992) The pathogenesis of coronary artery disease and the acute coronary syndromes. *New England journal of medicine* 326(5):310–318
- Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G (2018) Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine* 178(11):1544–1547, DOI 10.1001/jamainternmed.2018.3763, URL <https://www.ncbi.nlm.nih.gov/pubmed/30128552>
- Gittins JC, Glazebrook KD, Weber R, Weber R (1989) Multi-armed bandit allocation indices, vol 25. Wiley Online Library
- Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D’Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O’Donnell CJ, Robinson JG, Schwartz JS, Shero ST, Smith SC, Sorlie P, Stone NJ, Wilson PW (2014) 2013 acc/aha guideline on the assessment of cardiovascular risk. *Journal of the American College of Cardiology* 63(25 Part B):2935–2959, DOI 10.1016/j.jacc.2013.11.005, URL <https://www.onlinejacc.org/content/63/25.Part.B/2935>, <https://www.onlinejacc.org/content/63/25.Part.B/2935.full.pdf>
- Goldenshluger A, Zeevi A (2013) A linear response bandit problem. *Stochastic Systems* 3(1):230–261

- Hamburg MA, Collins FS (2010) The path to personalized medicine. *New England Journal of Medicine* 363(4):301–304
- Hansson GK (2005) Inflammation, atherosclerosis, and coronary artery disease. *New England Journal of Medicine* 352(16):1685–1695
- Ibrahim JG, Chen MH, Sinha D (2014) Bayesian survival analysis. Wiley StatsRef: Statistics Reference Online
- Imbens GW, Rubin DB (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, USA
- Kallus N (2017) Recursive partitioning for personalization using observational data. In: *International Conference on Machine Learning*, pp 1789–1798
- Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T (2017) Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology* 69(21):2657 – 2664, DOI <https://doi.org/10.1016/j.jacc.2017.03.571>, URL <http://www.sciencedirect.com/science/article/pii/S0735109717368456>
- Lagakos S (1979) General right censoring and its impact on the analysis of survival data. *Biometrics* 35(1):139–156
- Roeters van Lennep JE, Westerveld HT, Erkelens DW, van der Wall EE (2002) Risk factors for coronary heart disease: implications of gender. *Cardiovascular Research* 53(3):538–549, URL [http://dx.doi.org/10.1016/S0008-6363\(01\)00388-1](http://dx.doi.org/10.1016/S0008-6363(01)00388-1)
- Lesko L (2007) Personalized medicine: elusive dream or imminent reality? *Clinical Pharmacology & Therapeutics* 81(6):807–816
- Li L, Chu W, Langford J, Schapire RE (2010) A contextual-bandit approach to personalized news article recommendation. In: *Proceedings of the 19th international conference on World wide web*, ACM, pp 661–670
- Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) *Advances in Neural Information Processing Systems* 30, Curran Associates, Inc., pp 4765–4774, URL <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
- Nevin L, Editors PM, et al. (2018) Advancing the beneficial use of machine learning in health care and medicine: Toward a community understanding
- Omar AMS, Narula S, Rahman MAA, Pedrizzetti G, Raslan H, Rifaie O, Narula J, Sengupta PP (2017) Precision phenotyping in heart failure and pattern clustering of ultrasound data for the assessment of diastolic dysfunction. *JACC: Cardiovascular Imaging* 10(11):1291 – 1303, DOI <https://doi.org/10.1016/j.jcmg.2016.10.012>, URL <http://www.sciencedirect.com/science/article/pii/S1936878X16309792>
- Orfanoudaki A, Chesley E, Cadisch C, Stein B, Nouh A, Alberts MJ, Bertsimas D (2020) Machine learning provides evidence that stroke risk is not linear: The non-linear framingham stroke risk score. *PloS one* 15(5):e0232414
- Pearl J, et al. (2009) Causal inference in statistics: An overview. *Statistics surveys* 3:96–146
- Polonsky TS, McClelland RL, Jorgensen NW, Bild DE, Burke GL, Guerci AD, Greenland P (2010) Coronary artery calcium score and risk classification for coronary heart disease prediction. *Jama* 303(16):1610–1616
- Qian M, Murphy SA (2011) Performance guarantees for individualized treatment rules. *Annals of statistics* 39(2):1180
- Rejnmark L, Vestergaard P, Mosekilde L (2006) Treatment with beta-blockers, ace inhibitors, and calcium-channel blockers is associated with a reduced fracture risk: a nationwide case-control study. *Journal of hypertension* 24(3):581–589
- Ridker PM, Buring JE, Rifai N, Cook NR (2007) Development and Validation of Improved Algorithms for the Assessment of Global Cardiovascular Risk in Women The Reynolds Risk Score. *JAMA* 297(6):611–619, DOI 10.1001/jama.297.6.611, URL <https://doi.org/10.1001/jama.297.6.611>, <https://jamanetwork.com/journals/jama/articlepdf/205528/joc70004.611.619.pdf>
- Ron Kohavi FP (1998) Glossary of terms. *Machine Learning* 30:271–274
- Rosenbaum PR (2010) *Design of observational studies*, vol 10. Springer
- Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55, DOI 10.1093/biomet/70.1.41, URL <https://doi.org/10.1093/biomet/70.1.41>, <http://oup.prod.sis.lan/biomet/article-pdf/70/1/41/662954/70-1-41.pdf>
- Ross R (1999) Atherosclerosis—an inflammatory disease. *New England journal of medicine* 340(2):115–126

- Rubin DB (1990) Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* 5(4):472–480
- Schulz KF, Chalmers I, Hayes RJ, Altman DG (1995) Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Jama* 273(5):408–412
- Sedlis SP, Hartigan PM, Teo KK, Maron DJ, Spertus JA, Mancini GJ, Kostuk W, Chaitman BR, Berman D, Lorin JD, Dada M, Weintraub WS, Boden WE (2015) Effect of pci on long-term survival in patients with stable ischemic heart disease. *New England Journal of Medicine* 373(20):1937–1946, DOI 10.1056/NEJMoa1505532, URL <https://doi.org/10.1056/NEJMoa1505532>, PMID: 26559572, <https://doi.org/10.1056/NEJMoa1505532>
- Serruys PW, Morice MC, Kappetein AP, Colombo A, Holmes DR, Mack MJ, Stähle E, Feldman TE, van den Brand M, Bass EJ, Van Dyck N, Leadley K, Dawkins KD, Mohr FW (2009) Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease. *New England Journal of Medicine* 360(10):961–972, DOI 10.1056/NEJMoa0804626, URL <http://dx.doi.org/10.1056/NEJMoa0804626>, PMID: 19228612, <http://dx.doi.org/10.1056/NEJMoa0804626>
- Sianos G, Morel MA, Kappetein AP, Morice MC, Colombo A, Dawkins KD, van den Brand M, van Dyck N, Russell M, Serruys PW (2005) The syntax score: an angiographic tool grading the complexity of coronary artery disease. *EuroIntervention* 1(2):219–227, URL <https://www.pcronline.com/eurointervention/2nd.issue/36>
- Stoehlmacher J, Park D, Zhang W, Yang D, Groshen S, Zahedy S, Lenz H (2004) A multivariate analysis of genomic polymorphisms: prediction of clinical outcome to 5-fu/oxaliplatin combination chemotherapy in refractory colorectal cancer. *British journal of cancer* 91(2):344
- Stout KK, Daniels CJ, Aboulhosn JA, Bozkurt B, Broberg CS, Colman JM, Crumb SR, Dearani JA, Fuller S, Gurvitz M, et al. (2018) 2018 aha/acc guideline for the management of adults with congenital heart disease: a report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Circulation* pp CIR-0000000000000603
- Strom BL (2001) Data validity issues in using claims data. *Pharmacoepidemiology and drug safety* 10(5):389–392
- Tucker KL, Sheppard JP, Stevens R, Bosworth HB, Bove A, Bray EP, Earle K, George J, Godwin M, Green BB, et al. (2017) Self-monitoring of blood pressure in hypertension: A systematic review and individual patient data meta-analysis. *PLoS medicine* 14(9):e1002389
- 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
- Warnes CA (2017) Adult congenital heart disease: the challenges of a lifetime. *European Heart Journal* 38(26):2041–2047, URL <http://dx.doi.org/10.1093/eurheartj/ehw529>
- Wilson (2017) Estimation of cardiovascular risk in an individual patient without known cardiovascular disease. UpToDate, Waltham, MA
- Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB (1998) Prediction of coronary heart disease using risk factor categories. *Circulation* 97(18):1837–1847, DOI 10.1161/01.CIR.97.18.1837, URL <http://circ.ahajournals.org/content/97/18/1837>, <http://circ.ahajournals.org/content/97/18/1837.full.pdf>
- Zhou Y, Wilkinson D, Schreiber R, Pan R (2008) Large-scale parallel collaborative filtering for the netflix prize. In: *International conference on algorithmic applications in management*, Springer, pp 337–348

Appendix

Acronym	Acronym Definition
AHA	American Heart Association
ASA	Aspirin
AUC	Area Under the ROC Curve
BMC	Boston Medical Center
BMI	Body Mass Index
CABG	Coronary Artery Bypass Graft
CAD	Coronary Artery Disease
CART	Classification and Regression Trees
DMLA	Degree of ML Agreement
ECG	Electrocardiogram
EMR	Electronic Medical Records
FDA	US Food and Drug Administration
HDL	High-Density Lipoprotein
k-NN	k-Nearest Neighbors
LDL	Low-Density Lipoprotein
ML	Machine Learning
OCT	Optimal Classification Trees
ORT	Optimal Regression Trees
PE	Prescription Effectiveness
PR	Prescription Robustness
PCI	Percutaneous Coronary Intervention
ROC	Receiver Operator Characteristic
TAE	Time from diagnosis to a potential Adverse Event

List of all acronyms used in the manuscript in alphabetical order along with the corresponding definition.