Personalized treatment for CAD patients: A machine learning approach

AGNI ORFANOUDAKI
BOSTON INFORMS MEETING, JANUARY 2020
JOINT WORK WITH PROF. DIMITRIS BERTSIMAS & RORY B. WEINER
OPERATIONS RESEARCH CENTER, MIT
Overview

1. Introduction
2. Problem & Data Description
3. Addressing Censoring
4. Predictive Models
5. The Prescriptive Algorithm
6. Evaluating Counterfactuals
7. Results
8. The Online Application
9. Conclusions
Heart Disease is the No. 1 cause of death in the United States, killing over 370,000 people a year

- **Coronary Artery Disease (CAD)** is the medical term for the **buildup of plaque** in the arteries that could lead to a **heart attack** or a **stroke**.

- About **750,000** people in the U.S. experience **heart attacks** each year due to CAD.

- **Costs of CAD** total more than **$182 billion**.

Source: American Heart Association
Physicians choose among three options to treat CAD patients

1. Coronary Artery Bypass Surgery (CABG) + Medication
2. Percutaneous Coronary Intervention (PCI) + Medication
3. Medication

Sources: UptoDate, American Heart Association
The medical community has not established a globally accepted decision making process for the treatment of CAD

**Reasons:**

- Numerous parameters
- Multitude of complications
- Paucity of information for special subgroups
The magnitude and the repercussions of the disease give rise to our research question...

“What is the optimal treatment for a patient with Coronary Artery Disease in order to maximize the time from diagnosis to a potential adverse event?”
Analytics provides the key components to address it
Data: Electronic Health Records from the BMC
Models: Machine Learning & AI

Data

Models

Decisions
Decisions: Online Adaptive Support Tool
We developed a prescriptive algorithm that personalizes the treatment decision for patients using analytics.
The data from the Boston Medical Center

- BMC is a private, 487-bed, non-profit, academic medical center located in Boston.
- Our partnership with BMC gave us access to EMR for **1.1 million** patients from **1982 to 2016**.
Patients at BMC share a unique characteristic

BMC patients come predominantly from underprivileged socioeconomic backgrounds.

- In most cases they do not have the financial capability to support alternative health providers.
- Thus, they appeal to the BMC for healthcare services for the vast majority of their medical needs.

As a result, the vast majority of their EMR are in the BMC database.

We can follow their health trajectory from a single source.
We identified 21,460 patients who met at least one of our inclusion criteria

- There was not a distinct diagnosis code that labels all the CAD patients.
- We defined our own **inclusion criteria** using medical insight from our clinical collaborators:
  1. Patients who were administered **at least one CABG surgery or at least one PCI** in the BMC along with antihypertensive medication.
  2. Patients associated with **CAD risk of at least 10% based on the Framingham Heart Study** who were prescribed antihypertensive medication.

The Framingham Heart Study (FHS)
The underlying demographic characteristics of our sample are significantly different from the overall US population.
We identified the key risk factors that contribute to CAD based on the medical literature

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Treatment</th>
<th>Family History</th>
<th>Medical Records</th>
<th>Observed Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>CABG</td>
<td>Diabetes</td>
<td>BMI</td>
<td>Smoking</td>
</tr>
<tr>
<td>Gender</td>
<td>PCI</td>
<td>Hypertension</td>
<td>LDL</td>
<td>Time observed in EMR database</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Blockers (beta, alpha etc)</td>
<td></td>
<td>HDL</td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td>ACE inhibitors</td>
<td></td>
<td>Diastolic Blood Pressure</td>
<td></td>
</tr>
<tr>
<td>Marital Status</td>
<td>Nitrates</td>
<td></td>
<td>Systolic Blood Pressure</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diuretics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statins</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Antiarrhythmics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Angiotensin Agonists</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adrenergic Receptors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lipid Lowering medication</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cardiac Glycosides</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Muscle relaxants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Phosphodiesterase inhibitors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other antihypertensive</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The EHR database contained patients with incomplete information regarding the outcome of their treatment. Some patients disappeared from the EMR before the defined time frame of 10 years after diagnosis. The outcome of their treatment was unknown.
The $K$-Nearest Neighbors Method

- Heuristic method for unsupervised and supervised learning.
- For each observation $x$, the value of $\hat{Y}(x)$ is given by:
  \[ \hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \]
  where $N_k(x)$ is the set of $k$ points that belong in the neighborhood of $x$.
- We can use any distance metric to measure the distance between two points A and B,
The *K*-Nearest Neighbors Method (*K*=4)

- Heuristic method for unsupervised and supervised learning.
- For each observation *x*, the value of $\hat{y}(x)$ is given by:
  $$\hat{y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$
- where $N_k(x)$ is the set of *k* points that belong in the neighborhood of *x*.
- We can use any distance metric to measure the distance between two points A and B,
The $K$-Nearest Neighbors Method ($K=5$)

- Heuristic method for unsupervised and supervised learning.
- For each observation $x$, the value of $\hat{Y}(x)$ is given by:
  $$\hat{Y}(x) = \frac{1}{k} \sum_{i \in N_k(x)} Y_i$$
- where $N_k(x)$ is the set of $k$ points that belong in the neighborhood of $x$.
- We can use any distance metric to measure the distance between two points A and B,
We developed a novel k-NN based method to estimate the time to adverse event for the right censored population.

We considered the following sets of the population:

- **A**: patients that experienced an adverse event within 10 years
- **B**: patients that did not experience an adverse event within 10 years
- **C**: censored patients that did not have an adverse event within a time $t_0$ ($t_0 < 10$ years)

To estimate the time to adverse event for a patient $X$ in $C$, we consider patients within $A \cup B$ such that:
1. They have the same gender as $X$
2. They belong to the same age group as $X$
3. Their standard of care outcome metric is greater than or equal to the censoring time $t_0$ of $X$
We evaluated the accuracy of our k-NN estimation using the $R^2$ on the times of the non-censored population.

We introduce a new **approach** for **testing the performance** of our algorithm:

1. Select a sample of the population which was not censored (the time to adverse event is known).
2. Artificially generate a censoring time, sampled uniformly across the interval $[1, 3650]$ corresponding to a day in the 10 year time frame.
3. Apply the k-NN algorithm to estimate the time to adverse event.
4. Compare the results with the ground truth that is known.
5. Calculate the $R^2$.

**Prediction Accuracy:** $R^2 = 0.813$
Our method resulted in the creation of personalized survival curves for every censored patient.
We use Machine Learning to Build Predictive Models

Independent covariates from the BMC database:
\[ X_i = (x_{i,1}, x_{i,2}, ..., x_{i,p}) \] where \( i = 1, ..., N \)

Combination of known and estimated times to adverse event:
\[ T = (t_1, t_2, ... t_N) \]

1. **Binary** outcome for the occurrence of an adverse event for a 10-year time frame:
\[ Y = (y_1, y_2, ..., y_N) \]
\[ y_i = \begin{cases} 1 & \text{if } t_i < 3650 \\ 0 & \text{otherwise} \end{cases} \]

2. **Continuous** outcome for the occurrence of an adverse event for a 10-year time frame:
\[ T = (t_1, t_2, ... t_N) \]
Methods Overview: Logistic regression

- A Logistic Regression model predicts the **probability** of a patient experiencing an adverse event within 10 years, $P(Y = 1)$.

- Suppose that a patient has features $X_1, X_2, ..., X_p$.

- The Logistic Regression formula to calculate risk is:

$$P(Y = 1) = \frac{1}{1+e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}}.$$

The output of the Logistic Regression function is a set of values for all $\beta$ coefficients.

- Positive coefficients - increase in odds of having an adverse event (higher risk).

- Negative coefficients - decrease in odds of having an adverse event (lower risk).
Methods Overview: Tree Based Methods

- **Weight < 165 lbs**
  - **Yes**: 10%
  - **Yes**: 100%
  - **No**: 0%

- **Age < 51**
  - **Yes**: 0%
  - **No**: 100%
Methods Overview: Tree Based Methods

- **Low Risk Class:** p = 100%
  - **Weight:** < 165 lbs, **Age:** < 51
  - **Split 1:** Weight < 165 lbs
  - **Split 2:** Age < 51
- **High Risk Class:** p = 100%
  - **Weight:** < 165 lbs, **Age:** ≥ 51
  - **Split 1:** Weight < 165 lbs

- **Low Risk Class:** p = 90%
  - **Weight:** ≥ 165 lbs, **Age:** ≥ 51
Methods Overview: CART and OCT

- **CART**: Top-down approach to building a decision tree.
  - Greedy heuristic recursively starting with the full population in the top node and creating each subsequent split in isolation.

- **OCT**: Creates the entire decision tree at once through an optimization approach, resulting in more accurate results than its predecessor method.
Methods Overview: Ensemble Methods

- **Ensemble Methods**: Reliance on a large group of trees to generate predictions.
- Random Forest and XGBoost are the more characteristic examples.

![Diagram](image)

- The boundary to be learned (it cannot be recovered by a tree with parallel splits.)
- The boundary recovered by one parallel-split classification tree
- Smoother boundary found by an RF model averaged across the thresholds of many trees.
A. Binary Outcome: Risk of an Adverse Event

<table>
<thead>
<tr>
<th></th>
<th>Out-of-sample AUC</th>
<th>In-sample AUC</th>
<th>Out-of-sample Accuracy</th>
<th>In-sample Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OCT</strong></td>
<td>81.54%</td>
<td>81.35%</td>
<td>81.45%</td>
<td>81.36%</td>
</tr>
<tr>
<td><strong>CART</strong></td>
<td>73.33%</td>
<td>72.66%</td>
<td>80.23%</td>
<td>80.12%</td>
</tr>
<tr>
<td><strong>Random Forest</strong></td>
<td><strong>84.29%</strong></td>
<td><strong>83.29%</strong></td>
<td><strong>81.88%</strong></td>
<td><strong>82.35%</strong></td>
</tr>
<tr>
<td><strong>Logistic Regression</strong></td>
<td>80.83%</td>
<td>82.21%</td>
<td>80.55%</td>
<td>80.98%</td>
</tr>
<tr>
<td><strong>Gradient Boosted Trees</strong></td>
<td>81.43%</td>
<td>82.76%</td>
<td>81.03%</td>
<td>81.27%</td>
</tr>
<tr>
<td><strong>Baseline</strong></td>
<td>73.51%</td>
<td>73.51%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- We chose **OCT** as our main predictive model due to its **superior performance** and high interpretability.
- Random Forest yields better AUC results compared to OCT, although quite similar in terms of accuracy. However, it does not provide us with any insights on the key medical decision rules used to classify the patients.
Patients that performed surgery right after visiting the hospital are at high risk.

For those patients, that followed cholesterol treatment and HDL levels remained low, the prognosis is more pessimistic.
The models highlight the importance of some variables that are validated by the medical literature

- **Time in the database**: the time that the patient has been observed in the BMC database, indicating the information depth regarding his/her medical condition and history.

- **BMI**: higher BMI is a risk factor that increases the likelihood of experiencing an adverse event.

- **Race** seems to be a key differentiating factor to the development of the disease.

- **Older people** are more prone to have severe symptoms of CAD.

- **Diabetes** as well as *smoking* increase the chances of experiencing an adverse event within a 10-year time frame.

- The **type of medical and surgical treatment** received by the patient affect the development of his/her condition to a very significant level.
Each patient was administered a combination of different types of medication and potentially a surgical intervention.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>% of patients prescribed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nitrates</td>
<td>77.02%</td>
</tr>
<tr>
<td>Blockers (beta, alpha, etc.)*</td>
<td>68.03%</td>
</tr>
<tr>
<td>Statins*</td>
<td>58.78%</td>
</tr>
<tr>
<td>Diuretics</td>
<td>47.90%</td>
</tr>
<tr>
<td>ACE inhibitors</td>
<td>46.12%</td>
</tr>
<tr>
<td>PCI*</td>
<td>19.60%</td>
</tr>
<tr>
<td>Antiarrhythmics</td>
<td>13.65%</td>
</tr>
<tr>
<td>Angiotensin Agonists</td>
<td>13.62%</td>
</tr>
<tr>
<td>Other antihypertensive</td>
<td>11.37%</td>
</tr>
<tr>
<td>CABG*</td>
<td>7.01%</td>
</tr>
<tr>
<td>Adrenergic Receptors</td>
<td>6.38%</td>
</tr>
<tr>
<td>Lipid Lowering medication</td>
<td>5.29%</td>
</tr>
<tr>
<td>Muscle relaxants</td>
<td>4.81%</td>
</tr>
<tr>
<td>Phosphodiesterase inhibitors</td>
<td>3.59%</td>
</tr>
<tr>
<td>Cardiac Glycosides</td>
<td>2.45%</td>
</tr>
</tbody>
</table>
We selected as prescription options the treatment options that affect most significantly the trajectory of the patient.

**Prescription Options**
- CABG with medication
- PCI with medication
- Blockers with statins (Medication 1)
- Blockers without statins (Medication 2)
- Other medication without blockers (Medication 3)
We separate patients into groups based on the treatment they received and fit separate regression models:

- CABG with medication - N = 1,854
- PCI with medication - N = 4,042
- Blockers & Statins - N = 6,833
- Blockers – Statins - N = 3,767
- Other medication without Blockers - N = 4,964

- ORT
- CART
- Random Forest
- Linear Regression
- Gradient Boosted Trees
B. Continuous Outcome: Time to an Adverse Event

\[ R^2 \] metric on the testing dataset

<table>
<thead>
<tr>
<th></th>
<th>ORT</th>
<th>CART</th>
<th>Random Forests</th>
<th>Linear Regression</th>
<th>Gradient Boosted Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>CABG</td>
<td>73.14%</td>
<td>71.91%</td>
<td>83.00%</td>
<td>80.32%</td>
<td>80.06%</td>
</tr>
<tr>
<td>PCI</td>
<td>68.30%</td>
<td>67.73%</td>
<td>74.58%</td>
<td>73.21%</td>
<td>73.21%</td>
</tr>
<tr>
<td>Medication 1</td>
<td>78.64%</td>
<td>75.35%</td>
<td>83.92%</td>
<td>82.94%</td>
<td>82.48%</td>
</tr>
<tr>
<td>Medication 2</td>
<td>73.46%</td>
<td>72.56%</td>
<td>80.02%</td>
<td>79.98%</td>
<td>79.50%</td>
</tr>
<tr>
<td>Medication 3</td>
<td>67.10%</td>
<td>69.03%</td>
<td>77.71%</td>
<td>75.34%</td>
<td>75.29%</td>
</tr>
</tbody>
</table>

- Similar conclusions to what we observed in the binary classification models
- Basis for the prescriptive algorithm
Going from Prediction to Prescription: ML4CAD

Standard of care treatment: Medication 3

4.02 years

ML4CAD recommends CABG surgery with expected time 5.82 years and stand. deviation 1.01 years.
How do we evaluate a prescriptive algorithm when the counterfactuals are unknown?

For each patient X we only know the time to an adverse event for one treatment.

- It is impossible to evaluate the algorithm’s performance across all outcomes.
- How can we get an estimation of its performance?
  1. Assess its effectiveness across all cases.
  2. Consider Ground Truths and compare against those.
  3. Evaluate its accuracy in the cases for which we have data (20%).
  4. Compare with the treatment patterns from the BMC database.
The evaluation process of the prescriptive algorithm is requires the formal definition of the problem

Assessing the quality of the prescriptive algorithm poses a challenge.

We can define the problem in the following way. 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 predictive models;
- \( z_i \) be the treatment that patient \( x_i \) followed at the standard of care;
- \( t_i \) be the TAE for patient \( x_i \) for the treatment \( z_i \);
- \( \tau_i \) be the treatment recommendation of ML4CAD for patient \( i \);
- \( g_i^j(p) \) be the estimated TAE for patient \( i \) for treatment \( p \) from regression model \( j \);
- \( y_i(p) \) be the estimated TAE for patient \( i \) for the treatment \( p \) from ML4CAD.
Prescription Effectiveness

- Evaluates the algorithm’s effect against the standard of care.

\[ PE(ML4CAD) = \frac{1}{n} \sum_{i=1}^{n} y_i(\tau_i) - t_i, \quad i \in [n] \]

We can also calculate the prescription effectiveness of the \( \tau_i \) recommendation for a specific ML model \( j \)

\[ PE(ML_j) = \frac{1}{n} \sum_{i=1}^{n} g_i^j(\tau_i) - t_i, \quad i \in [n], \forall j \in [M] \]

Summarizes by how much we can improve on average the TAE and the impact of the algorithm's recommendations to the health condition of the patients under consideration.
Prescription Robustness

- Evaluates the algorithm’s effect against the standard of care under various ground truths.

\[ PR(ML4CAD) = \frac{1}{n} \sum_{i=1}^{n} y_i(\tau_i) - g^k_i(z_i), \quad i \in [n], \quad \forall k \in [M] \]

We can also calculate the prescription robustness of the \( \tau_i \) recommendation for specific ML models \( j \) and \( k \)

\[ PR(ML_{j,k}) = \frac{1}{n} \sum_{i=1}^{n} g^j_i(\tau_i) - g^k_i(z_i), \quad i \in [n], \quad \forall j, k \in [M] \]

That way we can increase our confidence regarding the algorithm’s performance.
## Prescription Effectiveness & Robustness Results

<table>
<thead>
<tr>
<th>Estimation Model</th>
<th>Baseline (BMC)</th>
<th>ORT</th>
<th>CART</th>
<th>Random Forest</th>
<th>Linear Regression</th>
<th>Boosted Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML4CAD</td>
<td><strong>1.101</strong></td>
<td>1.162</td>
<td>1.158</td>
<td>1.140</td>
<td>1.178</td>
<td>1.283</td>
</tr>
<tr>
<td>ORT</td>
<td>0.779</td>
<td>0.842</td>
<td>0.835</td>
<td>0.818</td>
<td>0.855</td>
<td>0.961</td>
</tr>
<tr>
<td>CART</td>
<td>0.923</td>
<td>0.983</td>
<td>0.979</td>
<td>0.965</td>
<td>0.999</td>
<td>1.105</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.757</td>
<td>0.818</td>
<td>0.813</td>
<td>0.796</td>
<td><strong>0.833</strong></td>
<td>0.939</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.485</td>
<td>0.546</td>
<td>0.541</td>
<td>0.524</td>
<td>0.561</td>
<td>0.667</td>
</tr>
<tr>
<td>Boosted Trees</td>
<td>0.591</td>
<td>0.652</td>
<td>0.647</td>
<td>0.63</td>
<td>0.667</td>
<td>0.773</td>
</tr>
</tbody>
</table>
We can compare the results for specific population subgroups

- **Hispanic** patients benefit the most: +1.86 years, 58.41% improvement.
- **Caucasian** patients: +1.16 years, 29.03% improvement.
- **Black** patients: +0.9 years, 17.09% improvement.
We can compare the results for specific population subgroups

- **65-80** patients: +0.99 years, 22.5% improvement.
- **80+** patients: +1.58 years, 46.9% improvement.
Prediction Accuracy – an adjusted $R^2$

- **ML4CAD**

$$\hat{R}^2(\text{ML4CAD}) = 1 - \frac{\sum_{i \in S} (v_i(z_i) - t_i)^2}{\sum_{i \in S} (t_{zi} - t_i)^2}, \quad S = \{i: \tau_i = z_i\}, \quad i \in [n]$$

Same structure as the well-known coefficient of determination $R^2$ but applied for each patient only to the prescriptive option whose outcome is known by the standard of care.

- **Individual Regression Models**

$$\hat{R}^2(\text{ML}_j) = 1 - \frac{\sum_{i \in S} (g_i^j(z_i) - t_i)^2}{\sum_{i \in S} (t_{zi} - t_i)^2}, \quad S = \{i: \theta_i^j = z_i\}, \quad i \in [n]$$

<table>
<thead>
<tr>
<th>Method</th>
<th>ML4CAD</th>
<th>ORT</th>
<th>CART</th>
<th>Random Forest</th>
<th>Linear Regression</th>
<th>Boosted Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{R}^2$</td>
<td>78.70%</td>
<td>72.68%</td>
<td>70.54%</td>
<td>77.25%</td>
<td>76.66%</td>
<td>76.59%</td>
</tr>
</tbody>
</table>
Comparing the treatment allocation patterns

- **Shift towards Medication 1**: Beta Blockers + Statins
- **Increase of CABG** and **decrease** of PCI
We bridge the gap with practitioners by offering an online interface.
Limitations

- The **censoring** problem was solved with **heuristic** methods, rather than an optimization technique.
- Patients were not **randomized** into treatment groups.
- Our data do not include socioeconomic factors or patient preferences that may be important in treatment decisions (i.e. income, fear of invasive treatment strategies).
- We can only estimate counterfactual outcomes.
- The study **population** of BMC is not representative of the general U.S. population.
- The **accuracy of the prediction model is limited**, though significantly better than the baseline model.
- We did not consider the different types of CABG surgery and PCI.
- We do not include drug specific recommendations.
We introduced the first personalized prescriptive algorithm that utilizes EMR data in the cardiovascular field

- **ML** may help **facilitate identification** of the **optimal treatment** strategy:
  - Reduce symptoms
  - Increase length of life for an individual patient.
- Our **findings** are **consistent** with **themes** that have emerged in **clinical trials**.
- The **ML4CAD** approach is **accurate**, highly **interpretable** and **flexible** in healthcare applications.
- Our **method accommodates alternative cardiovascular disease-management** approaches within specific disease subpopulations, such as arrhythmia and valvular disease management.
Thank you!

Questions?


Link to the App:
https://personalized.shinyapps.io/ml4cad/