Predicting Cardiovascular Risk using Electronic Health Records

Minh Nguyen
Department of Biomedical Informatics, Stanford University

Abstract

Atherosclerotic Cardiovascular Disease (ASCVDs) is the number 1 cause of death globally. Identifying at-risk individuals is important for early prevention and timely treatments. Although there is an abundance of prediction models for ASCVDs, the use of EHRs is still at the beginning. We utilized the large Electronic Health Records from the Stanford Translational Research Integrated Database Environment (STRIDE 8) to derive a valid cohort and build a gradient boosting model predicting CVD risks. Our model performed very well compared to the standard risk prediction scores and neural networks results from the literature, but raised questions regarding fairness when applying to subgroups with sensitive attributes such as gender, age, and race/ethnicity.

1 Introduction

Atherosclerotic cardiovascular disease (ASCVD), a common in the general population, generally refers to heart and blood vessel diseases where the blood vessels are narrowed or blocked due to plaque builds up. This can prevent blood flow and allow opportunities for blood clot formation when ruptured. The consequences including a heart attack and a stroke. Cardiovascular disease is a leading cause of morbidity and mortality worldwide (Wessler et al., 2015). In the United States, about 610,000 people die of heart disease every year, and that is 1 in every 4 deaths (cdc.gov).

In 2013, the American College of Cardiology and American Heart Association (ACC/AHA) released updated guideline on the assessment of cardiovascular risk, with a Pooled Cohort Equation (PCE) to predict 10-year risk of first ASCVD. Since then, there have been numerous studies evaluating the performance of PCEs with mixed results (Karmali, 2017). Early prevention is important and this involves identifying at-risk individuals. Although there is an abundance in predictions models developed in the recent years, the use of electronic health records (EHRs) in predicting ASCVD risks is still at the beginning. However, the rise of machine learning applied in big data provides an opportunity for mining EHRs to develop models predicting ASCVD risk.

This project aimed to build a model to predict ASCVD risk using EHRs from Stanford Translational Research Integrated Database Environment (STRIDE), version 8. The input to our algorithm is a data frame consists of about 40,000 features and outcome labels as ASCVD events. We then used a logistic regression model as our baseline and a final gradient boosting model to output predicted probabilities of having an ASCVD event on individual levels.

Recently, the notion of fairness in machine learning has increasingly gained attention. To simplify, a model is considered fair if it performs consistently similar for members of subgroups based on some defined fairness metrics. With this in mind, we will also evaluate the model performance for subgroups with sensitive attributes (such as race, gender, and age group) as an exploration step to motivate further study on fairness in modeling.

CS229: Machine Learning, Spring 2019, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Related work

- **Predictions without using EHR data:** Traditionally in medicine, prediction models select a specific and limited number of features. A systematic review of prediction models for cardiovascular disease risk in the general population were done in 2016 showed the median number of predictors was 7, and the range was 2-80 (Damen et al., 2016). Commonly used predictors include age, smoking, blood pressure, blood cholesterol measurements, diabetes, and body mass index. In terms of methods, most used Cox proportional hazards and logistic regression, with area under the receiver operating characteristic curve (AUROC) ranged from 0.61 to 1.00 (Damen et al., 2016). These simpler models are sometimes preferred because they are easier to understand and interpret. Despite the abundance of models, wide variation makes direct comparisons and validations difficult, and implementations are uncertain. Moreover, most prediction models were developed in North American and European populations (Damen et al., 2016). This suggests a need for models that include more diverse populations, tailor to more local settings, expand with new predictors, and utilize different sources of data such as EHRs.
- **Prediction with EHRs:** A different systematic review on the use of EHRs in risk prediction models suggests both opportunities and challenges. EHR data is larger with more patients, more features, and data points collected more frequently; thus, it seems to reflect more of real-world data. However, it lacks the standardization, strict controls and structures that large cohort studies have, so EHR data can be very messy. For example, outcomes and diagnoses vary depending on the source such as billing codes and problem lists, and heterogeneity in outcome definitions remains an issue (Goldstein et al., 2017). To our surprise, the use of Cox models and generalized linear regression also remains most commonly used with EHR data, and with a median 27 features. Other approaches include Bayesian methods, tree-based, and regularized regressions with LASSO and Ridge (Goldstein et al., 2017). Most of the reviewed studies used AUROC as the evaluation metrics, and model performance is worse the further out the prediction is done. A recent buzz-worthy paper (Rajkomar et al., 2018), employed a complex weighted RNN and LSTM model for predicting different medical outcomes using EHR data. However, in the supplemental readings, their baseline model's AUCROC is only 0.01 - 0.02 lower than the deep learning model. For example, with 30-day readmission, deep learning model performance for hospital B had an AUROC of 0.76 vs 0.75 for baseline logistic regression model trained using the Adam optimizer with early-stopping (Rajkomar et al., 2018).
- **Recent developments using EHRs in ASCVD risk predictions:** Using EHR data to assess the common risk score equations like QRISK2, FRS and PCE (Pike, 2016) and (Wolfson et al., 2017) evaluated these score performance with AUROC ranged from 0.6 - 0.75. (Weng et al., 2017) compared the performance of ACC/AHA risk score PCE to machine-learning algorithms, and concluded that machine-learning algorithms improved prediction. Whereas PCE achieved an AUROC of 0.728, random forest, logistic regression, gradient boosting and neural networks achieved AUROC of 0.745, 0.760, 0.761 and 0.764 respectively. (Zhao et al., 2019) compared the performance of PCE with had an average AUROC of 0.732 to those of machine learning models (logistic regression, random forest and gradient boosting trees) with AUROC of 0.765 – 0.782. When utilizing longitudinal features, gradient boosting trees and convolutional nerural networks achieved equally highest performance AUROC of 0.790 (Weng et al., 2019).

3 Dataset and Features

To explore and utilize the wealth of EHR data, we derived a large cohort from STRIDE8 to build a prediction model for ASCVD risk. STRIDE data contains clinical information on over 1.3 million pediatric and adult patients Stanford University Medical Center since 1995 (Lowe, 2009). We followed the cohort definition from the 2013 ACC/AHA guideline on the assessment of cardiovascular risk. More specifically, our inclusion/exclusion criteria are closely aligned with (Mutner et al., 2014) study.

We excluded patients who had any history of ASCVD incidence, based on ICD9 and ICD10 diagnose codes, such as myocardial infarction, coronary heart disease, stroke, atrial fibrillation, and heart failure. We also excluded patients who were on antilipidemic medications. This is a significant

change from the previously published paper of our lab (Pfohl, 2018). Outcome is defined as ASCVD event of myocardial infarction, stroke, or death within a year of a coronary heart disease diagnose. Due to the revised algorithm, our cohort has changed, with a significance reduction in positive outcome labels. Selected patients had ages ranged from 40 to 90. Random prediction times were chosen during valid time frames between their first visit and last visit such that patients have at least a year of history and a year of follow-up.

For feature extractions, we utilized the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM), as described in (Reps et al. 2018) to covert a series of time-stamped clinical elements across different concept domains (mapped to SNOMED) to a static presentation for each patient (Pfohl, 2018). This standard method has been used widely in the Observational Health Data Sciences and Informatics community, enabling model sharing and reproducibility. Features include diagnoses, conditions, procedures, medication orders, lab tests, measurements, clinical encounter types, departments, and other observations. Each feature is a binary value that represents whether a concept is present or missing at a certain time anywhere in the patient's history prior to the prediction time. The final complete data for modeling has 256,583 patients and 39,558 features in form of a sparse matrix of features. The percentage of positive outcome labels in our cohort is about 1.7%.

We explored the incidence rate in some subgroups with sensitive attributes such as gender, race/ethnicity and age groups. We found that although there are more female patients in our cohort, males have higher incidence of having ASCVD events compared to females (2.1% vs. 1.4%). Among different known race/ethnicity, White group is the majority, about 4 times more than Asians, 6.4 times Hispanic/Latino and almost 15 times Blacks. However, Blacks has the highest incidence rate of ASCVD events of 3.2% compared to the rest where their incidence rates are about 1.7 - 1.8%. For age groups, the ASCVD incidence rate is monotonically increasing with age.

We split the data into 80% of training and 20% of test. With 80% of the data for model training, we used 10-fold cross validation while training and evaluating.

4 Methods

We chose to use logistic regression and gradient boosting trees as two machine learning algorithms for our prediction task. This was decided from reviewing the literature to select good performing algorithms (that are not neural networks). Logistic regression is a fair choice as a baseline model, and gradient boosting trees seem to perform consistently well from literature review. There was also a practical purpose as glmnet and xgboost are two R packages that support sparse matrices. Our data is quite large and we had to use a sparse matrix feature for the purpose of storage and speed.

Baseline model - Logistic regression with regularization:

Our outcome is binary, indicating whether or not a patient has an ASCVD event after their prediction time. For modeling binary outcomes, logistic regression is probably the most commonly used model. Logistic regression, which is an extension of linear regression, models the probabilities for classification problems with categorical outcomes. By choosing a threshold between 0 and 1, we could then classify the predicted outcomes into 2 categories.

With a large number of features in our data, regularization is a necessary technique for preventing over-fitting to train a better model so that when we use the model on a completely unseen data, the model can be expected to generalize and perform reasonably well. The two well-known regularization techniques in regression are Lasso and Ridge, adding constrains to shrink the coefficient estimates, which reduces model complexity. In addition to shrinkage, Lasso also results in feature selection, which could be more useful on our high dimensional data set. Regularization therefore helps reduce the variance of the model without substantially compromise bias.

The objective function for the regularized logistic regression is:

$$\min_{(\eta_0,\beta)\in\mathbb{R}^{p+1}} -\left[\frac{1}{N}\sum_{i=1}^{N} y_i \cdot (\beta_0 + x_i^T\beta) - \log(1+e^{(\beta_0+x_i^T\beta)})\right] + \lambda\left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\right] \quad (1)$$

When applying regularization, Ridge retains all of the features but shrinks the coefficients whereas Lasso sets some correlated variables to zero. A combination of both, also known as elastic net, can be used to achieve better performance.

Gradient Boosting Decision Trees, with regularization - XGboost

Gradient boosting model uses a decision tree ensemble consisting of a set of classification and regression trees. These trees in boosting are built sequentially such that each subsequent tree learns from the trees that lead to it and aims to reduce the errors of the previous trees. Although each tree can be a weak learner (also called a model, or a regression tree), it still contributes valuable information to the overall prediction task. In addition, an ensemble of these weak learners can build a stronger learner. Each tree has 2 kinds of parameters: weight of each leaf and number of leaves. When building a tree, to find the best split for each tree, greedy algorithm is used for small and medium data set. With large data set, approximate algorithm is used instead (Chen and Guestrin, 2016).

The model trained in an additive manner, in which the learning objective function is a set of additive regression trees, where each tree contains continuous scores on the leaves. These are used to calculate the final predicted probabilities by summing up the scores in the corresponding leaves (Chen and Guestrin, 2016).

The objective function to be minimized is:

$$\mathcal{L}(\theta) = \sum_i^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad \text{where} \quad \Omega(f) = \gamma T + \tfrac{1}{2}\lambda \|w\|^2 \qquad (2)$$

A gradient descent algorithm is used to minimize the loss when adding new regression trees. Here, ℓ is a differentiable convex loss function that measures the difference between the predicted values and the truth. Decision trees are one of the most easily interpretable models, but usually with high variance. And $\Omega(f)$ is a regularised term that discourages the complexity of the model to help reduce variance (Chen and Guestrin, 2016).

We used XGBoost as an implementation of gradient boosting decision trees, which is designed for speed, scalability, and performance. Xgboost also implements L1 and L2 regularization as in Lasso and Ridge regression, with many hyperparameters to tune. We did not use regularization with column sub-sampling.

5 Experiments/Results:

We chose the area under the receiving operating characteristic curve (AUROC) as our evaluation metric as it does not depend on the threshold for classification, and is helpful in case of binary outcomes, especially with imbalanced classes. Another commonly used evaluation metric for imbalanced classes is the area under precision and recall curve (AUPRC). However, from our literature review, most used AUROC, so we also used AUROC as our primary evaluation metric for comparison purpose.

Hyperparameter Tuning: We split the data into 80% training and 20% testing. With the training data, tuning was done using 10 fold cross validation for both logistic regression and gradient boosting trees, minimizing the objective loss functions. AUROC was used as the evaluation metric for the validation data during cross validation and for evaluating model performance on the test data subsequently.

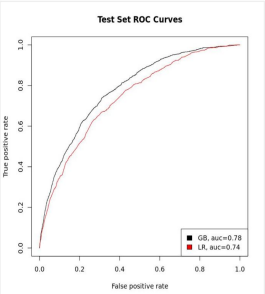
Logistic regression: We modeled logistic regression with 3 cases of regularization: Ridge only ($\alpha = 0$), Lasso only ($\alpha = 1$), and both (α in between 0 and 1). The regularization hyperparameter λ was tuned. It turned out that using an elastic net, which is a mix of L1 and L2 norms or a compromise between Lasso and Ridge, shrinks and does a sparse feature selection simultaneously. This yielded better results than just using either one of the two techniques alone. The AUROC for the 3 models are: 0.728, 0.741, 0.731 for Ridge, elastic net (α = 0.5), and Lasso respectively.

Gradient boosting trees: We first tuned manually and then a grid search over more than 100 models, tuning selected hyperparameters to avoid over-fitting. For maximum number of trees to grow, it was done in parallel with early stopping to stop the training of the model once the performance on the validation set has not improved after 5 iterations. For maximum depth of a tree, although our dataset is large, we used small numbers of depth (4, 6, 10), making the model less complex. After getting the weights of new features at each boosting step, learning rate (0.025, 0.05, 0.1) shrinks the these weights to make the boosting process more conservative. Last, we also tuned lambda and alpha (0.2, 0.5, 0.8), controlling for L2 regularization (Ridge regression) and L1 regularization (Lasso regression) on weights respectively.

To address a major property of our data, which is extremely imbalanced outcome labels where the ASCVD incidence rate is only 1.7%, we looked into different approaches, but did not find anything every satisfactory. With xgboost, we tuned the maximum delta step to help with convergence but void re-balancing the dataset and aim for predicting the right probabilities.

When we evaluated the final model performance on subgroups with sensitive attributes such as gender and race, we found that there are moderate gaps among these groups, which suggests that the model is not fair. This can have negative social and health implication when it comes to decision making.

Overall, with both logistic regression and gradient boosting trees, we tuned aggressively to avoid over-fitting and be conservative with our models considering we had a large number of features. Judging from the very narrow gap between average validation and test set AUROC, we think our model did not over-fit.



Confusion Matrix (threshold is 3rd quantile of predicted probabilities)

Prediction	Reference 0	1
0	38193	294
1	12275	554

Other metrics	
Sensitivity	**0.6533**
Specificity	**0.7568**
Accuracy	**0.7511**

Model (built)	AUROC	Model (literature review)	AUROC
Logistic Reg (w/ Ridge)	**0.728**	Risk score eq (QRISK2, FRS, PCE)	**0.60 - 0.75**
Logistic Reg (w/ Lasso)	**0.731**	Logistic regression, Tree-based	**0.765 - 0.782**
Logistic Reg (w/ elastic net)	**0.741**	GBM with longitudinal data	**~ 0. 790**
Best tuned GBM	***0.782***	NN with longitudinal data	***~ 0.790***

AUROC evaluated on subgroups w/ sensitive attributes (*gender*, race/ethnicity)			
Male	***0.744***	Asians	**0.792**
Female	***0.794***	Black	**0.781**
White	**0.767**	Hispanic/Latino	**0.743**

6 Discussion:

To handle the extreme class imbalance issue, we explored up and down sampling, Synthetic Minority Over-sampling Technique (SMOTE), and adding weights in loss functions. We avoided up-sampling as this would result in dependent observations and add more computation burden. Using the original data, it took over 6 hours to train a logistic regression model with 10 fold cross-validation. With down sampling, the results did not improve as much. We attempted this with binomial deviance metric and did not have enough time to redo with AUC metric for comparison. SMOTE is a clever technique in which it creates synthetic observations from the existing minority observations using the k-nearest neighbors, couples with under-sampling the majority class. Although we explored this technique, we did not implement it as there was no support for sparse matrix with SMOTE in R. These techniques however introduce a bias as they select more or less samples from one class than from another. Since our gradient boosting models with original data do reasonably well compared to those in literature, we focused our efforts in tuning our gradient boosting tree models. Xgboost also has an option for scaling the number of observations to handle imbalance issue. So we stayed true to the data we derived from STRIDE. However, when testing on similar data with even lower incidence rate, AUROC decreased. So approaches to handle class imbalance well is worth exploring further. Overall, Xgboost performance is consistent and in terms of speed, it is superior compared to glmnet.

7 Future Work:

Another approach we would like to consider is the inverse probability of censoring weighted (IPCW) to penalize more for under-performance in predicting rare positive labels. In order to implement IPCW, we would need to constraint the length of time, for example, 10 years, for evaluating outcomes to avoid the exploding weight problem for some right censored observations. When evaluating our model performance on the subgroups, it is clear to us why the notion of fairness in machine learning has gained momentum in the last few years. (Pfohl, 2018) attempted to create a fair model predicting ASCVD risk using adversarial networks which showed to reduce variability across subgroups with sensitive attributes. When it comes to decision making, this will hopefully lessen the degree of mis-estimation for minority groups in order to avoid harm.

References

[1] Wessler, Benjamin S., et al. "Clinical prediction models for cardiovascular disease: tufts predictive analytics and comparative effectiveness clinical prediction model database." Circulation: Cardiovascular Quality and Outcomes 8.4 (2015): 368-375.

[2] https://www.cdc.gov/heartdisease/facts.htm

[3] Damen, Johanna AAG, et al. "Prediction models for cardiovascular disease risk in the general population: systematic review." bmj 353 (2016): i2416.

[4] Karmali, Kunal N., and Donald M. Lloyd-Jones. "Implementing cardiovascular risk prediction in clinical practice: the future is now." (2017): e006019.

[5] Muntner, Paul, et al. "Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations." Jama 311.14 (2014): 1406-1415.

[6] Pike, Mindy M., et al. "Improvement in cardiovascular risk prediction with electronic health records." Journal of cardiovascular translational research 9.3 (2016): 214-222.

[7] Wolfson, Julian, et al. "Use and customization of risk scores for predicting cardiovascular events using electronic health record data." Journal of the American Heart Association 6.4 (2017): e003670.

[8] Goldstein, Benjamin A., et al. "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review." Journal of the American Medical Informatics Association 24.1 (2017): 198-208.

[9] Zhao, Juan, et al. "Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction." Scientific reports 9.1 (2019): 717.

[10] Weng, Stephen F., et al. "Can machine-learning improve cardiovascular risk prediction using routine clinical data?." PloS one 12.4 (2017): e0174944.

[11] Rajkomar, Alvin, et al. "Scalable and accurate deep learning with electronic health records." NPJ Digital Medicine 1.1 (2018): 18.

[12] Lowe, Henry J., et al. "STRIDE–An integrated standards-based translational research informatics platform." AMIA Annual Symposium Proceedings. Vol. 2009. American Medical Informatics Association, 2009.

[13] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016.

[14] Pfohl, Stephen, et al. "Creating Fair Models of Atherosclerotic Cardiovascular Disease Risk." arXiv preprint arXiv:1809.04663 (2018).

8 Acknowledgements:

Special thanks to:

Stephen Pfohl and **Shah lab**: for mentorship throughout this project.
Gael Colas: for extensive help with everything during weekly office hours.
Anand Avanti: for providing extremely helpful feedback during project office hours.

Github repo: https://github.com/Minh084/cs229