Cardiovascular disease prediction: a novel risk-stratification tool

Stylianos Serghiou

Contact details: sstelios@stanford.edu

1. Introduction

Motivation

Cardiovascular disease (CVD) is the number one cause of death in the US. As such, prediction of CVD and implementation of measures to prevent or delay its occurrence are critical.[1] However, currently recommended risk stratification tools do not take into account significant predictors of CVD; as such, they are estimated to provide incorrect estimates to at least 31.6 million Americans per year.

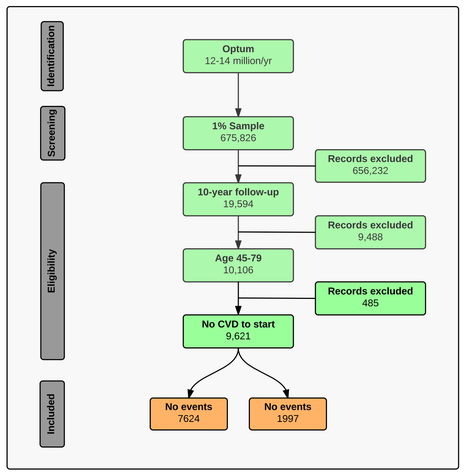
Our solution

Utilize methods of machine learning to develop a new risk stratification tool incorporating all important predictors of CVD.

2. Methods

Data acquisition

Data were provided by Optum, which maintains a longitudinal dataset of health plan data for 12-14 million annual lives across the US.

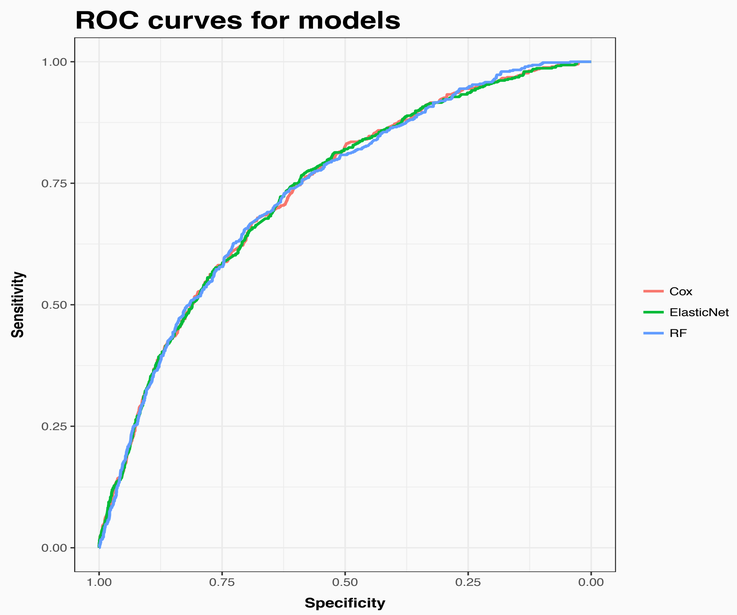


3. Results

Modelling

We used the following attributes: event of heart attack or stroke, time to event, age, sex, first digit of ZIP code, use of anti-diabetic medication, use of anti-hypertensive medication, use of cholesterol-lowering medication, use of anti-rheumatic medication and evidence of chronic kidney disease. All parameters were tuned by cross-validation. We used a 70-30 split for train-dev set (6734 patients with 1402 events in train set, 2887 with 595 events in dev set). All models were assessed using area under the curve (AUC). A Receiver Operating Characteristic (ROC) curve illustrates the performance of our best 3 models.

Model	Train AUC	Dev AUC
Currently used model (PCE)[2]	-	72.0%
Cox Proportional Hazards Regression	72.5%	72.4%
Elastic Net Regularized Cox Regression	73.8%	73.2%
Survival tree	70.0%	69.2%
Survival Forest (50 trees)	**80.1%**	**73.2%**
DeepSurv	(pending)	(pending)



4. Conclusions

Discussion

1. Advantages. Our models achieved a surprisingly good predictive ability for the amount of data we used. This was done in the context of a limited parameter space and without using traditionally highly-regarded predictors (e.g. smoking). The long-advocated geospatial data missing from all tools to date were the most powerful predictors after age.
2. Disadvantages. Our models' predictive ability would benefit by access to attributes such as tobacco smoking and obesity that are currently unavailable to us. Limitations in using socioeconomic data in combination with geospatial data due to identifiability concerns almost certainly restricted the potential performance of our models.

Future directions

1. *Sample expansion:* expand our sample to incorporate any individual with one year of follow up or more, any individual >30 yrs old
2. *Feature exploration:* create more features and utilize features designed by the Agency of Healthcare Research and Quality (AHRQ) to enhance the predictive ability of our models.
3. *Time-varying predictors:* use predictors the status of which does not remain at baseline, but rather changes over time.
4. *Web app:* deploy a web app so that individuals and doctors in the US can easily estimate risk of CVD.

References

1. Lloyd-Jones DM. Cardiovascular risk prediction: basic concepts, current status, and future directions. Circulation. 2010 Apr 20;121(15):1768-77. doi: 10.1161/CIRCULATIONAHA.109.849166.
2. Muntner et al. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. JAMA. 2014 Apr 9;311(14):1406-15. doi: 10.1001/jama.2014.2630.