Cardiovascular disease prediction: a novel risk-stratification tool

Abstract

Cardiovascular disease (CVD) accounts for 1 in 3 deaths worldwide. However, current "state-of-the-art" prediction tools annually misdiagnose 31.6 million Americans. We propose a novel risk stratification tool by applying methods of machine learning to health claims data. Our neural network outperformed the current state-of-the-art in terms of area under the curve (AUC) and illustrated that area of residence, which is currently neglected by alternative tools, is in fact one of the strongest predictors of CVD.

Introduction

Being the number one killers in the world, heart attack and stroke collectively account for an approximate 750,000 fatalities per year in the US alone.[1,2] Heart attack and stroke are manifestations of a broader class of morbidity, known as cardiovascular disease (CVD). The incidence of CVD to a large extent can either be completely prevented or delayed with appropriate dietary, lifestyle and drug intervention. However, such intervention is costly and at times harmful, for which reason risk stratification tools are essential.

Risk stratification tools attempt to predict occurrence of CVD in the next ten years in individuals with no signs of current CVD and treat those at a significantly high risk. The first such tool, the Framingham Risk Score, emerged in the early 1990's,[3] since which an additional 362 potential prediction models have been published.[4] The current "state-of-the-art" prediction tool used in the US is known as Pooled Cohort Equations (PCEs), which like most other tools, was developed by applying Cox proportional hazards regression to fit known CVD predictors to a longitudinal cohort of patients.[5-6] However, this tool does not take into account known significant predictors of CVD, including major CVD-promoting co-morbidities, such as rheumatoid arthritis and chronic kidney disease (CKD). As a result, an estimated 31.6 million Americans receive an incorrect CVD risk prediction and consequently do not receive appropriate preventive treatment under the current treatment guidelines.

Our aim is to mitigate the number of Americans being on the wrong treatment by developing a new CVD risk prediction tool to address this critical and unmet need. We plan to do so, by (1) systematically applying methods of machine learning to developing a new risk stratification algorithm for CVD, (2) incorporating in our algorithm features we believe are important that had not been included in the PCEs (such as, area of residence and rheumatoid arthritis) and (3) using a large database with much broader coverage than that originally used to develop the PCEs. We then plan to compare our algorithms' performance against the currently recommended risk stratification system.

Methods

Study design

We examined commercial health plan data and historic claims for Medicare Advantage members of a large national managed healthcare company affiliated with Optum. Optum maintains longitudinal data for 12 to 14 million annual lives across all 50 states between January 1, 2003 and December 31, 2016. Access to these data was approved and provided by the Stanford Center for Population Health Sciences (PHS). In reporting our work, we will be using the guidelines issued by the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) statement.

Study population

Members of Optum were eligible for our study if they were 45-79 years old upon enrollment into Optum, excluding those with no continuous enrollment for at least five years. All records were obtained from Optum and included 549 variables related to demographic characteristics, socioeconomic status, area of residence (ZIP-coded), medical claims (majority), medication, laboratory results, hospitalization and details of healthcare provider. No validations of the data held by Optum have been done to ascertain misclassification of exposure or outcome. Our analyses were done on a 1% random sample of the Optum dataset for relative computational convenience. Because of identifiability concerns, we were not allowed to combine data regarding socioeconomic status and area of residence, for which reason we did not use any such data.

Outcomes

Our primary end-point was identical to that of PCEs: nonfatal myocardial infarction (i.e. heart attack) or coronary heart disease death, or fatal or nonfatal stroke, over a 10-year period among people free from CVD at the beginning of the period. This was defined by the International Classification of Diseases 10th revision, Clinical Modification (ICD-10-CM) codes I21.xx (acute myocardial infraction (MI), i.e. heart attack) and I60.xx, I61.xx, I63.xx, I65.xx, I66.xx or I67.89 (cerebrovascular disease, e.g. stroke). Wherever these were not available, we used ICD-9-CM codes "410.xx (MI) and stroke events were defined by an ICD-9 diagnosis code of 430.xx, 431.xx, 433.xx, 434.xx, or 436.x. These definitions have positive predictive values of greater than 90%.[6]

Statistical analysis

First, we divided our sample data into a training set (80%), a dev set (10%) and a test set (10%). Then, we developed the following three sets of features at baseline (i.e. upon enrollment into Optum):

Set 1: Original features. These are the features used in fitting the PCEs, apart from smoking status, which Optum does not collect: gender, age, race, total cholesterol, HDL (High-Density Lipoprotein) cholesterol, administration of anti-hypertensive medication and presence of diabetes.

Set 2: Theoretical features. These are the features originally used in fitting the PCEs (apart from ethnicity, which cannot be used in conjunction with area of residence data), as well as features with theoretically significant predictive ability for CVD: first-digit ZIP code, use of anti-diabetic medication, use of anti-rheumatoid medication, use of lipid lowering medication and MDRD-predicted glomerular filtration rate (a marker of CKD based on serum creatinine). All of these features, apart from first-digit ZIP code, were engineered using context-based knowledge. Other important features, such as obesity, were not available on Optum.

Set 3: Augmented features. This included two separate subsets of features. The first subset consisted of all aforementioned demographic and medical history features, as well as any of 428 classes of medication identified by Optum. The second set of features was identical to the first, but classes of medication were analyzed using Principal Component Analysis (PCA), from which we extracted the first 50 principal components.

Hyper-parameter tuning was done using cross-validation. Models were assessed against each other and the PCEs using the following two methods: (1) assessment of the proportional-hazards assumption using Shoenfeld residuals (violations of this assumption affect accuracy of risk estimates among subgroups); (2) assessment of discriminating ability using the cross-validated area under the Receiver Operating Characteristic (ROC) curve (AUC) in the train and dev sets.[7] Our best-performing model was then evaluated on the test set.

All analyses were done in R (R Foundation, version 3.4.2) and Python (Python Software Foundation, version 3.6) on PHS Windows servers. All code will be available at https://github.com/serghiou.

Results

Patients

Of 675,826 unique patients in our 1% sample of Optum, we identified 104,105 patients with at least five years of continuous follow up. Of these, only 46,263 were between 45-79 years of age; these included 21,448 male patients (46%) of mean age 59.6 years (standard deviation (SD), 10.0 years) and 24,815 female patients (54%) of mean age 60.7 years (SD, 10.2). Most of these patients resided in the West Cost (8,383) with only 1,733 patients in the Northeast.

Of eligible patients, 7,169/46,263 (16.5%) had an event we classified as CVD. Most patients with an event were male (49% vs 45%) and at baseline tended to be older (66 vs 59 years old) and suffer from more comorbidities, such as high blood pressure (64 vs 46%) and diabetes mellitus (8 vs 4%). Among patients with no event, the median follow up time was 2556 days (interquartile range (IQR), 2101-3287 days) and among patients with no event, the median follow up time was 2770 days (IQR, 2191-3712 days).

Modelling

We modelled our data using ten different models (Table 1). All models performed similarly well in terms of AUC and in all models train AUC was very similar to that in dev set. Performance in classification models was approximately identical to that of survival models, even though Cox regression violated the proportional hazards assumption (P-value < 0.001). Radial SVM significantly underperformed, even though this was in the context of parameter tuning within a smaller randomly selected cohort of 2,000 individuals due to computational limitations. The best performing model was the neural network, even though its mean performance was only marginally higher than that of survival forest. In its final form, this neural network included one hidden layer with 50 units, each of which units used a rectified linear unit (ReLU) activation function with a sigmoid activation function in the output layer; this was trained for 100 epochs. Given that this neural network was our selected model, this was also evaluated in our test set, in which it performed less well than the dev set, but similarly well (AUC, 0.744; 95% CI, 0.726-0.763). We then attempted to model the predictors of set 3 using a neural network, which was the best performing approach in step 2. The neural network for all medications had two hidden layers of 64 units each and the neural network for the 50 principal components had one hidden layer with 15 units (no improvement in performance with alternative configurations).

Method	Train AUC	95% CI	Dev AUC	95% CI
PCE	-	-	0.72	0.70-0.75
		Classification		
Logistic	0.740	0.734-0.747	0.750	0.731-0.770
Logistic Elastic	0.740	0.734-0.747	0.750	0.731-0.770
SVM radial	0.634	0.628-0.644	0.621	0.596-0.646
RF	0.766	0.760-0.772	0.751	0.733-0.770
NN	0.745	0.739-0.752	**0.755**	**0.737-0.774**
NN (all)	0.653	0.645-0.661	0.640	0.616-0.664
NN (PCA)	0.743	0.736-0.750	0.734	0.713-0.755
		Survival		
CPH	0.740	0.733-0.747	0.750	0.731-0.770
CPH Elastic	0.740	0.733-0.747	0.750	0.731-0.770
Survival forest	0.763	0.756-0.769	0.754	0.735-0.773

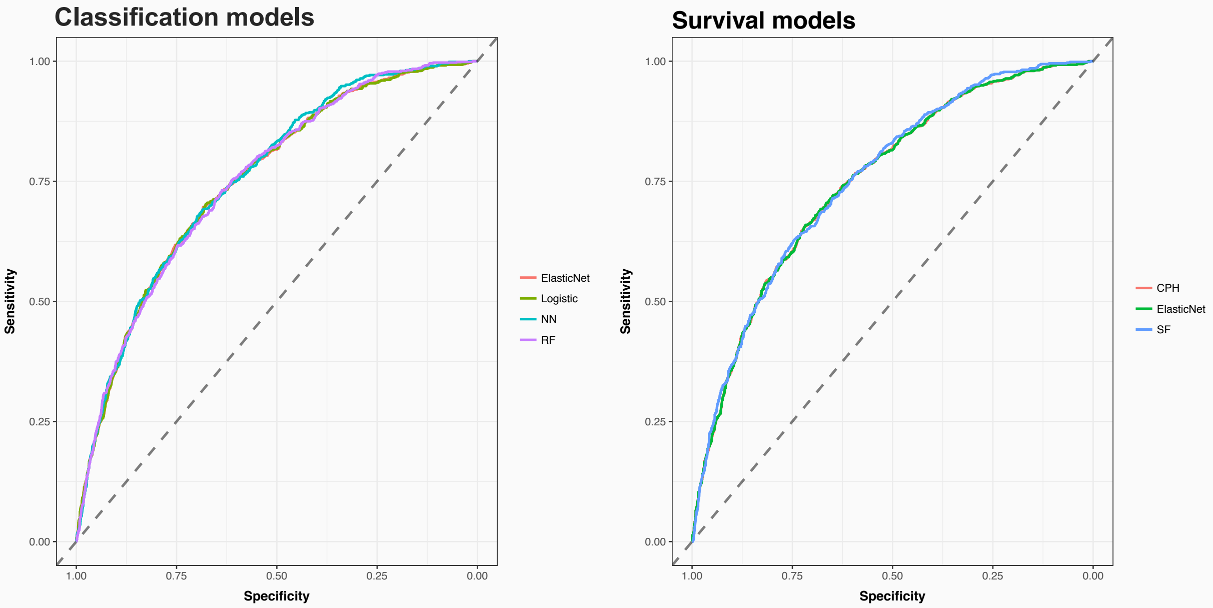

Figure 1. Receiver Operating Characteristic (ROC) curves. ROC curves for the survival models were calculated by considering the binary outcome CVD within a maximum of ten years within enrollment versus not.

In attempting to understand the performance of our models, we also quantified relative predictor importance, by quantifying the relative weight given to the scaled predictors. Cox regression allocated its top three weights to diabetic medication, first ZIP digit = 9 and decade of age. This allocation was very similar to that of the random and survival forests, where, by cumulative decrease in the Gini impurity criterion, the three most important predictors were age, first-digit of ZIP code and diabetic medication. Lastly, in the best-performing neural network, estimating predictor importance from weight parameters,[7] the three most important predictors were, first ZIP digit = 4, anti-diabetic medication and male gender.

Discussion

Summary

In our attempt to predict 10-year CVD using health claims data, we were able to devise models capable of performing at a level significantly higher than that of the established model in terms of AUC in our test set. Interestingly, we were able to do so without access to known important predictors, such as tobacco smoking and obesity, which makes our model more broadly applicable. Agnostic models incorporating all possible medication performed well, but not as well as models on predictors based on prior knowledge. Lastly, even though area of residence is not a component of any CVD risk stratification tool in the US, in our sample this was consistently one of the most important predictors.

Benefits and limitations

Our analysis was not without limitations. First, it is almost certain that we would have achieved a better fit with access to known important predictors, such as tobacco smoking and obesity. Nevertheless, despite their absence, our model was able to outperform models using these missing predictors. Second, there was an astonishing amount of missingness in our database. We attempted to mitigate this problem by diagnosing disease, such as CKD, on the basis of medication taken and lab results, but many other diagnoses could not be identified reliably. Third, we were unable to run a few of the analyses we had planned (e.g. calibrated radial SVM, lasso with all medications, etc.) because of computational limitations. We tried to circumvent this issue by working with a smaller sample size, vectorization, parallelization to multiple cores and using packages suited to larger data (e.g. data.table for data management and keras for fitting neural networks in R), but this was not successful for some of our analyses.

Future work

A number of significant steps remain in producing the best possible risk stratification tool. First, we need to scale our code to include the whole Optum dataset; we have already started working on this by migrating parts of our data manipulation code to Apache Spark and of our data analysis code to the Google Cloud Platform. Second, we need to consider time-varying covariates; these are potentially important indicators of disease, which may improve our predictive ability beyond that of baseline characteristics. Third, we are interested in testing the performance of a recently developed neural network for survival analysis, called DeepSurv.[8] This was attempted, but further work is needed.

Acknowledgements

This work was done under supervision by Prof. Sanjay Basu, MD, PhD.

References

1. Centers for Disease Control and Prevention (CDC). Heart Disease Facts. 2017 [cited 13 Dec 2017]. Available from: https://www.cdc.gov/heartdisease/facts.htm
2. Centers for Disease Control and Prevention (CDC). Stroke Facts. 2017 [cited 13 Dec 2017]. Available from: https://www.cdc.gov/stroke/facts.htm
3. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. Am Heart J1991;121:293-8. doi:10.1016/0002-8703(91)90861-B.
4. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, Lassale CM, Siontis GC, Chiocchia V, Roberts C, Schlüssel MM, Gerry S, Black JA, Heus P, van der Schouw YT, Peelen LM, Moons KG. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 2016; 353: i2416. doi: 10.1136/bmj.i2416
5. Muntner P, Colantonio LD, Cushman M, Goff DC Jr, Howard G, Howard VJ, Kissela B, Levitan EB, Lloyd-Jones DM, Safford MM. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. JAMA. 2014 Apr 9;311(14):1406-15. doi: 10.1001/jama.2014.2630.
6. Goff DC Jr, 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 Jr, Sorlie P, Stone NJ, Wilson PW, Jordan HS, Nevo L, Wnek J, Anderson JL, Halperin JL, Albert NM, Bozkurt B, Brindis RG, Curtis LH, DeMets D, Hochman JS, Kovacs RJ, Ohman EM, Pressler SJ, Sellke FW, Shen WK, Smith SC Jr, Tomaselli GF; American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2014 Jun 24;129 (25 Suppl 2):S49-73. doi: 10.1161/01.cir.0000437741.48606.98.
7. Garson DG. Interpreting Neural-network Connection Weights. AI Expert. 1991 April; 6(4):46-51
8. Katzman J, Shaham U, Bates J, Cloninger A, Jiang T, Kluger Y. DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network. arXiv. 2017 Aug 17 (3rd Edition).