CS229 Project: Building a Better Risk Prediction Model for ASCVD

Aditya Kanukurthy (sang33t1)
Jonas Kemp (jbkemp7)

Introduction

Risk prediction for cardiovascular disease continues to be an active area of research. In 2013, the American Heart Association (AHA) and the American College of Cardiology (ACC) released a new set of guidelines for statin therapy based on risk estimates from the atherosclerotic cardiovascular disease pooled cohort equations (ASCVD PCEs).[1] Risk calculators based on these equations are freely available online.[2] The ASCVD PCEs were fit to data from five separate cohort studies, improving precision and generalizability over previous models.

However, recent validation studies of these equations have suggested nontrivial overestimation of risk in new multiethnic cohorts, particularly among minority groups.[3,4] Given the increasingly widespread use of the model in guiding treatment decisions, overestimates of risk could lead to significant overtreatment for heart conditions, resulting in unnecessary costs or harms from side effects. Here, we extend ongoing work by Dr. Sanjay Basu (School of Medicine, Prevention Research Center) that seeks to improve on the existing PCEs through statistical techniques to reduce overfitting and account for cohort effects. We attempt to develop an alternative risk model using machine learning methods, adopting a classification approach to predicting the occurrence of an ASCVD event in a 10-year period based on demographic and biometric patient data. Although our results did not show clear improvement over the baseline model, we explore some of the challenges of modeling these data and propose potential directions for additional research.

Methods

Data

Our dataset included patient data from eight cohort studies. Five of these cohorts were used to fit the original ASCVD equations: (i) the Atherosclerosis Risk in Communities Study (ARIC); (ii) the Cardiovascular Health Study (CHS); (iii) the Coronary Artery Risk Development in Young Adults Study (CARDIA); (iv) the Framingham Heart Study; and (v) the Framingham Offspring Study. These cohorts were dominated by white patients, but the three additional cohorts provided a more ethnically diverse sample: (i) the Women's Health Initiative Observational Study (WHI-OS), (ii) the Multi-Ethnic Study of Atherosclerosis (MESA), and (iii) the Jackson Heart Study (JHS). Using the same inclusion criteria as for development of the original equations, patients were included if they were 40-79 years old, white or black, and free from previous heart problems. Outcomes were defined as an ASCVD event (heart attack, stroke, or death from coronary heart disease) within the first 10 years of any study. Covariates collected consistently across all eight cohorts included age, gender, race, BMI, cholesterol levels, blood pressure, hypertension treatment status, smoking status, and diabetes status. Features considered for each model included individual covariates (or their logarithm, if continuously valued), a quadratic term for age, interactions with age, and interactions between blood pressure and hypertension treatment. The full set of pooled cohorts included 29,096 total patients.

Analysis of Original Five-Cohort Dataset

The original ASCVD PCEs consist of four separate Cox proportional hazards models, one fitted for each combination of race (white or black) and gender (male or female). Although Cox models estimate only relative risk (i.e. a log hazard ratio, as a linear function of the predictors), absolute risk estimates can be constructed using information on baseline survival rates, per the following equation[5]:

$$\hat{p}_i = 1 - S_0(t)^{\exp(\beta^T(X_i - \bar{X}))} \quad (1)$$

$S_0(t)$ is the baseline survival rate to time t, β is the vector of parameters estimated by the model, X_i is the vector of predictor values for patient i, and $\bar{X}$ is the vector of mean predictor values in the population. In clinical practice, these Cox risk predictions can be transformed into decision rules for care. For example, the AHA/ACC guidelines recommend initiation of statin therapy in individuals with an estimated 10-year risk over 7.5%. In our analysis, we use baseline survival rates provided by Dr. Basu, derived from CDC data.

We first reconstructed the black women's PCE on the original five cohorts and performed error analysis, evaluating sensitivity, specificity, and error using the WHI cohort as the test set. Patients were predicted to have an event if they fell in the high-risk group, corresponding to the 7.5% threshold specified by clinical guidelines on the above 10-year risk score. Evidence from this analysis corroborated previous claims of minority risk overestimation, with extremely poor generalization error driven by a high false positive rate (Fig. 1). Dr. Basu proposed two initial hypotheses for this phenomenon: 1) overfitting, due to the small proportion of black patients in the original five-cohort sample, or 2) cohort effects not accounted for in the basic model. To test these hypotheses, we fit both L_1-regularized and mixed-effects Cox models (with random slope and intercept by cohort) and evaluated performance using 10-fold cross-validation on the original five cohorts. Regularization parameters were selected using a nested cross-validation procedure. We computed our analyses in R[6], using the packages coxme[7] for mixed-effects Cox models, coxnet (from glmnet)[8] for regularized Cox models, caret[9] for cross-validation, and AUC[10] for performance metrics.

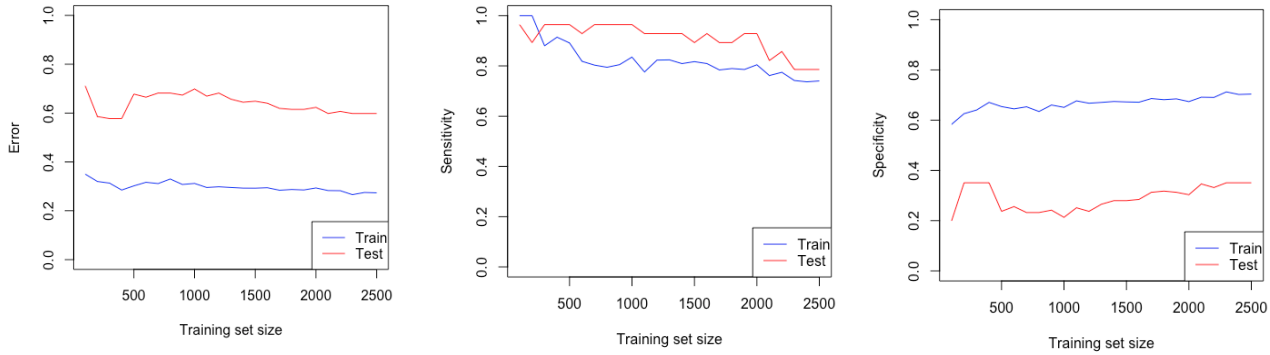

Figure 1. Train vs. test error (left), sensitivity (middle), and specificity (right) for the reconstructed black women's PCE.

Analysis of Full Eight-Cohort Dataset

Once we gained access to all eight cohort datasets, including MESA and JHS, we combined all cohorts into a single pooled dataset and split into train and test sets (75% – 25%). Error analysis on Cox models refit to

these data no longer showed evidence of significant minority risk overestimation, but still suggested a substantial bias problem (Fig. 2):

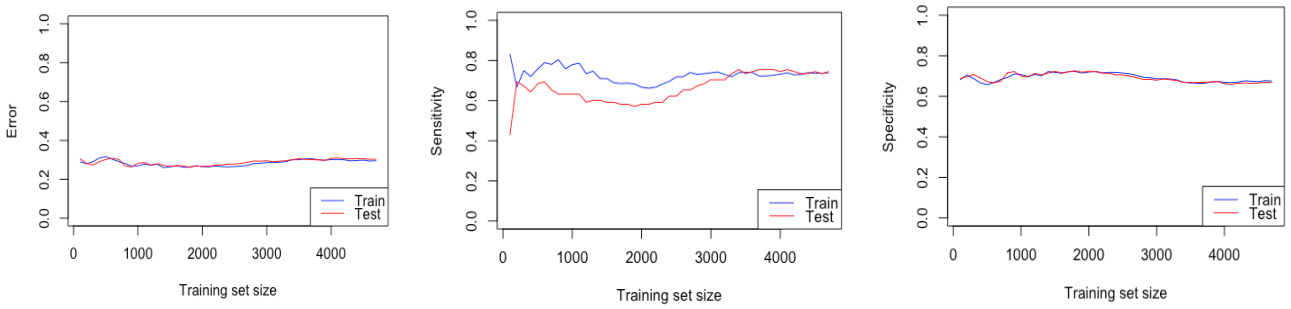

Figure 2. Train vs. test error (left), sensitivity (middle), and specificity (right) for a Cox model fit to a training set of black women from all eight cohorts.

We attempted to fit three higher-variance models on the full training set to mitigate this problem: SVMs with polynomial and RBF kernels, and random survival forests. Although SVMs have significant limitations in this context, as they neither handle the survival (time-to-event) aspect of the data nor naturally produce interpretable risk scores, kernelized implementations are readily available and we felt they might therefore offer some information about whether a higher-dimensional feature space could improve discrimination.

Random survival forests extend Breiman's random forest method to handle right-censored survival data. In this algorithm, decision trees are grown on bootstrap samples of the training data, under the condition that each terminal node must contain at least $d_0 > 0$ unique events. These trees can be used to estimate a cumulative hazard function using the Nelson-Aalen estimator, averaged over all trees to reduce overfitting. (See Ishwaran et al. for full mathematical details.[11]) Cumulative hazard and survival are related by the equation

$$S(t) = \exp(-H(t)) \tag{2}$$

Therefore, we computed the random forest risk score for each patient at time t as $1 - S(t)$.

We also experimented with gradient- and likelihood-based boosting methods for fitting Cox models. The former method performs coordinate descent on the negative partial log-likelihood, where each update (the "weak estimator") is penalized; the latter performs coordinate ascent on the L_2-penalized partial log-likelihood modified with a specific offset term. (See De Bin for a detailed comparison of the two methods, which in the case of Cox regression are quite similar but not generally equivalent without a particular choice of penalty parameters.[12]) These methods increase bias when few boosting steps are used, but approach the maximum partial likelihood estimate of the standard Cox model as the number of steps goes to infinity. We fit both models to the full dataset in order to compare results across a wider spectrum of bias/variance tradeoffs.

Hyperparameter for the SVMs (regularization terms, RBF kernel bandwidth, polynomial kernel degree) and for the boosted models (number of boosting steps, penalty terms) were selected using 10-fold cross-validation. Unfortunately, because the random survival forest was both expensive to compute and had

several hyperparameters (including number of candidate variables for splitting, number of trees, maximum tree depth, minimum number of events per node, etc.), we fit using only the default values in Ishwaran's implementation. We used the packages caret to fit SVMs, randomForestSRC[13] to fit random survival forests, and mboost[14] and CoxBoost[15] for gradient- and likelihood-boosted Cox models, respectively.

Results

We compared models on sensitivity and specificity at the 7.5% clinical threshold, as well as on AUC. Note that we do not report accuracy, as we do not find it a useful comparison metric here: due to the high class imbalance in our data, high accuracy can trivially be achieved by always predicting the negative class.

Five-Cohort Analysis

Across all four race and gender combinations, differences between the baseline Cox model and mixed-effects models were negligible. The regularized model for black men was also negligibly different from baseline; however, the other regularized models decreased sensitivity by 0.02-0.05 and increased specificity by 0.01-0.04. AUC for these regularized models decreased by up to 0.01.

Eight-Cohort Analysis

To simplify the full eight-cohort analysis, we trained just two versions of each model (one each for men and women) and included race in the models as a predictor. The results of this analysis are presented in table 1.

	Men						***Women***					
	Train			*Test*			*Train*			*Test*		
	Sens.	Spec.	AUC	Sens.	Spec.	AUC	Sens.	Spec.	AUC	Sens.	Spec.	AUC
Cox (Baseline)	0.89	0.43	0.74	0.88	0.42	0.72	0.78	0.65	0.79	0.71	0.66	0.72
Random Forest	1	0.68	1	0.85	0.41	0.70	1	0.84	1	0.65	0.69	0.72
RBF SVM	1	0	0.99	0.99	0	0.60	1	1	1	0.66	0.59	0.63
Polynomial SVM	1	0	0.58	1	0	0.58	0.69	0.58	0.69	0.71	0.58	0.70
Gradient-Boosting	0.86	0.39	0.70	0.86	0.38	0.68	0.49	0.85	0.76	0.42	0.84	0.69
Likelihood-Boosting	0.86	0.44	0.73	0.87	0.43	0.72	0.70	0.71	0.78	0.65	0.70	0.73

Table 1. Comparison of five algorithms to baseline Cox model on sensitivity and specificity at 7.5% risk threshold and on AUC. Train-test splits and fitting were performed separately for men and for women.

Discussion and Future Directions

One of the most important results we present here is also one of the simplest: we find that overestimation of minority risk can be reduced merely by training on additional data from modern, ethnically diverse cohorts. We failed to improve performance by statistically accounting for cohort effects, and found that regularization reduced the false positive rate at the clinically relevant threshold but at the cost of additional false negatives, with no net improvement in discrimination (as measured by the AUC). Though we operated under an event classification paradigm, further research is needed to understand whether such trade-offs correspond to clinically appropriate changes in the actual risk score estimates. However, our error analysis results from merging all eight cohorts clearly demonstrate that generalization error essentially vanishes given an expanded minority sample. Perhaps unsurprisingly, sophisticated modeling techniques are no substitute for better data.

None of our models trained on the merged cohort data offered a clear improvement over the basic Cox model. Notably, the RBF-kernelized SVM severely overfit the training data despite regularization, while the polynomial-kernelized SVM showed little generalization error but high bias, particularly for men. Part of this poor performance may simply result from the fact that SVMs do not explicitly handle the survival aspect of the data; though we could not implement one due to time constraints, a kernelized Cox model might be more appropriate for the problem. However, the results may also imply that additional features beyond what is readily available in the data could be necessary to improve discrimination. Of course, conducting a large-scale cohort studies to collect such additional survival data would likely be prohibitively expensive, so perhaps other clever feature engineering techniques beyond kernel methods should be explored.

Our random survival forest also heavily overfit the training data, though performance on the test set was still roughly on par with the baseline model. Our exploration of this model was unfortunately limited, and further examination of different hyperparameter settings would certainly be warranted – other configurations might better translate the model's power into generalizable gains. The likelihood-boosted Cox model, meanwhile, matched or even slightly outperformed the baseline model on AUC. In fact, like the regularized models in the initial analysis, the boosted models for women seem to favor specificity over sensitivity (perhaps too dramatically, in the gradient-boosted case). Again, whether and to what extent such trade-offs are suitable for practice is a matter for clinical expertise.

Throughout all our approaches, this dataset proved remarkably challenging to model well. For example, despite only moderate gender imbalance (57% women to 43% men) and differences in event rates (8% for women, 12% for men) in the merged dataset, the false positive rate for men remained stubbornly high across all models. Another possible direction for future research might include unsupervised analysis to try to understand whether any hidden structure in the data could be contributing to these difficulties, or could be exploited to improve on future models.

Acknowledgements

Thanks to Dr. Basu for providing access to the data, an overview of his preliminary analysis, and ongoing guidance throughout our project.

References

[1] Stone N.J., et al. (2013). 2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults. *Circulation*, *134*(21). doi:10.1161/01.cir.0000437738.63853.7a

[2] ASCVD Risk Estimator. http://tools.acc.org/ASCVD-Risk-Estimator/

[3] DeFilippis A., et al. (2015). An Analysis of Calibration and Discrimination Among Multiple Cardiovascular Risk Scores in a Modern Multiethnic Cohort. *Annals of Internal Medicine, 162*(4): 266-275. doi:10.7326/M14-1281

[4] DeFilippis A., et al. (2016). Risk score overestimation: the impact of individual cardiovascular risk factors and preventive therapies on the performance of the American Heart Association-American College of Cardiology-Atherosclerotic Cardiovascular Disease risk score in a modern multi-ethnic cohort. *European Heart Journal*. doi:10.1093/eurheartj/ehw301

[5] D'Agostino R., et al. (2008). General cardiovascular risk profile for use in primary care. *Circulation, 117:* 743-53. doi:10.1161/CIRCULATIONAHA.107.699579

[6] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[7] Therneau T. (2015). coxme: Mixed Effects Cox Models. R package version 2.2-5. https://CRAN.R-project.org/package=coxme

[8] Simon N., Friedman J., Hastie T., Tibshirani R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software, 39*(5): 1-13. http://www.jstatsoft.org/v39/i05/

[9] Kuhn M., et al. (2016). caret: Classification and Regression Training. R package version 6.0-73. https://CRAN.R-project.org/package=caret

[10] Ballings M., Van den Poel D. (2013). AUC: Threshold-independent performance measures for probabilistic classifiers. R package version 0.3.0. https://CRAN.R-project.org/package=AUC

[11] Ishwaran H., et al. (2008). Random survival forests. *Annals of Applied Statistics 2*(3): 841-60. doi:10.1214/08-AOAS169

[12] De Bin R. (2016). Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost. *Computational Statistics 31*(2): 513-31. doi:10.1007/s00180-015-0642-2

[13] Ishwaran H., Kogalur U.B. (2016). Random Forests for Survival, Regression and Classification (RF-SRC), R package version 2.4.1.

[14] Hothorn T., et al. (2016). mboost: Model-Based Boosting, R package version 2.7-0, https://CRAN.R-project.org/package=mboost.

[15] Binder H. (2013). CoxBoost: Cox models by likelihood-based boosting for a single survival endpoint or competing risks. R package version 1.4. https://CRAN.R-project.org/package=CoxBoost.