Predictive Models for Cardiovascular Health Based on Social Determinants of Health

Noreen Mayat[1]

***Abstract*— How may we improve existing predictive metrics for cardiovascular health outcomes using machine learning?**

I. Introduction

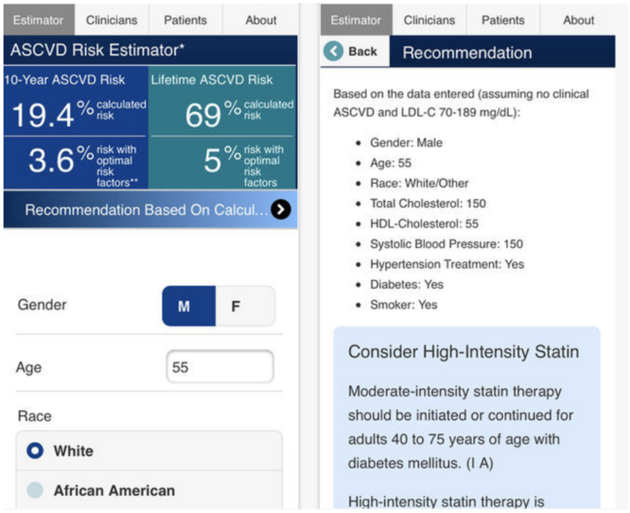
One of the questions people have been grappling with is whether or not to incorporate race (or other genetic markers), or social determinants of health (SDOH) correlated with race to predict cardiovascular health outcomes. Sometimes, incorporating race into these predictive equations for different health outcomes has adverse effects; for example, existing risk scores for kidney disease, such as eGFR, may disproportionately disadvantage Black communities, as Black patients with similar creatinine levels to white counterparts can be scored as having "healthier" kidneys, which may lead to under-treatment by health professionals [1]. Such issues point to why using social determinants of health alongside race in these predictive models may allow for a more nuanced understanding of an individual's health risks by considering the broader picture of their environment, socioeconomic status, BMI, mental health, and access to care, which may all also be critical factors in determining an individual's health risks in addition to their race.

Our research goal is to build a machine learning model incorporating other variables, such as social determinants of health, not currently being leveraged in computing existing risk scores for cardiovascular health outcomes, and to analyze how such a machine learning model compares to existing risk scores in predicting cardiovascular health outcomes.

II. Background and Related Work

A. ASCVD Risk Scores

There are various existing risk scores used for predicting cardiovascular health outcomes: the first is a risk score known as the Framingham score, developed by the Framingham Heart Study [2]. Other risk scores for predicting cardiovascular health outcomes include the Systematic Coronary Risk Evaluation (SCORE) algorithm in Europe, the QRISK3 in England and Wales, and the risk score for atherosclerotic cardiovascular disease (ASCVD), developed by the American College of Cardiology/American Heart Association using pooled cohort equations (PCE): the ACC/AHA 2013 pooled cohort risk equation [2]. This is typically the ASCVD risk score referenced in most studies.



ASCVD Risk Estimator*

10-Year ASCVD Risk
19.4% calculated risk
3.6% risk with optimal risk factors**

Lifetime ASCVD Risk
69% calculated risk
5% risk with optimal risk factors

Gender ● M ○ F

Age 55

Race
● White
○ African American

Recommendation

Based on the data entered (assuming no clinical ASCVD and LDL-C 70-189 mg/dL):

- Gender: Male
- Age: 55
- Race: White/Other
- Total Cholesterol: 150
- HDL-Cholesterol: 55
- Systolic Blood Pressure: 150
- Hypertension Treatment: Yes
- Diabetes: Yes
- Smoker: Yes

Consider High-Intensity Statin

Moderate-intensity statin therapy should be initiated or continued for adults 40 to 75 years of age with diabetes mellitus. (I A)

High-intensity statin therapy is

Fig. 1. Score demonstrating the variables used in the current ASCVD Risk Estimator Calculator using the ACC/AHA PCE to compute ASCVD risk scores for a 55 year-old white male [5].

PCEs leverage a combination of cohort studies in public health where they recruited patients from various demographics and followed their cardiovascular health for varying amounts of time, and "pool" the studies together to increase the diversity of the sample used to develop the metric. The current health variables leveraged to estimate ASCVD risk using the ACC/AHA PCE includes age, sex, race, total cholesterol, HDL cholesterol, systolic blood pressure, blood pressure lowering medication use, diabetes status, and smoking status [3]. According to a 2016 study, researchers found that in a large, multi-ethnic population, the ACC/AHA Pooled Cohort Risk Equation for ASCVD substantially overestimated actual 5-year risk in adults without diabetes, overall and across sociodemographic subgroups [4].

B. Social Determinants of Health

The World Health Organization defines social determinants of health to be the non-medical factors that influence health outcomes; they are the "conditions in which people are born, grow, work, live, and age, and the wider set of forces and systems shaping the conditions of daily life" [6]. Past research has demonstrated that integrating individual-level social determinants of health into electronic health records can assist in overall risk assessment models and in predicting holistic healthcare utilization and health outcomes, motivating efforts to collect and standardize patient-level social determinants of health information to benefit all types

[1] Noreen Mayat is a senior majoring in data science at Barnard College, Columbia University New York, NY. nm3224@barnard.edu

of risk scores, not just cardiovascular risk scores [7].

C. *Impacts of Social Determinants of Health*

Researchers have been exploring if incorporating social determinants of health into existing risk score equations for cardiovascular health outcomes improves the accuracy of these risk score predictions. A 2020 systematic review analyzing articles reporting on the use of machine learning models for cardiovascular disease prediction, which incorporated social determinants of health, found that most studies that compared performance with or without social determinants of health showed increased performance with them [8]. The most commonly included social determinants of health variables in these studies were gender, race/ethnicity, marital status, occupation, and income. The researchers note that there were a limited variety of sources and data in the reviewed studies, and thus, there is not as much research on how other social determinants of health variables, such as environmental ones, are known to impact cardiovascular disease risk, would impact model performance. Recording such data in electronic databases, as previous studies have also recommended, would enable their use. Further, a 2022 study found that adding social determinant of health risk factors alongside the existing variables used in ACC/AHA PCE to train a model in fact improved ASCVD risk prediction in specifically an African American cohort, a historically disadvantaged group in the healthcare system [9]. In this study, social determinants of health such as BMI, depression, weekly stress, insurance status, family income, and neighborhood violence were determined as the most important for prediction in this demographic, and were independently associated with 10-year ASCVD risk [9]. Other studies have found similar results in ASCVD risk prediction models leveraging social determinants of health such as education, income, and employment in addition to the existing variables used in PCEs for ASCVD risk prediction in both Black populations and non-Black female populations [10].

D. *Leveraging Indexes: Social Disadvantage Score*

Other studies have investigated the effects of social determinants of health on ASCVD risk scores by establishing a baseline Social Disadvantage Score (SDS) and examining its relationship with atherosclerotic cardiovascular disease (ASCVD) and overall mortality, as well as its influence on the prediction of ASCVD risk scores. The SDS ranged from 0 to 4, and was calculated by tallying the following social factors: (1) household income less than the federal poverty level; (2) educational attainment less than a high school diploma; (3) single-living status; and (4) experience of lifetime discrimination. However, this study found that Although SDS is independently associated with incident ASCVD and all-cause mortality, it does not improve 10-year ASCVD risk prediction beyond pooled cohort equations [11]. This may mean some social determinants of health may improve ASCVD risk score prediction, but not all, so we must be careful which ones to include in future models based on this existing research.

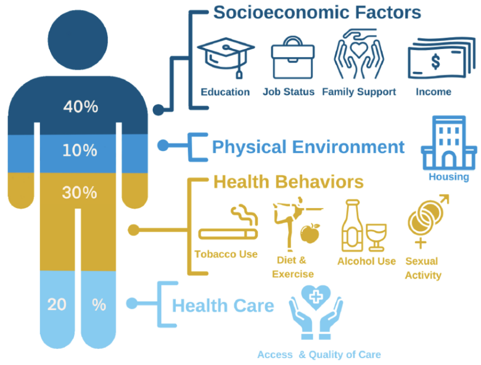


Fig. 2. Visualization of various social determinants of health and how they can contribute to overall health outcomes, developed by UCLA Health [12].

We must note that previous studies have found that removing race from machine learning models predicting ASCVD risk scores did not improve model performance in any subgroup, while various studies have found including race alongside other social determinants of health have improved model performance [13]. This information points to the idea that race is definitely still a significant variable in computing ASCVD risk, and should continue to be included in any future models predicting ASCVD risk, alongside other social determinants of health.

E. *NHANES*

The National Health and Nutrition Examination Survey (NHANES) is a downloadable public use data set used to document health care utilization, health status of various age groups, and related personal and lifestyle characteristics [14]. The data files are prepared and disseminated through the Centers for Disease Control and Prevention (CDC) to provide full access to data. More specifically, NHANES is a population-based survey designed to collect information on the health and nutrition of the US household population.

F. *Machine Learning Models*

Predictive machine learning models analyze datasets to predict a specific target variable: such datasets are composed of multiple data points, or samples, where each data point represents an entity we want to analyze [15]. Each of these entities has a list of various features associated with it: these features can be categorical (predefined values of no particular order like male and female), ordinal (predefined values that have an intrinsic order to them like a disease stage), or numerical (e.g., real values), and they are used to train the model and predict the target variable [15].

A model would analyze these feature variables and learn which variables are generally correlated/significant for predicting ASCVD using a training set, which is usually 75 percent of the data, and then makes predictions on a test set it hasn't seen before, which is usually 25 percent of the data. Different metrics are then used to evaluate model performance and accuracy, such as root mean square error, mean absolute percentage error, and r-squared value [15].

Popular machine learning models include KNN, Random Forest, and Decision Trees [16]. Past research in Taiwan has used machine learning models with appropriate transfer learning as a tool for the development of cardiovascular risk prediction (ASCVD) models for Asian populations [17]. Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned [18].

III. Design / Implementation / Algorithm

A. High-Level Overview

In this project, I am leveraging the data in NHANES to incorporate different social determinant of health variables into existing risk score metrics, such as the ACC/AHA 2013 pooled cohort risk equation, to predict cardiovascular health outcomes. I want to pull all relevant data to social determinants health and data used in existing ASCVD risk score calculators, such as general health and demographic data, as well as data surrounding cardiovascular health outcomes.

For our research purposes, our target variable is negative cardiovascular health outcomes: these outcomes include congenitive heart failure, stroke, heart attack, and coronary heart disease. Our feature variables used to train our model are various numerical and categorical values related to demographic data (age, gender, sex), general health data used to compute current ASCVD scores (total cholesterol, HDL cholesterol, systolic blood pressure, blood pressure lowering medication use, diabetes status, and smoking status), and data on various social determinants of health, such as: a more stratified race variable with more granualr categorization (current ASCVD only includes white and African American), household income, poverty ratio, food security index, fast food intake, pressure to buy low cost meals, and an inability to afford balanced meals. Below is a visualization of different types of feature variables, both those used in to predict current ASCVD scores, and new features I plan implement, that will be used to train a machine learning model to successfully predict our target variable, CVD health outcomes, in adult patients.

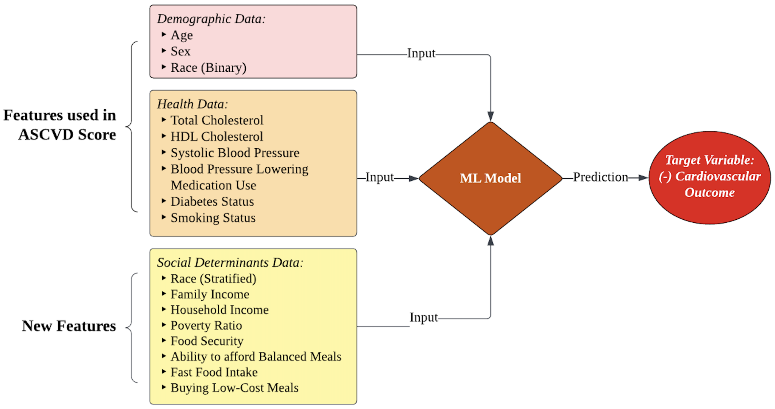


Fig. 3. Machine Learning Pipeline representing the feature variables to be used to train the model to predict CVD outcomes in adult patients.

We will build two models: one using the existing ASCVD variables as a control, and one with the ASCVD variables + the social determinants of health outlined above. After building both models and outputting predictions for each, we will compare the models' predictions of CVD health outcomes comapre to each other, as well as analyze how these predictions compare to the ASCVD risk scores, in order to observe which is more accurate (my model(s), or the risk score) in predicting cardiovascular health outcomes.

B. Machine Learning Pipeline

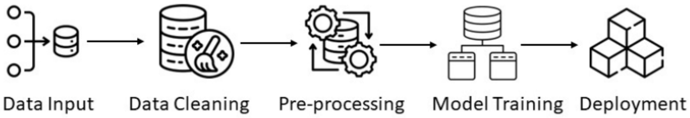


Fig. 4. Machine Learning Pipeline representing the various steps required to build a predictive machine learning model [19].

1) Data Pre-Processing: The first step in building a machine learning model is data pre-processing and preparation [**?**]. This usually begins with data cleaning, which can either remove or impute missing values, correcting errors, and removing outliers [**?**]. Later steps may also include data integration, which may require merging and joining various datasets; all of the data I need from NHANES cannot be downloaded in one file, it will require downloading various files. Each of these files will need to be cleaned and filtered to only retain the relevant data for my project, and joined into one, cleaned comprehensive dataset for training.

2) Data Transformation: Data transformation involves converting data into a format that is more appropriate for modeling by normalizing the data (scaling all numeric attributes in the dataset to a specific range, ex. converting them into a proportion or percentage) and transformation.

For these next several steps in data transformation and model training/tuning, and evaluation, we will be utilizing the approaches by Wiemken and Kelly as outlined in *Machine learning in epidemiology and health outcomes research* [20]. For categorical values, this involves transforming and encoding categorical data into numerical data (0s and 1s, or 0s, 1s, 2s, 3s, etc.) using techniques like one-hot encoding. Larger datasets may also be reduced to include less features using techniques such as dimensionality reduction (ex. PCA Principal Component Analysis).

Finally, the dataset is split into training, validation, and test sets. Typically, about 70-75 percent of the data is used to train the model, and the remaining 25-20 percent is used to validate and test. The training set is used to train the model, the validation set is used to tune the hyperparameters, and the test set is used to evaluate the model's performance.

3) Model Training and Tuning: Next, the training set is used to train the model [20]. This involves feeding the the model our data, and allowing it to learn from the data by adjusting its parameters. We evaluate the model's performance using the validation datasets to pick an optimal hyperparameter. Evaluation metrics include accuracy, precision, recall, F1 score, and root mean squared error. These metrics allow us to compare various hyper-parameters and adjust the model's hyperparameters to improve performance based on which hyper-parameters result in lower error scores and higher accuracy scores. This can be done manually or through automated processes like grid search or random search. We will explain what these hyper-parameters are in more depth in our section outlining the chosen model for our project: Random Forest.

4) Model Evaluation/Deployment: Lastly, these same metrics are used to evaluate the final model's performance and accuracy when ran on new data, the test set [20]. We will compare the model's predictions and accuracy to the computed ASCVD risk scores to see which predictive metric is better; for example, if our model successfully predicts a negative CVD outcome in a patient that was said to have a low ASCVD risk score, we can conclude our model is a better predictor for CVD outcomes in this patient than the ASCVD risk score. This would have to be the case for more than half, or the majority, of our dataset for this conclusion to be true. We can evaluate how well our model compares to ASCVD score predictions for each demographic group as well to see if it performs better on some groups but worse than others and gain more insights.

C. Random Forest

Random Forest is an ensemble machine learning method used for both classification and regression tasks [21]. It operates by constructing a random "forest" of decision trees during training, and outputting the class that is the average of the classes of the individual trees [21]. Health datasets can be large and include many variables: we choose random forest for its many advantages, listed below, and primarily because it can handle high dimensional data without the need for feature reduction, making it suitable for analyzing comprehensive health data sets. A decision tree resembles a flowchart, where the root node represents a sample row in the dataset containing feature variables, and each node in the tree leads to a different path based on the features to predict the final class [21]. Each of the child nodes of the root tree considers a different subset of the training data–this technique is known as bagging. Bagging results in a wide diversity that generally results in a better performing model.

After this initial bagging of the training data, the following levels of nodes consider different subsets of the training data's *features* as well to predict a class [21]. This means that at each split in each tree, the algorithm is randomly selecting from the set of all features, from its respective subset of training data, and continuously splitting the data based on the "best" feature to predict a class [21]. For example, some trees may only consider age, BMI, and income data, while other trees might include an entirely different set of 3 features–if the tree finds that "age" is the most significant feature out of the subset of features it chose for predicting CVD, it will split on the age variable and choose another subset of 3 features (ex. smoking status, sex and cholesterol) to again choose which feature is most helpful in predicting the target variable, based on values such as information gain, and repeat the process [21].

In this example, the size of the feature subset considered at each split is fixed; this is where hyper-parameter tuning may come into play. We may find the model performs better when more than 3 features are considered at each split based on the evaluation metrics–in that case, our final model may consider not just 3, but 5 features at each split. Hyper-parameter tuning may also include altering the number of decision trees used to train the model, and the minimum number of samples (subset of our training data) to be considered at each split. Each of these hyper-parameters have the potential to impact our final model's performance, which is what makes validating our dataset and choosing optimal hyper-parameters for our final model so important.

Finally, each child node from the root outputs a final target outcome (in our case, a CVD outcome) for that row of data, and the model outputs "average" outcome as the final target variable for the row. Below is a visual representation of random forest and decision trees.

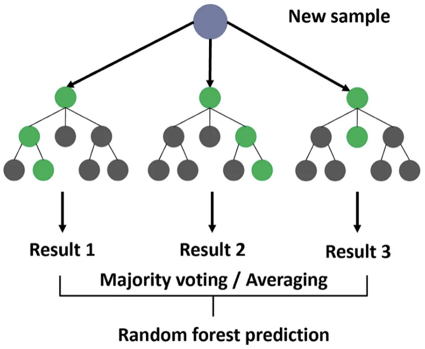


Fig. 5. Random Forest diagram representing how the model uses and averages various decision trees to output a prediction [**?**].

1) Advantages of Random Forest: The random forest model has various advanatages: for example, it can handle both classification and regression tasks, as well as deal with large datasets with higher dimensionality and successfully estimate which variables are important in the classification by nature of the algorithm's splitting process [21]. This will be useful for our dataset, which is relatively large and contains a lot of features.

2) Disadvantages of Random Forest: Some disadvtanages to random forest are that it can be complex and computationally intensive, and may overfit to datasets, making it difficult to generalize on a new dataset [21]. This is particularly harmful when working with health data, where datasets may drastically differ among different patient types and demographics. This is important to consider in the scope of our project.

IV. Results

A. Exploratory Data Analysis

1) Dataset Imbalance: Our dataset has 39156 observations and is incredibly imbalanced, with roughly 52 percent of people fortunately not having experienced a negative cardiovascular outcome, and roughly 6 percent unfortunately having experienced one. 42 percent of patients did not respond, leading to a lot of missing data in our response variable column as well. The amount of NAs burdening the dataset and imbalance makes our model more prone to potential overfitting and not being able to accurately predict the positive class, since it doesn't have enough data to "learn" it. Figure 6 helps visualize this imbalance.

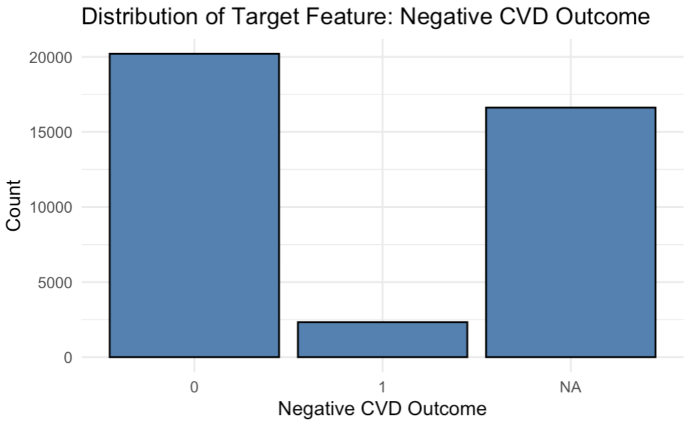


Fig. 6. Bar graph demonstrating the imbalance in our dataset pertaining to our target variable: cardiovascular health outcomes: negative = 1 vs. non-negative = 0.

Possible solutions to these issues include using SMOTE to create synthetic data and balance our dataset, but this is generally not recommended when working with health data as it means we are no longer evaluating model performance *solely* on real, clinical, patient data, but rather also introducing bias by evaluating its accuracy on synthetically generated data as well.

Our dataset also had various feature variables with lots of missing data: to handle this missing data, we used an imputation technique called MICE. Mice uses the other variables in the dataset to predict the missing values in the selected variable. This is typically done using a regression model, but the choice of model can vary depending on the nature of the variable (e.g., logistic regression for binary variables, linear regression for continuous variables) [22].

While imputation is generally *not* recommended to use on a target variable, imputing feature variables is generally more acceptable as inaccuracies and bias in features can be mitigated during model training. Further, the target variable is directly used to measure outcomes, and so errors in the target variable will directly impact the model's credibility and our ability to draw accurate conclusions by interpreting it more heavily; contrastingly, errors made among feature variables are less impactful and severe, and any biases would relate to the features' distributions and relationships with other variables, instead of reflecting a distorted relationship between features and the target variable as a bias would be when imputing the target variable. These issues surrounding imputing the target variable make it especially more problematic in critical fields like healthcare, where we are using machine learning to improve decisions surrounding patient care.

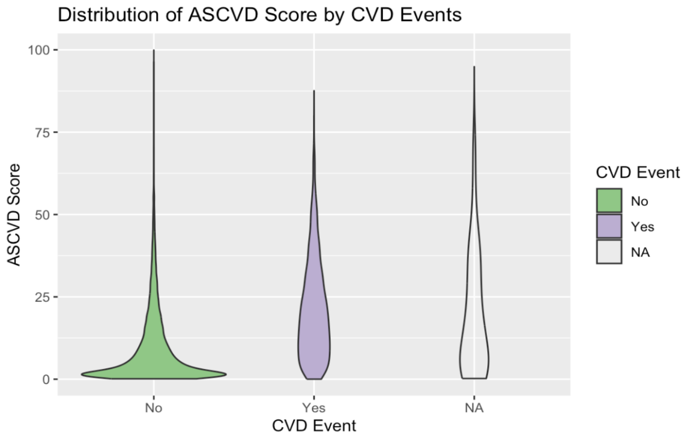


Fig. 7. Violin plot demonstrating how the ASCVD score distribution varies between our two cardiovascular health outcomes: negative = 1 vs. non-negative = 0.

2) ASCVD vs. Cardiovascular Events: We can observe via Figure 7 that for our target variable (CVD events), the distribution of ASCVD scores, which is the existing metric for predicting negative cardiovascular events, varies tremendously. Those who experienced a negative cardiovascular event appear to have higher ASCVD scores and a wider distribution of values as opposed to those who did not. Those who did not experience a negative cardiovascular event generally have lower ASCVD scores closer to 0, indicating low risk. This data demonstrates that the existing metric with its current variables is relatively accurate in predicting negative cardiovascular events, and that our target variable can be used as an accurate benchmark for cardiovascular events relative to the ASCVD score.

3) ASCVD vs. Race: We note in Figure 8 (on the following page) that none of the error bars for the average ASCVD score for each racial demographic overlap, indicating that these differences across racial categories in average ASCVD score *are* indeed statistically significant. The current metric for predicting cardiovascular events only includes races white, Black, and other. This figure suggests that a new model, which includes a more stratified race variable that accounts for these statistically significant differences in the ASCVD score used to predict cardiovascular events, may be more effective in predicting cardiovascular events.

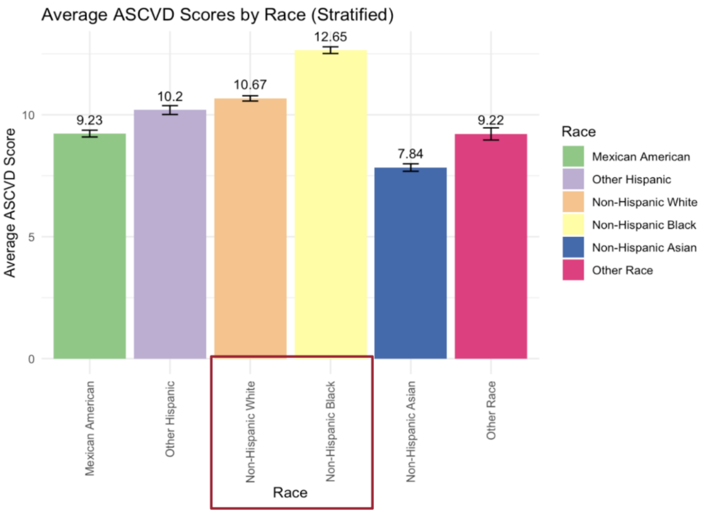


Fig. 8. Bar plot demonstrating how the average ASCVD score varies across different racial demographics.

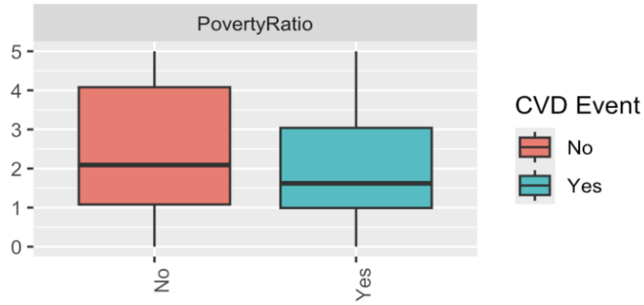


Fig. 9. Box plot demonstrating how the ASCVD score distribution varies across our two cardiovascular health outcomes: negative = 1 vs. non-negative = 0.

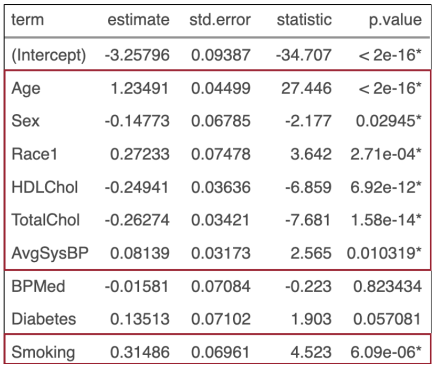
4) ASCVD vs. Poverty Ratio: Similarly: we note in Figure 9 that our interquartile ranges and standard errors for median poverty ratio are overall lower for those who experienced a negative cardiovascular event ("yes" category in blue) than for those who didn't ("no" category in red). Poverty ratio is defined as the total family income divided by the poverty threshold [23]. A lower poverty ratio usually indicates a higher likelihood that a household with that given income is living in poverty.

This finding suggests a correlation between experiencing a negative cardiovascular event, and an individual's poverty ratio. Consequently, a new future model that includes income, or poverty ratio as a metric related to income/socio-economic status, may be more effective in predicting cardiovascular events.

B. Logistic Regression: Existing Variables

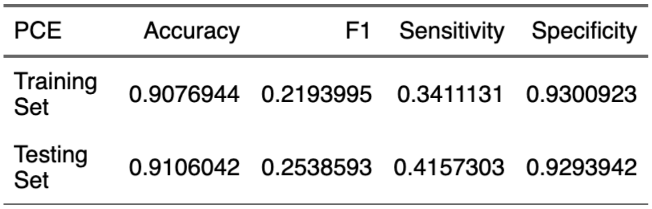
Although we have not fully trained and tuned our random forest model, we have trained two basic logistic regression models: one using the existing ASCVD variables as a control to understand our current dataset's strengths and weaknesses, and one using the existing ACSVD variables and our newly introduced social determinants of health.

We note as seen in Figure 10 that almost all the existing variables used in the ASCVD calculator to predict cardiovascular health outcomes are indeed significant in predicting the target variable, as is evident by their p-values below 0.5, with the most significant coefficients stemming from the age, sex, race, both cholesterol, blood pressure, and smoking variables. These results indicate that these variables should not be removed from any future models, as they significantly impact the model's ability to predict cardiovascular outcomes.



term	estimate	std.error	statistic	p.value
(Intercept)	-3.25796	0.09387	-34.707	< 2e-16*
Age	1.23491	0.04499	27.446	< 2e-16*
Sex	-0.14773	0.06785	-2.177	0.02945*
Race1	0.27233	0.07478	3.642	2.71e-04*
HDLChol	-0.24941	0.03636	-6.859	6.92e-12*
TotalChol	-0.26274	0.03421	-7.681	1.58e-14*
AvgSysBP	0.08139	0.03173	2.565	0.010319*
BPMed	-0.01581	0.07084	-0.223	0.823434
Diabetes	0.13513	0.07102	1.903	0.057081
Smoking	0.31486	0.06961	4.523	6.09e-06*

Fig. 10. Significant coefficient results from our logistic regression model, indicating feature variables which were significant in predicting our target variable: cardiovascular health outcomes



PCE	Accuracy	F1	Sensitivity	Specificity
Training Set	0.9076944	0.2193995	0.3411131	0.9300923
Testing Set	0.9106042	0.2538593	0.4157303	0.9293942

Fig. 11. Table demonstrating model performance with **existing ASCVD variables**: includes accuracy, F1 score, sensitivity, and specificity.

This regression model also has an overall high accuracy, performing at close to 91 percent accuracy on both the training and testing sets, as seen in Figure 11. This means our model did not overfit, which is good; however, the low F1 score and low sensitivity suggest that the model may not be performing very well in classifying the "positive cases" (occurence of negative CVD outcomes). This is likely due to the lack of data available for negative CVD outcomes. The high specificity indicates that the model is useful for confirming the absence of the negative cardiovascular outcomes, but not for detecting its presence, which is precisely what we need it for, indicating we may need more data to build a better model. Further, the area under the curve, as seen in Figure 12, is 0.814, which is relatively close to 1 and indicates overall good model performance. We used a threshold of 0.3 on our model.

C. Logistic Regression: Social Determinants of Health

We also trained a basic logistic regression model using the existing ASCVD variables *and* our chosen social determinants of health to compare to our model control. Similar

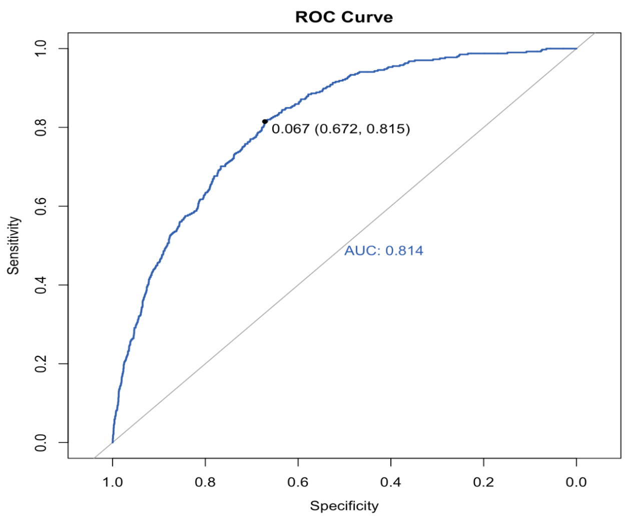
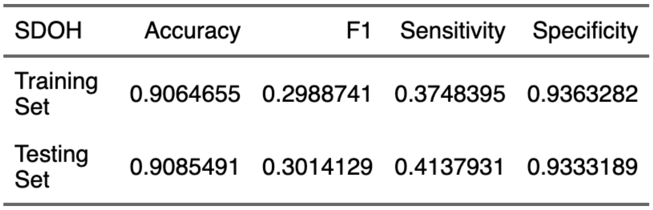


Fig. 12. ROC curve demonstrating model performance and optimal threshold for model using only the existing variables. The area under the curve (AUC) is 0.814. Optimal threshold optimizing both sensitivity and specificity is identified to be 0.672.

to our model with just the ASCVD variables, as observed in Figure 12, the area under the curve was equal to .831 indicating our overall model performance is good and better than that of a random classifier.

We note here again, although not included due to size limitations in the output, that almost all the existing variables used in the ASCVD calculator to predict cardiovascular health outcomes were indeed significant in predicting the target variable in this model as well, and had significant coefficients below 0.5. Out of all the new variables reflecting social determinants of health introduced, only the stratified race variables and poverty ratio had significant coefficients at the 0.05 level. We again used a threshold of 0.3 on our model.



SDOH	Accuracy	F1	Sensitivity	Specificity
Training Set	0.9064655	0.2988741	0.3748395	0.9363282
Testing Set	0.9085491	0.3014129	0.4137931	0.9333189

Fig. 13. Table demonstrating model performance including **social determinants of health variables**: includes accuracy, F1 score, sensitivity, and specificity.

This model also seems to perform well in terms of overall accuracy, also at 91 percent as observed in Figure 13, similar to that of the previous model excluding social determinants. It also appears to struggle with correctly identifying negative cardiovascular outcomes, as again evidenced by the low F1 score and low sensitivity, meaning the model is not equivalently good at predicting both positive and negative case outcomes. The specificity is again high, again indicating a strong ability to correctly identify negatives, but not the same for positive (the occurence of a negative cardiovascular event). Similar to the findings with our previous model, it appears more data may be needed to build a better performing model that is also able to predict negative cardiovascular health outcomes at a higher rate.

V. Conclusion and Future Work

We conclude the logistic regression model with the PCE variables, excluding the SDOH variables, performs mildly better. Due to our dataset imbalance, more data is necessary to produce a model that is better at identifying positives (the occurence of a negative cardiovascular event). All variables currently being used to predict cardiovascular events in the existing ASCVD score were found to be significant, with social determinants related to income (ex. food security and poverty ratio) as well as race (stratified for more demographic categories) also found to be significant at the 0.05 significance level.

Next steps include generating more data to balance our dataset potentially using SMOTE, removing the social determinant variables we included if they do not appear to be improving model performance based on coefficient scores, relative to the existing ASCVD variables being leveraged, and re-running our logistic regression models with only those found to be significant as well as new SDOH variables.

We would like to train future models to use an optimal threshold where the percent of predicted negative outcomes in the training set equals the percent of actual negative outcomes in the whole dataset, optimizing sensitivity.

Furthermore, we also need to build and hyper-parameter tune our random forest model next and visualizing those results to see if they are better than our current logistic regression model.

References

[1] J. W. Tsai, “Evaluating the impact and rationale of race-specific estimations of kidney function: Estimations from u.s. nhanes, 2015-2018,” in *EClinicalMedicine*, vol. 42, no. 101197, 2021.

[2] P. L. Temporelli, “Risk scores, atherosclerotic cardiovascular disease and the crystal ball,” in *European Journal of Preventive Cardiology*, vol. 28, no. 14, 2020, pp. 14–15.

[3] “Ascvd risk calculator,” in *American College of Cardiology and American Health Association*.

[4] J. Rana, G. Tabada, and M. Solomon, “Accuracy of the atherosclerotic cardiovascular risk equation in a large contemporary, multiethnic population,” in *J Am Coll Cardiol*, 2016, p. 2118–2130.

[5] “Ascvd risk estimator,” in *MDedge Federal Practitioner*, vol. 31, no. 5, 2014.

[6] “Social determinants of health,” in *World Health Organization*.

[7] M. Chen, X. Tan, and R. Padman, “Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review,” in *Journal of the American Medical Informatics Association*, vol. 27, no. 11, 2020, p. 1764–1773.

[8] Y. Zhao, E. Wood, N. Mirin, S. Cook, and R. Chunara, “Social determinants in machine learning cardiovascular disease prediction models: A systematic review,” in *American Journal of Preventive Medicine*, vol. 61, no. 4, 2021, pp. 596–605.

[9] J. Hammond, R. Waken, M. Sims, K. Henderson, and K. J. Maddox, “The addition of social determinants of health improves the predictive accuracy of the pooled cohort equations for 10-year ascvd events in african americans,” in *Epidemiology, Big Data and Precision Medicine: Social, Structural and Systemic Determinants of Cardiovascular Disease Risk*, 2022.

[10] M. Xia, J. An, M. M. Safford, L. Colantonio, P. Muntner, K. Reynolds, A. E. Moran, and Y. Zhang, "Atherosclerotic cardiovascular disease risk associated with social determinants of health at individual and area levels," in *Epidemiology, Big Data and Precision Medicine: Neighborhood and Multi-Level Social Determinants of Health in Association with Cardiovascular Health*.

[11] A. Hammoud, H. Chen, A. Ivanov, J. Yeboah, K. Nasir, M. Cainzos-Achirica, S. U. K. Alain Bertoni, M. Blaha, D. Herrington, and M. D. Shapiro, "Implications of social disadvantage score in cardiovascular outcomes and risk assessment: Findings from the multi-ethnic study of atherosclerosis," in *HomeCirculation: Cardiovascular Quality and Outcomes*, vol. 16, 2023.

[12] U. Health, "Social determinants of health," in *Sustainability*, 2018.

[13] A. Ghosh, S. Venkatraman, and M. G. Nanna, "Risk prediction for atherosclerotic cardiovascular disease with and without race stratification," in *JAMA Cardiol*, vol. 9, no. 1, 2024, pp. 55–62.

[14] J. A. Fain, "Nhanes: Use of a free public data set," in *European Journal of Preventive Cardiology*, vol. 43, no. 2, 2017.

[15] S. Badillo, B. Banfai, F. Birzele, I. I. Davydov, L. Hutchinson, T. Kam-Thong, J. Siebourg-Polster, B. Steiert, and J. D. Zhang, "An introduction to machine learning," in *Clinical Pharmacology Therapeutics*, 2020.

[16] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," in *PeerJ Computer Science*, 2021.

[17] Y.-C. Hsiao, C.-Y. Kuo, F.-J. Lin, Y.-W. Wu, T.-H. Lin, H.-I. Yeh, J.-W. Chen, and C.-C. Wu, "Machine learning models for ascvd risk prediction in an asian population — how to validate the model is important," in *Acta Cardiol Sin*, vol. 39, no. 6, 2023.

[18] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, 2010.

[19] V. S. D. Prasad, "The importance of a machine learning pipeline," in *Softnautics*, 2023.

[20] T. L. Wiemken and R. R. Kelley, "Machine learning in epidemiology and health outcomes research," in *Annu Rev Public Health*, vol. 41, no. 1, 2020, pp. 21–36.

[21] A. Jehad, "Random forests and decision trees," in *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 5, 2012, pp. 272–278.

[22] K. Woźnica, "ow to predict with missing values in r?"

[23] U. S. C. Bureau, "How the census bureau measures poverty," in *Guidance for Poverty Data Users*, 2023.