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 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, disproportionately disadvantage Black communities because Black patients are generally scored as having "better" kidneys, and thus their kidney health concerns are sometimes brushed off and not taken as seriously by health professionals as they should be.

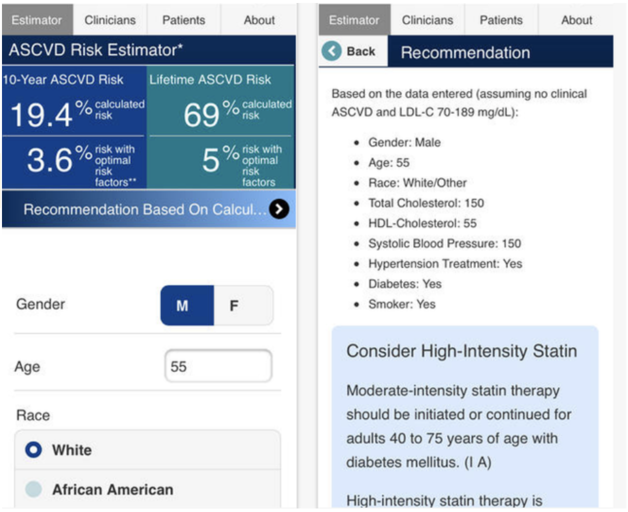
Our research goal is to build a machine learning model incorporating other variables, such as social determinants of health, not currently being leveraged in computing existing risk scores for cardiovascular health outcomes and analyze how such a machine learning model compares to existing risk scores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

II. 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 [1]. 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 [1]. This is typically the ASCVD risk score referenced in most studies.

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 [2]. 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 [3].



ASCVD Risk Estimator*

10-Year ASCVD Risk	Lifetime ASCVD Risk
19.4 % calculated risk	69 % calculated risk
3.6 % risk with optimal risk factors**	5 % risk with optimal risk factors

Recommendation Based On Calcul...

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. This is an image 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.

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," [4]. 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 of risk scores, not just cardiovascular risk scores [5].

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 [6].

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

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 [7]. 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 [7]. 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 [8].

However, other research studies have also explored social determinants of health impacts on ASCVD risk scores using data from the Multi-Ethnic Study of Atherosclerosis created an index of baseline Social Disadvantage Score (SDS) to explore its association with incident atherosclerotic cardiovascular disease (ASCVD) and all-cause mortality and impact on ASCVD risk prediction. 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 [9]. 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.

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 [10]. 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.

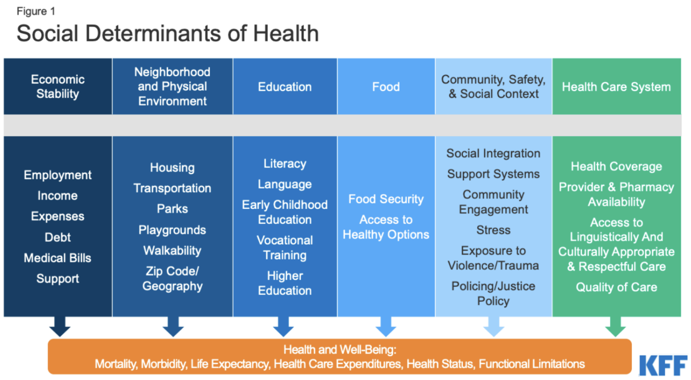


Figure 1
Social Determinants of Health
Economic Stability
Neighborhood and Physical Environment
Education
Food
Community, Safety, & Social Context
Health Care System
Employment
Income
Expenses
Debt
Medical Bills
Support
Housing
Transportation
Parks
Playgrounds
Walkability
Zip Code/ Geography
Literacy
Language
Early Childhood Education
Vocational Training
Higher Education
Food Security
Access to Healthy Options
Social Integration
Support Systems
Community Engagement
Stress
Exposure to Violence/Trauma
Policing/Justice Policy
Health Coverage
Provider & Pharmacy Availability
Access to Linguistically And Culturally Appropriate & Respectful Care
Quality of Care
Health and Well-Being:
Mortality, Morbidity, Life Expectancy, Health Care Expenditures, Health Status, Functional Limitations
KFF

Fig. 2. This is an image demonstrating various social determinants of health and how they contribute to overall health outcomes, developed by KFF.

III. Background

A. NHANES

The National Health and Nutrition Examination Survey (NHANES) is a downloadable public use data set used ~~by diabetes educators~~ to document health care utilization, health status of various age groups, and related personal and lifestyle characteristics [11]. 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. I am interested in 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, data used in existing ASCVD risk score calculators, such as general health and demographic data, and cardiovascular health outcomes.

B. 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 [12]. 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 [12]. For our research purposes, our target variable would be ~~the ASCVD risk s~~ and our feature variables would be various numerical and categorical values related to age, gender, social determinants of health, blood pressure levels, diabetes status, etc. The risk score we will be focusing on for our research purposes is a 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.

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 [12]. Popular machine learning models include KNN, Random Forest, Decision Tree, and ~~Linear Regression~~ [13]. 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 [14]. 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 [15].

I would want to compare the model's predictions to ASCVD risk score predictions and see how they compare, as well as which is more accurate for known cardiovascular health outcomes.

IV. TBD Design / Implementation / Algorithm

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

V. Results

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

VI. Conclusion and Future Work

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

References

[1] P. L. Temporelli, "Risk scores, atherosclerotic cardiovascular disease and the crystal ball," vol. 28, no. 14, 2020, pp. 14–15.

[2] "Ascvd risk calculator," in *American College of Cardiology and American Health Association*.

[3] J. Rana, G. Tabada, and M. Solomon, "Accuracy of the atherosclerotic cardiovascular risk equation in a large contemporary, multiethnic population," 2016, p. 2118–2130.

[4] "Social determinants of health," in *World Health Organization*.

[5] 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," vol. 27, no. 11, 2020, p. 1764–1773.

[6] Y. Zhao, E. Wood, N. Mirin, S. Cook, and R. Chunara, "Social determinants in machine learning cardiovascular disease prediction models: A systematic review," vol. 61, no. 4, 2021, pp. 596–605.

[7] 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," 2022.

[8] 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."

[9] 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," vol. 16, 2023.

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

[11] J. A. Fain, "Nhanes: Use of a free public data set," vol. 43, no. 2, 2017.

[12] 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," 2020.

[13] 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," 2021.

[14] 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," vol. 39, no. 6, 2023.

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