JOURNAL OF
EMERGING INVESTIGATORS

Article

Using broad health-related survey questions to predict the presence of coronary heart disease

Aiden Chavda[1], Jason Hyun[2]

[1] Los Altos High School, Los Altos, California

[2] Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, California

SUMMARY

Coronary heart disease (CHD) is the leading cause of death in the United States and was responsible for the deaths of almost 700,000 people in 2021. CHD is influenced by a variety of factors, including genetics and behavioral patterns. It is a dangerous disease characterized by a clogging of the arteries, which can cause myocardial infarction if left unchecked. CHD can develop without showing any symptoms, making its prediction all the more important. However, current methods can only predict CHD accurately using expensive clinical equipment and tests. Past machine learning projects aimed at predicting and preventing CHD typically depended on these inconvenient clinical procedures. This study tests the hypothesis that CHD can be predicted by applying machine learning to demographic, clinical, and behavioral data provided by survey responses. Trained on over 300,000 samples from the CDC's 2022 Behavioral Risk Factor Surveillance System, binary classification models predicting CHD and myocardial infarction history achieved Matthews correlation coefficients (MCCs) ranging from 0.299 to 0.313 and accuracies ranging from 0.716 to 0.726 during 5-fold cross validation. Individual demographic-specific models were also trained and could achieve MCCs of up to 0.504. Lastly, interpretation of these models using coefficient weights recovered associations between CHD and behavioral, clinical, and demographic variables that were consistent with previous studies. This study demonstrates a proof of concept for predicting the presence of CHD by looking solely at data provided by responses to broad health-related survey questions.

INTRODUCTION

The most common cause of death in the United States is heart disease, responsible for the deaths of almost 700,000 people per year (1). Coronary heart disease (CHD) is a type of heart disease that arises when the arteries of the heart become restricted and fail to supply sufficient oxygen-rich blood to the heart itself (2). CHD has a variety of causes and current prevention techniques include living a heart-healthy lifestyle involving dietary guidelines and regular exercise along with cholesterol, blood pressure, or blood-thinning medication, all intended to reduce the risk of clotting. However, these prevention methods are not guaranteed to succeed in stopping the development of CHD (2). CHD may present with a variety of symptoms, including shortness of breath and chest pain (3). However, an important characteristic of CHD is its tendency to develop silently, showing few to no symptoms prior to inducing a potentially fatal myocardial infarction, also known as a heart attack (4–5). The often-silent nature of CHD makes it especially challenging to diagnose and treat; while various routine lab tests such as those for blood cholesterol levels are often used to gauge CHD risk, they do not provide an actual diagnosis of CHD and can require stressful procedures such as a blood draw. As a result, both doctors and patients are often left with very few warning signs or potential indicators that can help them diagnose CHD (4). CHD is also not yet curable, and once a person has been diagnosed with CHD, additional intervention is needed to manage the risk of a potentially fatal heart attack (6). As a result, while it is possible to manage progressed CHD, preferable approaches would emphasize combating CHD with earlier, less-invasive, preventative measures (6).

One way to prevent the development of CHD is by identifying high risk individuals and alerting them. This intervention can allow for an earlier diagnosis of CHD and earlier adoption of lifestyle changes that may reduce the risk of life-threatening complications. Predicting and diagnosing heart disease may be sensitive to many types of variables, including demographic, clinical, and behavioral ones. For example, researchers have known for years that black adults are more likely to get diagnosed with heart disease due to socioeconomic disadvantages that lead to a worse standard of living and decreased access to sufficient healthcare (7). While this is an example of an informative demographic variable, demographics alone do not provide enough information to make accurate predictions on an individual level. These variables are often combined with clinical variables, such as other health conditions, as well as behavioral ones in order to sufficiently gauge a patient's risk of CHD. Past researchers have often attempted to utilize these demographic, behavioral, and clinical features along with informative laboratory data as well as machine learning and artificial intelligence in an attempt to predict heart disease on an individual level. This approach has been seen in many recent studies which have used laboratory tests such as data from CT scans and blood tests to create highly accurate models for diagnosing CHD (8–10).

While past projects using machine learning have been successful in predicting and diagnosing CHD, they have often relied on laboratory tests ranging from the simple yet often stressful blood draw to more expensive examinations such as CT scans, to obtain the necessary data, causing significant inconveniences for both researchers and patients (8). Despite these limitations, machine learning can nevertheless be utilized along with the vast amount of available health data to train models as tools for cardiologists to accurately assess CHD risk (9). Many of the previous studies in this field, un-

https://doi.org/10.59720/24-006

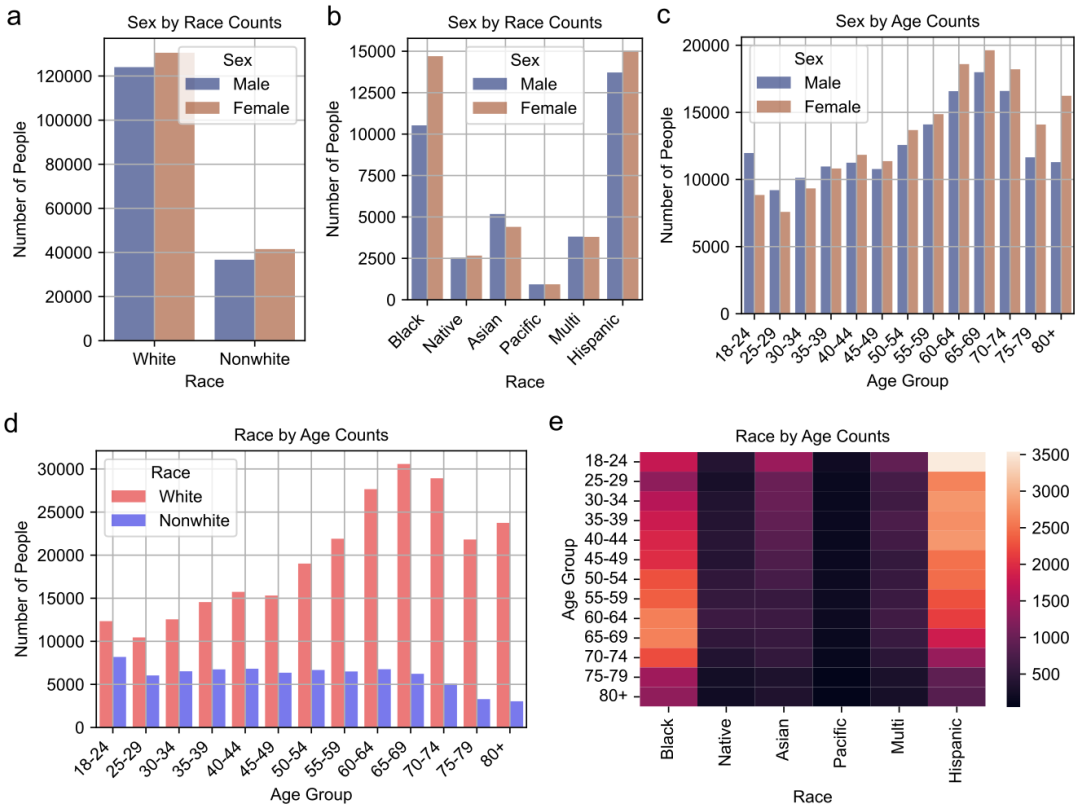

Figure 1: Summary of BRFSS 2022 participants by demographic. a) Proportion of male and female respondents for 'white' and 'nonwhite' categories. b) Proportion of male and female respondents for all 'nonwhite' race categories. c) Proportion of male and female respondents for different age groups. d) Proportion of 'white' and 'nonwhite' respondents for different age groups. e) Proportion of each 'nonwhite' race category for different age groups.

der the restriction of requiring laboratory data stemming from clinical tests, have not taken advantage of the volume of other patient data types that have recently become available to the public. Some of the most recent studies applying machine learning to the diagnosis of heart disease have used datasets with less than one thousand samples (10–13). While these studies included additional features such as laboratory data, they are limited by their scale, as models trained on fewer samples are often less accurate while also representing a smaller and potentially more biased portion of the population. Instead of using data with more informative features at a smaller-scale, we instead chose to utilize a larger, publicly available dataset containing a variety of demographic, clinical, and behavioral data in an attempt to address sample-size issues which have affected some of these previous studies.

This study uses the Center for Disease Control's (CDC) Behavioral Risk Factor Surveillance System (BRFSS) 2022 phone survey as a large source of data for future machine learning projects (14). The advantages of this dataset are that it is easily accessible to the public and contains many samples covering many different demographics, including patient age, sex, and race. Conclusions drawn from this data could be applied to new patients with relative ease, as they would only have to answer a couple of questions to use the model. This is in contrast to existing models, which, along with data which can be easily obtained such as blood pressure, also subject the patient to the burden of various laboratory tests. In this study, we tested the hypothesis that coronary heart disease could be predicted by applying machine learning to broad, health-related survey data. We developed a workflow for training and interpreting models to predict CHD from survey responses at a scale of over 300,000 samples. These models were moderately accurate and, after demonstrating more promise than previous survey-based studies at smaller scales, can be seen as evidence demonstrating the feasibility of predicting a diagnosis of CHD using only behavioral, clinical, and demographic data derived from survey responses. Additionally, the models suggested associations between several behavioral patterns and the presence of CHD. These associations could become the subject of future experiments and, if proven true, could help contribute to prevention processes such as preventative screening.

RESULTS

Dataset

A preprocessed dataframe of 340,200 samples and 32 features, including behavioral, clinical, and demographic data, was built from the CDC's 2022 BRFSS survey (14). The dataset was balanced when it came to sex (48.3% male, 51.7% female) but was highly skewed towards white (76.5% white, 23.5% nonwhite) and elderly (64.9% above age 50) populations and lacked data in some of the younger and nonwhite demographics **(Appendix, Figure 1A–C)**. For example, individuals of Pacific Islander descent were very rare, representing 0.54% of all samples. In addition, many nonwhite demographics had decreasing sample count with respect to

https://doi.org/10.59720/24-006

age, especially for groups above the age of 65 **(Figure 1D,E)**. However, despite these limitations, there were at least 1000 samples for every race group, which we deemed to be sufficient for training and interpreting models.

The target variable was built from the same dataset by identifying any individuals who had self-reported as previously diagnosed with CHD or suffered from a myocardial infarction. A holdout dataset consisting of 10% of randomly selected samples (34,020 samples) was set aside, and the remaining samples were used to train and test Random Forest, Naive Bayes, Support Vector Machine, and Logistic Regression binary classifiers. The best binary classifier was defined as the model achieving the highest test set Matthews correlation coefficient (MCC) over 5-fold cross validation. The performance of global models was evaluated using both accuracy and MCC across all binary classifiers. MCC is an evaluation metric that ranges from -1 to 1, with -1 representing a model that is always wrong, 0 representing a model that is randomly guessing, and 1 representing a model that is always correct. MCC differs from accuracy in that it considers all four possible binary classification outcomes evenly (true positives, true negatives, false positives, false negatives), instead of focusing on the total number of correct predictions, like accuracy does. An alternative metric is necessary for unbalanced datasets such as this one, as only around 10% of the samples were positive for CHD, meaning a model would have an accuracy of 90% if it predicted no CHD for every sample. This metric was useful when evaluating methods with deceptive results, such as Naive Bayes, which had a much higher accuracy than the other models, yet a much lower MCC, demonstrating how accuracy was often not representative of a model's true performance **(Figure 2A,B)**. The best model type was identified and used to train both global models (using all samples) as well as demographic-specific models (using samples from specific race and/or sex cohorts). Models were interpreted by identifying the features with the highest and lowest weights **(Figure 3)**.

Global Model

By looking at the accuracy and MCC for each classifier, Logistic Regression was identified as the top model choice, with mean values of 0.725 and 0.305 for the accuracy and MCC, respectively **(Figure 2A,B)**. This model's performance was evaluated on the previously unseen holdout test set and achieved an accuracy of 0.724 and MCC of 0.307 **(Figure 2A,B)**. These values suggested a moderate correlation between the model's prediction and whether or not the patient self-reported as having CHD. The four highest weighted variables of the final, global model were the person's age, average sleeping time, height, and sex **(Figure 2C)**. Interpretation of these variables suggests that people who slept more, were older, were taller, or were male were more likely to be diag-

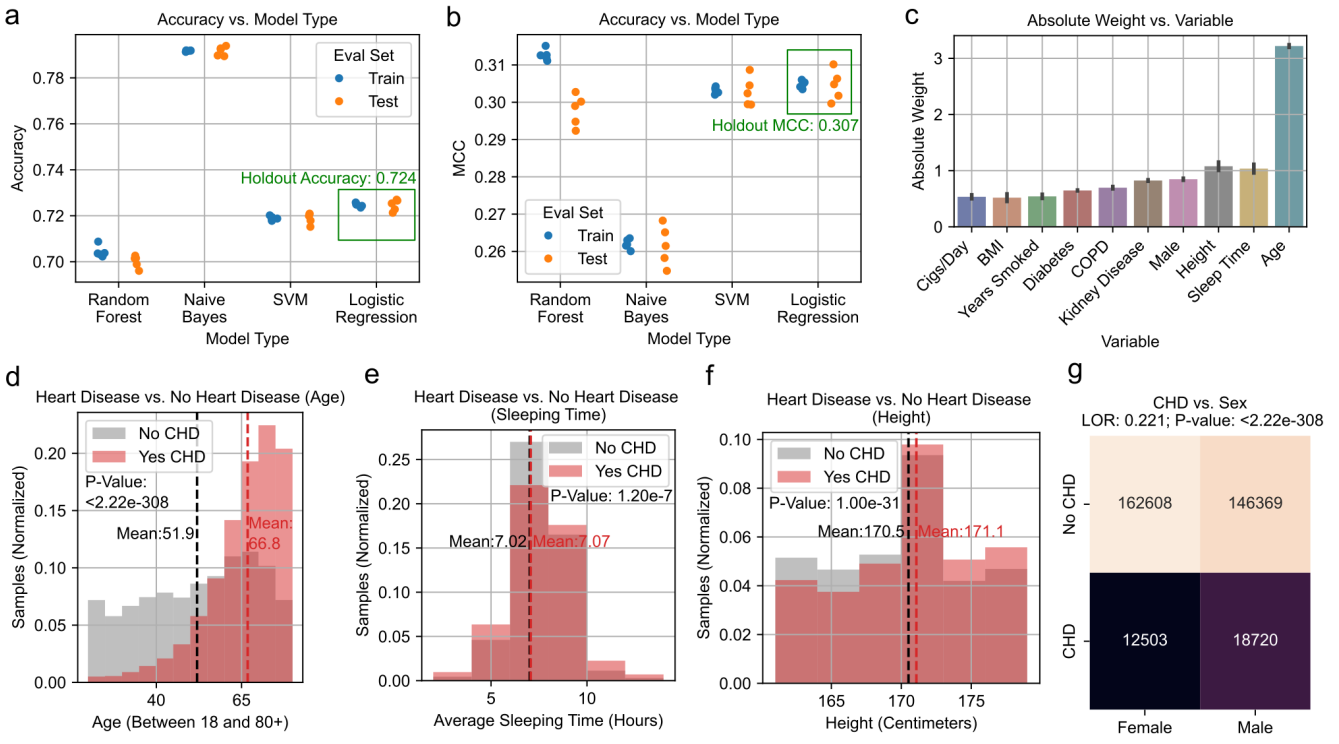

Figure 2: Performance and interpretation of global CHD models. a-b) Performance of each model type from 5-fold cross validation. Train and test sets were obtained using a stratified K-fold approach. MCC is the Matthews correlation coefficient. Holdout accuracy and MCC were calculated by evaluating the final model on 10% of the data which was withheld from cross validation. c) Highest weighted variables from a logistic regression model connecting features to a diagnosis of CHD. Absolute weight was determined by taking the absolute value of each variable's feature weight for each of 5 folds. Bars represent means, and error bars represent standard deviations. d-g) Distributions for CHD positive vs. negative participants for the top four most highly weighted variables. Log odds ratio (LOR) was determined by taking the log base 10 of the odds ratio. P-values, means, and LORs are displayed on the graphs. Two-sided Mann-Whitney U-tests (n=340,200) were applied for continuous variables Sleeping Time and Age, while two-sided Fisher's exact test was used for binary variable Sex. Dotted lines represent means for each dataset (black: No CHD, red: Yes CHD).

https://doi.org/10.59720/24-006

nosed with CHD **(Figure 2D,G)**. Both age (p<2.22e-308, two-sided Mann-Whitney U-test) and sex (p<2.22e-308, two-sided Fisher's exact test) were significantly associated with CHD to family-wise error rate (FWER) < 0.05 (two-sided Fisher's exact test, 55 tests, Bonferroni correction, threshold=0.0009) and had large effect sizes. Being male over female had a LOR (log10 odds ratio) of 0.221 for CHD, and there was an approximate 15-year difference in mean age between CHD and no CHD samples (means of 51.9 years with no CHD vs. 66.8 years with CHD). While sleeping time (p=1.20e-7, two-sided Mann-Whitney U-test) and height (p=1.00e-31, two-sided Mann-Whitney U-test) were also found to be statistically significant to FWER < 0.05 (two-sided Fisher's exact test, 55 tests, Bonferroni correction, threshold=0.0009) and had large model weights, the effect sizes were only a difference of 0.05 hours (3 minutes) in mean sleep per night (means of 7.07 hours CHD vs. 7.02 hours no CHD) and 0.6 centimeters in mean height (means of 171.1 centimeters CHD vs. 170.5 centimeters no CHD) between CHD and no CHD samples.

Individual Demographics

The dataset was then further split based on both race and sex and the resulting datasets were used to train additional demographic-specific models and similarly evaluated through 5-fold cross validation. As expected, model performance varied by demographic. For example, our models had higher MCCs when predicting CHD in male-only datasets compared to female-only datasets across every race. In addition, models for Native (individuals of Native American descent), Pacific Islander, and multiracial samples saw higher MCCs of around 0.30, while those for Black, Hispanic, and Asian samples reported relatively low MCCs around 0.25 **(Figure 4A,B)**. Further examination of the precision and recall of each model showed the effects of this 0.05 MCC difference. The mean precision, a measure of the fraction of predicted positive diagnoses that were correct, of the Black, Hispanic, and Asian samples ranged from 0.037 to 0.1 lower than those of the global model (Black - 0.180; Asian - 0.117; Hispanic - 0.165 vs. Global - 0.217), indicating that its predicted positive CHD diagnoses were wrong 3-10% more often than the global model. While lower precision is sometimes associated with higher recall, a measure of the fraction of true CHD positive samples which were accurately predicted, in this case the Asian, Black, and Hispanic demographics did not display higher recall than the global model (Black - 0.741; Asian - 0.776; Hispanic - 0.754 vs. Global - 0.774). Overall, the impact of the lower MCC of these models can be seen in the higher rate of false positive predictions without a higher rate of true positive CHD cases diagnosed.

Throughout the demographic-specific models, interpretation of model weights indicated that the variables with the strongest positive correlations with the presence of CHD were demographic and clinical ones. Age remained the most informative, with the highest mean weight and mean absolute value of weight throughout all of the individual models **(Figure 4C)**. However, these models also suggested that the presence of either COPD (chronic obstructive pulmonary disease) or kidney disease were both major correlates with CHD. These variables had LORs ranging from 0.6 to 0.8, p-values ranging from less than 2.22e-308 to 3.20e-07, and were significantly associated with CHD to FWER < 0.05 (two-sided Fisher's exact test, 55 tests, Bonferroni correction, threshold=0.0009) throughout every demographic, suggesting a consistently large and significant effect **(Figure 4D,E)**. Other variables such as sex (for race-specific models which included both sexes) and sleeping time remained consistently highly weighted in most models.

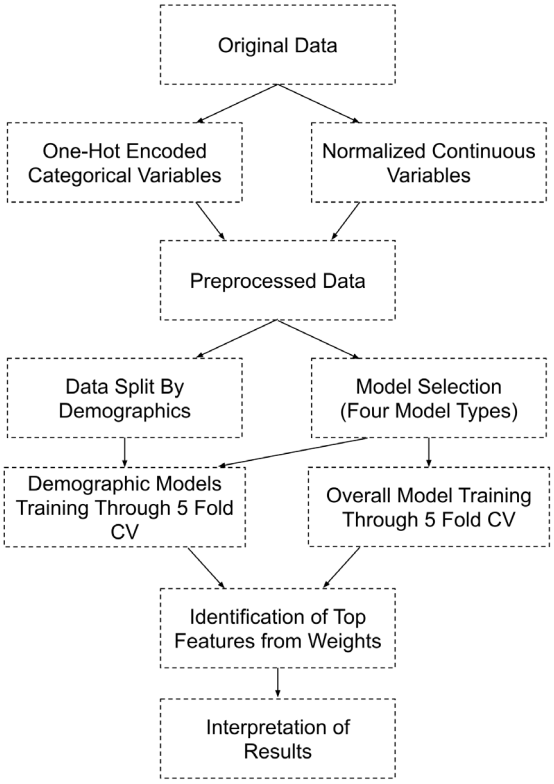

Figure 3: Machine learning workflow. Data splitting was done based on race and sex. Top features were identified by looking at the average absolute value of each feature's weight. Cross validation is abbreviated as "CV".

Upon further examination of the individual demographic weights, some variables had differing effects across demographics. For example, variables indicating the number of days one had consumed alcohol in the last month (30 days), the average amount of alcohol consumed each time, and the number of years smoked had differing associations across demographics **(Figure 5A–C)**. While the number of days an individual drank in the last 30 days usually had a small negative correlation with CHD, this negative correlation increased noticeably in the Asian demographic. In this demographic, individuals without CHD drank, on average, over half a day more in the last month than those with CHD (means of 2.79 days CHD vs. 3.35 days no CHD) **(Figure 5D)**. An analogous phenomenon occurred with the average amount of alcohol drank variable in the Pacific Islander demographic, in which patients diagnosed with CHD drank on average about a third of a glass less per occasion than those without CHD (means of 1.22 drinks CHD vs. 1.64 drinks no CHD) **(Figure 5E)**. Lastly, while years smoked had a small positive correlation with the diagnosis of CHD in most of the demographics, that positive correlation with CHD diagnosis saw an increase in the Native demographic. Individuals with CHD reported 11 more years smoked on average than those without CHD in this specific demographic (means of 23.5 years CHD vs. 12.1

https://doi.org/10.59720/24-006

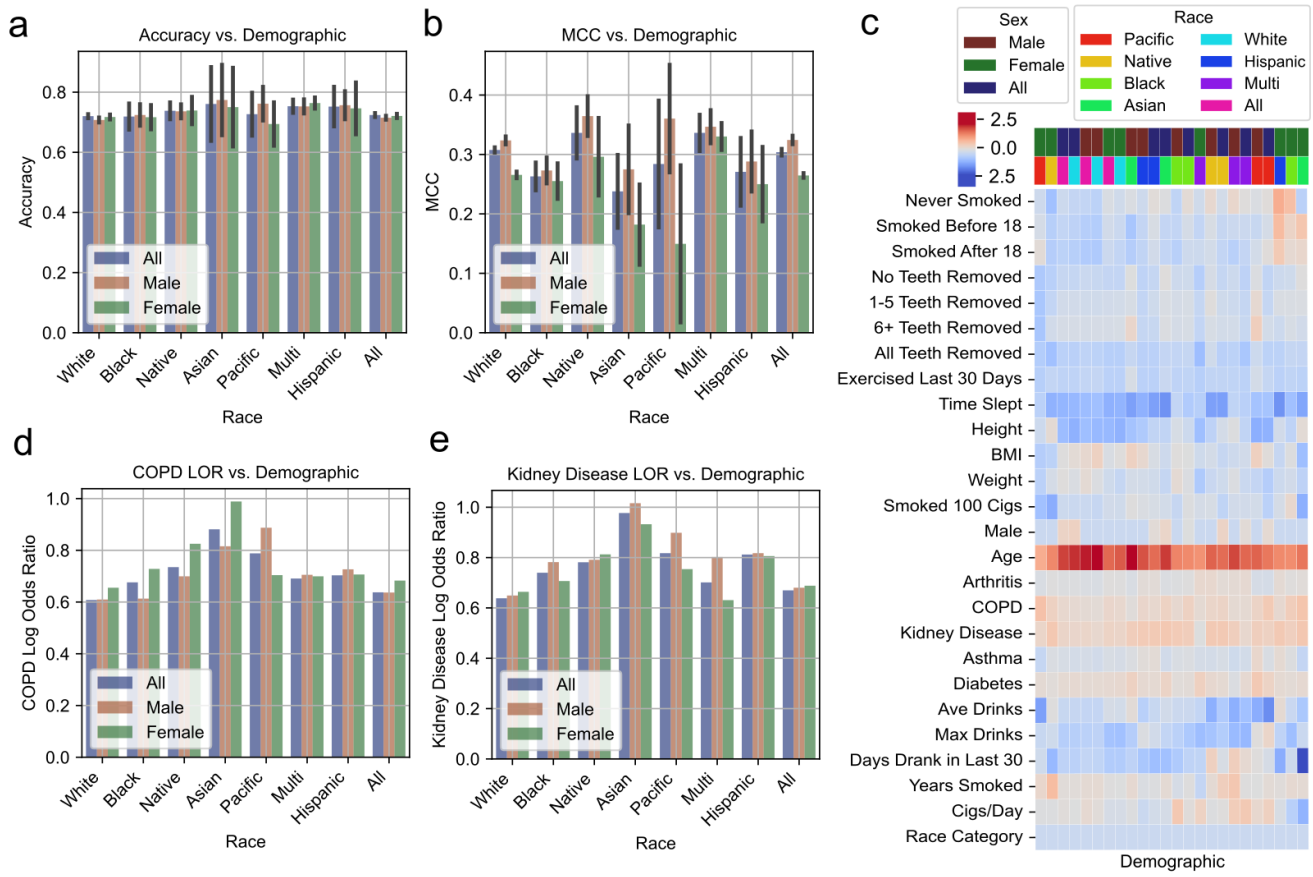

Figure 4: Performance and interpretation of demographic-specific CHD models. a-b) Performance of the model across each racial demographic. Bars represent means, and error bars represent standard deviations. Logistic Regression was used as the model type for all tests. Accuracies and Matthews correlation coefficients are from test sets during 5-fold cross validation. c) Average weight of each variable across each demographic over 5-fold cross validation. Rows and columns were ordered by using hierarchical clustering with average linkage applied to Euclidean distances on the average weight of each variable across each demographic. d-e) Log10 odds ratios (LORs) of COPD (d) and kidney disease (e) that were highly weighted across all demographics. COPD represents Chronic obstructive pulmonary disease.

years no CHD), explaining its high variable weight **(Figure 5F)**. While the effect sizes of these variables were large, it is important to note that, possibly because of sample size, the association between the average number of drinks variable and CHD in the Pacific demographic was given a p-value of 0.00362 by the two-sided Mann-Whitney U-test and found not statistically significant after the Bonferroni correction (55 tests, FWER < 0.05). However, both days drank in the last 30 days (p=7.99e-07, two-sided Mann-Whitney U-test) and average years smoked (p=6.35e-36, two-sided Mann-Whitney U-test) were significantly associated with CHD at FWER < 0.05.

DISCUSSION

CHD was one of the leading global causes of death in 2023 (1). The correlations drawn here between the features and the target variable suggest that there is a robust relationship between behavior and the diagnosis of CHD. The high weights and effect sizes of the age and sex variables are consistent with existing literature regarding correlations between CHD and demographic variables, as older people and males are known to have higher risks of CHD (2). On the other hand, the high feature weights of height and sleeping time are not supported by existing literature, and their small effect sizes indicate that their significance could be a result of other correlations. For example, males are taller on average, so the correlation between taller individuals and a CHD diagnosis may be mediated by sex. The high feature weight of the average sleeping time variable may be related to the CHD diagnosis itself, as people with health conditions such as CHD might require more rest. Lastly, one final group of variables which demonstrated statistically significant yet unconfirmed associations are the alcohol related features, which contribute to the currently open debate on whether a moderate amount of alcohol may help lower the risk of cardiovascular disease (15, 16). These correlations demonstrate that, even if data from self-reported questions are not perfectly predictive of CHD, it may be possible to use survey questions to identify relationships between certain behavioral patterns and the presence of CHD.

Our models consistently performed better when predicting CHD in men compared to women. One possible explanation for this is that women are less likely to be diagnosed with CHD (2). This explanation was supported by the data, in which around 11.3% of men self-reported as having been diagnosed with CHD compared to just 7.1% of women. Fewer positive answers to the target question in the female demo-

https://doi.org/10.59720/24-006

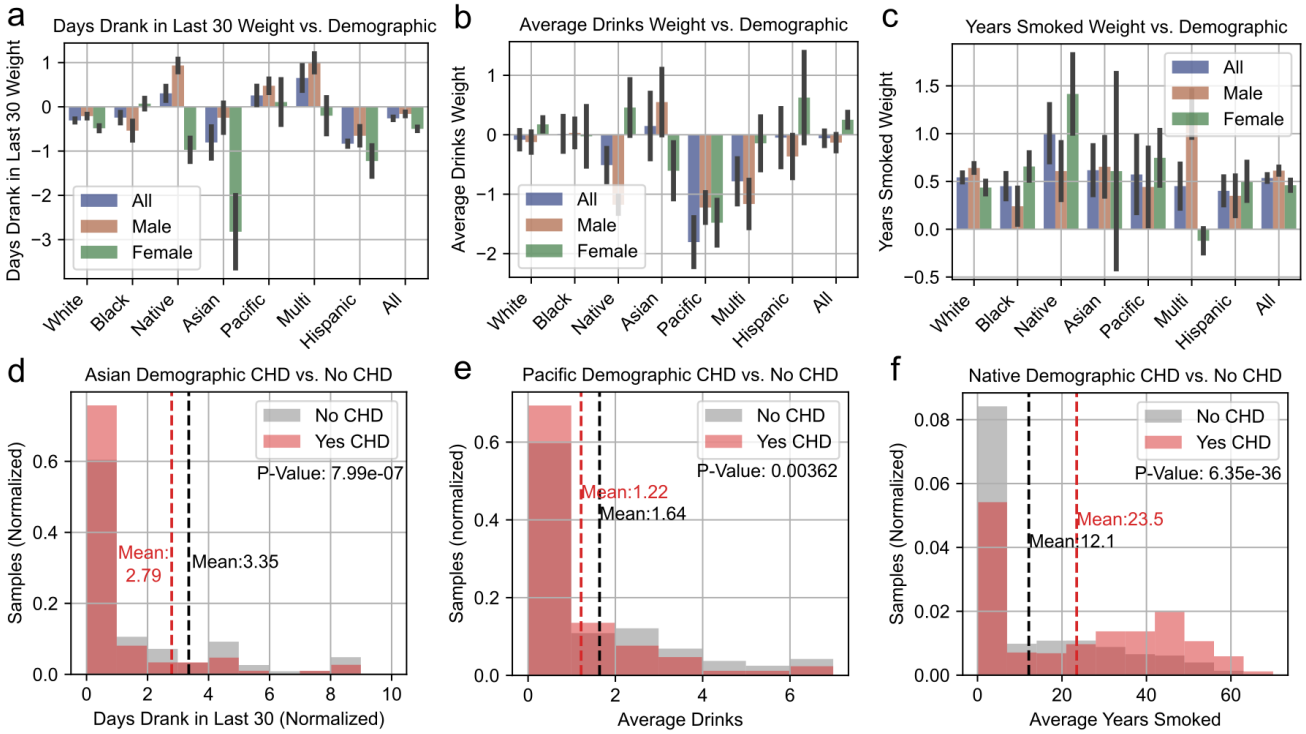

Figure 5: Features with variable weight across demographic-specific CHD models. a-c) Weights of (a) days drank in the last 30, (b) average drinks per sitting, and (c) years smoked, which each showed differing effects across different demographic-specific models. Logistic Regression was used as the model type for all cases. All feature weight values were computed over 5-fold cross validation. Bars represent means, and error bars represent standard deviations. d-f) Distribution of (d) days drank in the last 30, (e) average drinks per sitting, and (f) years smoked, in the Asian demographic, Pacific demographic, and Native demographic, respectively. P-values (two-sided Mann-Whitney U-test; n=9,577, 1,867, and 5,208, respectively) and means are displayed. Dotted lines represent means for each dataset (black: No CHD, red: Yes CHD).

graphic would have resulted in fewer positive cases on which to train the model, further aggravating the imbalance in the data. These results are supported by the evaluation metrics, which show that, although models trained on the female demographic had comparable accuracies to those of the male demographic, their MCCs were noticeably lower. This pattern is consistent with a model that is able to diagnose the disease-free majority correctly but struggles to accurately predict the few true positives.

Our analysis also encountered high variability across models when evaluating specific demographic groups. While the models performed well on the Native and Pacific demographics, they struggled with others such as Asian and Hispanic demographics. One explanation for this could be geographic diversity within each demographic: the worst performing models involved demographics which were inherently broad (such as Asians, which includes all of Asia), while the better performing ones involved more specific regions (such as Pacific or Native). Geographic diversity often correlates with genetic and environmental differences, which would have introduced extra variation that the model could not account for with provided data (such as East Asian vs. South Asian).

In addition to finding correlations between CHD and clinical, demographic, and behavioral patterns, this study also demonstrated a proof of concept that survey questions like those in the BRFSS are informative when discussing the potential diagnoses of specific, behavior-influenced diseases such as coronary heart disease. We used simple machine learning models to identify specific risk factors for individual demographics regarding heart disease when trained solely using broad questions taken from a study relating to overall health. The performance of these predictors could be enhanced by training them using additional CHD-related data or by testing further model types to maximize the performance with currently available data. Improvements such as these could lead to a more accurate predictor, potentially improving on existing first line of defense approaches and helping better identify and notify high-risk individuals, allowing them to see a doctor and get a more accurate diagnosis.

There were many limitations of this study. The data itself were a limitation, as there were relatively few behavior-related variables that could potentially relate to heart disease. The data were also heavily skewed towards white and elderly populations and thus may limit the generalizability of the findings to the broader US population and beyond. Additionally, there are likely issues regarding the reliability of self-reported survey responses that may have affected the reliability of the data and, as a result, the models it was used to train. For example, some individuals that had not been diagnosed with CHD may have nonetheless had the disease without knowing it. This would have resulted in both fewer positives for the model to train on as well as rendering some of the currently negative samples mislabeled. Another limitation arose due to issues of patient privacy in a public dataset with personally

https://doi.org/10.59720/24-006

identifiable information, such as the specific ethnicity or exact age of each sample. For example, some of the issues described above relating to geographic diversity may have been solved by providing the model with more meaningful ethnicity labels, such as dividing the world's largest population group (Asians) into multiple groups (South Asians, East Asians, etc.). In addition, the lack of exact age data introduced issues where the model considered two people aged 55 and 59 as the same age (placed in the 55-59 age group), whereas two people aged 59 and 60 were placed in separate categories and labeled as having different ages. Lastly, the scope of this research was limited by the availability of computational resources, which rendered the use of higher complexity models, such as neural networks, infeasible.

For future experiments, we recommend spending more time to either construct or find a large and specialized dataset. Although such a specially constructed dataset would probably be smaller than BRFSS, the ability to tailor questions specifically to the diagnosis of CHD would be invaluable. A more specific study asking questions more focused on heart disease could yield much better results while costing a potential patient much less than undergoing a CT scan or other traditional methods for diagnosing heart disease. For example, a specific area of questions that pertain heavily to CHD are diet-related questions, which were absent from the BRFSS questionnaire (14). Future studies could include many more behavioral questions, allowing for better assessment of the role of behavior in CHD. In addition, another aspect that could potentially be added to a future study is laboratory measurements that are widely accessible and can be taken outside of a clinical setting, such as blood pressure measurements. This type of data can potentially be collected at home and would grant a major piece of important information to the model while still costing less and remaining more convenient than more complex laboratory tests such as a CT scan.

The goal of this study was to demonstrate the promise of using survey-based prediction to diagnose CHD and recover correlations between behavioral factors and heart disease using solely large datasets of survey questions. This approach can not only be cheaper and easier to perform for researchers compared to running inconvenient clinical tests, but any developments or models successfully trained can then be made easily accessible to patients, only needing to answer the same survey questions that yield accurate models. This would allow medical professionals to determine which individuals need treatment the most, allowing them to give priority to patients at higher risk of CHD and potentially save numerous lives in the process. Hopefully, this study is seen by other AI/ML practitioners in the field and influences others to also adopt the useful and promising methods of survey-based prediction.

MATERIALS AND METHODS

Data

Data were obtained from the CDC's annual public BRFSS 2022 phone survey (14). The data were downloaded in SAS format, and the corresponding materials such as the codebook were used to interpret the data.

Preprocessing

All code written for this project was done using Python 3 and the Jupyter Notebook IDE, available at github.com/aidenc08/Diagnosing-CHD-With-Survey-Questions. The initial dataset, containing 445,132 samples and 326 features, was imported using the pandas module (18), and the pandas DataFrame object was used for the remainder of the project. Relevant variables were identified by manually flagging any variables that could be classified as behavioral patterns (smoking, alcohol use, exercise, sleep time), previous health conditions (kidney disease, diabetes, COPD, asthma, oral health), or demographic data (race, age, gender). After identifying the relevant variables, those variables were prepared for model training through five preprocessing steps. The first step was classifying each variable as either categorical or continuous. All binary variables were considered categorical and stored in one variable with 0 indicating no presence and 1 indicating presence. The next step was expressing the categorical variables with multiple binary variables each expressing a single possible category of the variable (one-hot encoding). In addition, one continuous variable, when a person had first started smoking, was one-hot encoded into different categories including never smoked, first cigarette before 18, and first cigarette after 18. The third step was preprocessing continuous variables. Questions left blank because of negative answers to prior questions (e.g. "What is the most number of drinks you've ever had?" asked to a person who had never drank) had their blank entries turned into zeroes. Values indicating missing or invalid responses (frequently either two or three-digit numbers composed of only 7s or 9s) were replaced with the numpy module's (19) Not a Number variable to make them easier to identify. One variable, which indicated either the number of times a person had drank in the last week or in the last 30 days, needed additional preprocessing. Each value that represented the last week was divided by 7 and multiplied by 30 to reflect 30-day values. Finally, technically categorical variables with many categories that reflected a continuous nature were kept continuous, such as an age variable that classified people into 13 different, 5-year age groups. The fourth step was linearly scaling the continuous variables to range from 0 to 1, which was done using scikit-learn's (20) MinMaxScaler. Prior to this, if a variable did not seem to be normally distributed (evaluated by looking at a representative matplotlib histogram), the variable was log transformed using a natural log base. If the variable had values of zero, one was added to each value before taking the natural log. The final step was to combine all of the preprocessed data into one dataframe and drop samples with a Not a Number value for any variable. After these steps, a final dataframe with 340,200 samples and 32 features remained. The target variable was obtained by taking a binary variable that indicated the self-reported past diagnosis of a myocardial infarction (heart attack) or CHD in a sample and reindexing the original target array using the preprocessed dataframe.

Model Training

For model training, multiple different binary classification models from the scikit-learn module were considered (20). These models were Support Vector Machine (LinearSVC), Naive Bayes, Logistic Regression, and Random Forest Classifier. Naive Bayes was run with default parameters while LinearSVC and LogisticRegression were run with a 'balanced' class_weight and a max_iter of 10000. RandomForestClassifier was also run with a 'balanced' class_weight and with a min_samples_leaf of 125. Training and test sets were gen-

https://doi.org/10.59720/24-006

erated using 5-fold stratified cross validation (equal ratio of positive to negative cases in all folds) before being used to fit and evaluate different model types. Model performance was assessed by averaging the Matthews Correlation Coefficient on both training and test datasets across all 5 folds. Logistic Regression with a "balanced" class-weight and a max iterations of 10000 was determined to be the best performing model by mean test set MCC. Cross validation experiments were conducted on 90% of the data, with the remaining 10% used as a holdout dataset to evaluate the final model choice. This model choice was then applied to secondary datasets generated by splitting based on the demographic of variables race and sex.

Interpretation of Models

Interpretation of models was done using the .coef_ field to determine the weights of each feature passed into the model. The most significant features were found by taking the absolute value of each weight, and determining which ones had the highest average weight over all five folds from 5-fold cross validation. The same was done for the secondary dataset generated for race and sex demographics. The weights of each variable across all demographics were then analyzed to determine mean weight, standard deviation, and max weight in order to assess how impactful and how consistent each variable was across all demographics. Variables with a high standard deviation or a large discrepancy between the mean weight and the max weight were then further analyzed to identify demographics with outlier results. Variables with the highest and lowest mean weights, as well as those with the highest standard deviations and discrepancies between mean weight and max weight, were further analyzed through statistical tests. Associations with CHD were determined using two-sided Fisher's Exact Test for binary variables and the two-sided Mann-Whitney U-test for continuous ones. The Bonferroni correction for multiple hypothesis testing was used with a total of 55 tests run to obtain a threshold of around 0.0009 for statistical significance (FWER < 0.05). Effect size was also assessed using a log10 odds ratio for binary variables and difference in means for continuous ones. Both p-values and effect size were considered when determining whether a variable's association with CHD was both statistically significant and practically meaningful.

Received: December 22, 2023
Accepted: April 09, 2024
Published: August 23, 2024

REFERENCES

1. "Heart Disease Facts." *Centers for Disease Control and Prevention*, www.cdc.gov/heart-disease/data-research/facts-stats/index.html. Accessed 16 Dec. 2023.
2. "What Is Coronary Heart Disease?" *NHLBI, NIH*, www.nhlbi.nih.gov/health/coronary-heart-disease. Accessed 15 Dec. 2023.
3. *Coronary Heart Disease - Illnesses and Conditions | NHS INFORM*, www.nhsinform.scot/illnesses-and-conditions/heart-and-blood-vessels/conditions/coronary-heart-disease. Accessed 16 Dec. 2023.
4. "Symptoms." *NHLBI, NIH*, www.nhlbi.nih.gov/health/coronary-heart-disease/symptoms. Accessed 15 Dec. 2023.
5. Ojha, Niranjan. and Amit S. Dhamoon. "Myocardial Infarction." *StatPearls*, StatPearls Publishing, 8 August 2023.
6. *NHS Choices*, NHS, www.nhs.uk/conditions/coronary-heart-disease/. Accessed 17 Dec. 2023.
7. Javed, Z., et al. "Race, Racism, and Cardiovascular Health: Applying a Social Determinants of Health Framework to Racial/Ethnic Disparities in Cardiovascular Disease." *Circulation. Cardiovascular Quality and Outcomes*, vol. 15, no. 1, Jan. 2022, https://doi.org/10.1161/CIRCOUTCOMES.121.007917.
8. Nakanishi, Rine, et al. "Machine Learning Adds to Clinical and CAC Assessments in Predicting 10-Year CHD and CVD Deaths." *JACC: Cardiovascular Imaging*, vol. 14, no. 3, Mar. 2021, pp. 615–625, https://doi.org/10.1016/j.jcmg.2020.08.024.
9. Javaid, Aamir, et al. "Medicine 2032: The Future of Cardiovascular Disease Prevention with Machine Learning and Digital Health Technology." *American Journal of Preventive Cardiology*, vol. 12, Dec. 2022, https://doi.org/10.1016/j.ajpc.2022.100379.
10. Özbilgin, Ferdi, et al. "Prediction of Coronary Artery Disease Using Machine Learning Techniques with Iris Analysis." *Diagnostics*, vol. 13, no. 6, Mar. 2023, p. 1081, https://doi.org/10.3390/diagnostics13061081.
11. Hassan, Ch Anwar ul, et al. "Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers." *Sensors*, vol. 22, no. 19, Oct. 2022, https://doi.org/10.3390/s22197227.
12. Sayadi, Mohammadjavad, et al. "A Machine Learning Model for Detection of Coronary Artery Disease Using Noninvasive Clinical Parameters." *Life*, vol. 12, no. 11, Nov. 2022, https://doi.org/10.3390/life12111933.
13. Srinivasan, Saravanan, et al. "An Active Learning Machine Technique Based Prediction of Cardiovascular Heart Disease from UCI-Repository Database." *Scientific Reports*, vol. 13, no. 1, 21 Aug. 2023, https://doi.org/10.1038/s41598-023-40717-1.
14. "CDC - 2022 BRFSS Survey Data and Documentation." *Centers for Disease Control and Prevention (CDC)*, 29 Aug. 2023, www.cdc.gov/brfss/annual_data/annual_2022.html. Accessed 10 Oct. 2023.
15. "Heart Disease - Symptoms and Causes." *Mayo Clinic*, 2018, www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118. Accessed 14 Dec. 2023.
16. "Alcohol and Heart Health: Separating Fact from Fiction." *John Hopkins Medicine*, 2023, www.hopkinsmedicine.org/health/wellness-and-prevention/alcohol-and-heart-health-separating-fact-from-fiction. Accessed Dec. 15 2023.
17. Kromhout, Daan, et al. "Prevention of Coronary Heart Disease by Diet and Lifestyle." *Circulation*, Feb. 2002, https://doi.org/10.1161/hc0702.103728.
18. The pandas development team. "Pandas-Dev/pandas: Pandas." *Zenodo*. 2023. https://doi.org/10.5281/zenodo.10304236.
19. Harris, Charles R., et al. "Array Programming with NumPy." *Nature*, vol. 585, no. 7825, 16 Sept. 2020, pp. 357–362, https://doi.org/10.1038/s41586-020-2649-2.
20. Pedregosa, Fabian, and Alexandre Gramfort. "Scikit-

https://doi.org/10.59720/24-006

Learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, no. 85, 2011, pp. 2825–2830, jmlr.org/papers/v12/pedregosa11a.html. Accessed 20 Dec. 2023.