Data-Driven Discovery of Risk Factors in Cardiovascular Disease: The Role of Demographics, Socioeconomic Status, and Lifestyle Factors

Benjamin Luo
Department of Management Science and Engineering
University of Waterloo
Waterloo, Canada
b33luo@uwaterloo.ca

***Abstract*—This paper performs a data analysis on the Behavioral Risk Factor Surveillance System 2022 dataset to investigate hidden patterns in cardiovascular disease, with a focus on demographic, socioeconomic, and lifestyle factors. Multiple data analyses were exercised to showcase both conventional and machine learning methods on big data, including their limitations.**

The findings generally reinforce clinical knowledge on cardiovascular disease risk factors. In particular - age, health coverage, and self-assessed physical health were found to be the most important factors from the 3 respective categories, with an emphasis on age past 60 years.

Keywords—Cardiovascular disease, machine learning, big data analytics

I. Introduction

Cardiovascular diseases (CVD) are the leading cause of death worldwide; the World Health Organization (WHO) estimated that in 2022, one death occurred every 33 seconds for a total of 702,880[1]. CVDs are characterized as diseases affecting the heart and blood vessels, including coronary heart disease, strokes, peripheral arterial disease, and aortic disease[2]. According to the National Health Services (NHS), key risk factors include high blood pressure, smoking, high cholesterol, diabetes, kidney disease, inactivity, and obesity[2]. There are also notable differences observed across biological sex and ethnic groups which may be attributed to genetic predisposition[2].

By studying health implications, especially through large scale datasets, research can identify hidden patterns in CVD prevalence. These risk factors can inform governmental policies such as enforcing stricter health and safety standards for select industries. Understanding demographic factors can also provide insight into vulnerable populations in clinical settings so that physicians can tailor their treatment approaches and provide targeted interventions, ultimately reducing health disparities and improving CVD outcomes in high-risk groups. Finally, examining behavioral and lifestyle factors has the potential to help researchers develop evidence-based recommendations for public health campaigns and preventive measures which encourages individuals to adopt healthy habits and reduce their risk of developing CVD. Overall, this multifaceted approach can lead to more effective prevention and management strategies for CVD.

One challenge in working with big data is that selecting the most suitable analytical approach can be daunting, particularly when dealing with complex datasets like the Behavioral Risk Factor Surveillance System (BRFSS). To address this challenge, employing a combination of traditional statistical methods and machine learning techniques can provide a more comprehensive understanding of the data. Conventional statistics can offer insights into population-level trends, including correlations, while machine learning algorithms can uncover hidden patterns and nonlinear relationships. For instance, logistic regression can be used to estimate the odds of developing CVD based on demographic and behavioral factors, while machine learning algorithms like random forests or neural networks can be employed to predict individual-level risk and identify complex interactions between variables.

This investigation is driven by a threefold purpose: firstly, to pinpoint vulnerable populations within healthcare settings, secondly, to inform policies that foster health equity and improve access to care, and thirdly, to encourage individuals to adopt healthier lifestyles. The following research questions will guide the exploration:

RQ1) What is the prevalence of CVD across various age, sex, and ethnic groups?

RQ2) Do socioeconomic factors such as income, education, and insurance status significantly impact CVD prevalence?

RQ3) How do behavioral factors like smoking, alcohol consumption, and physical activity influence CVD prevalence?

II. Methodology

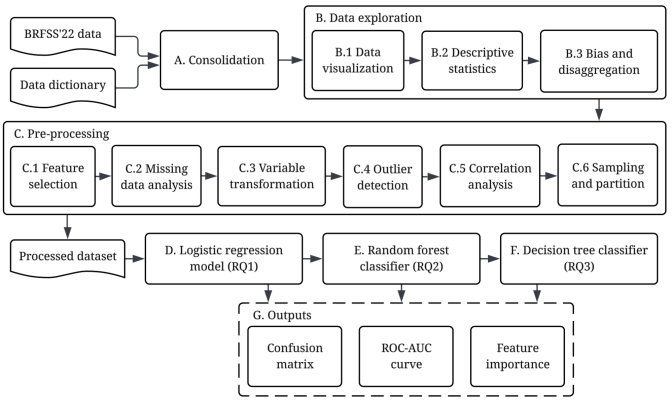


Fig. 1. Overview of the methodology from preprocessing to analysis

Dataset

The Behavioral Risk Factors Surveillance System (BRFSS) is an annual telephone survey conducted across the United State of America (US) by the US Centers for Disease Control and Prevention (CDC). The 2022 dataset[3] was used for this analysis, and contains 445,132 responses with 328 questions. These questions include a wide range of demographic, socioeconomic, and lifestyle factors, in addition to self-declared diagnoses. Due to the poor disaggregation of

CVDs, the analysis operationalized the dependent variable, CVD, as the presence of angina, coronary heart disease (CHD), and heart attack.

All responses were considered by default, but pre-processing pruned the irrelevant questions (C.1) and responses (C.2) to approximately 130k datapoints and 36 features, including control variables. Figure 1 overviews the high-level methodology for exploring, processing, and analyzing the data.

A. Consolidation

By default, the BRFSS'22 has encoded values where each value is assigned a numeric code and can be decoded through the provided data dictionary. Feature names are also encoded and can be similarly decoded. Appendix S1 provides a glimpse into the BRFSS'22 dataset. The majority of features were in fact categorical, but not necessarily Boolean as respondents had the option of responding "Don't know" or directly refusing to respond.

Consolidation refers to the process by which the data dictionary was applied to decode values in the raw dataset. The purpose of this was to make data more interpretable during exploration.

B. Data exploration

Data exploration was initially done using visualizations at the univariate, bivariate, and multivariate levels. Univariate analysis through boxplots captured distributions in key variables such as age and BMI, which informed appropriate statistical testing. Bivariate and multivariate visualizations attempted to identify correlations between data for potential further investigation.

Descriptive statistics were used to identify means and variances in the data to get a rough estimate of potential disparities between groups in terms of CVD prevalence. For example, appendix S2 shows disaggregation done for demographic risk factors, which revealed disparities across biological sex, age, race/ethnicity, and marital status. This revealed a clear trend between age and CVD prevalence, but it is insufficient to establish statistical significance and effect size.

Finally, disaggregation is done to identify any imbalances in the dataset which may introduce bias into the models. For example, disaggregation with CVD revealed that only ~9.32% of respondents reported being affected by CVD. Similarly, the majority (~73.68%) of respondents identified as being racially white (Appendix S3). Although these proportions are similar to the expected prevalence and demographics in the US, they may introduce bias to the model so careful preprocessing is required (C.6 Sampling and partition).

C. Pre-processing

Pre-processing attempted to clean the data in preparation for analysis, with a focus on pruning unnecessary data to avoid the issue of garbage-in-garbage-out. Feature selection identified relevant features for each of the research focuses (i.e. demographics, socioeconomics, and behavioral/lifestyle factors), as well as the dependent variables (CVDs) and control variables (physical conditions). A LLM was prompted to categorize the list of 328 features into each of the above categories, and results were manually validated. This resulted in the selection of 36 features.

Next, missing data analysis first removed questions where over 80% of responses were missing or refused, and then datapoints with any missing or refused responses. This step is crucial for certain statistical models that fail in the presence of NaN values. Mean/median imputation was considered for interpolating responses, but the Boolean nature of the responses may have exaggerated certain responses and the final dataset size was reasonable at ~272k datapoints (approximately half of the original dataset). Note that 'Other' responses (ex. for race/ethnicity) were not removed.

Univariate data analysis was repeated to confirm the overall distributions and trends have not drastically changed after removing incomplete data points; thus reducing the risks of missing not at random (MNAR), especially for the questions that were often refused or answered as 'Unsure'. Afterward, variables (ex. sleep hours, alcohol consumption, weight, height, BMI) were transformed from strings to numbers so that outlier detection could be performed via histograms. This step also involved scaling the data to a consistent time measure as certain questions asked about lifestyle habits on the weekly or monthly basis (ex. During the past 30 days, how many days *per week or per month* did you have at least one drink of any alcoholic beverage?). In this case, the response encoding can be seen in appendix 4.

Outliers were then observed using histograms but no outliers were removed. There were very few outliers due to the initial binning inherent to the BRFSS (ex. income $200k+).

Correlation analysis was performed within each of the demographic, socioeconomic, and behavioral/lifestyle factors to remove highly correlated data. Spearman's correlation was used on numeric data as most values were not normally distributed. Cramer's V is used to study correlations for categorical variables. Inter-class correlations were neglected, such as age and marital status, but were controlled for during C.6 (Sampling and partition) to avoid confounding variables. The final variable selection is given in Table I.

TABLE I
SELECTED VARIABLES BY TYPE

Variable	Category	Type
Cardiovascular disease	Dependent	Binary
General Health Self-Assessment	Control	Categorical
Diabetes		Binary
Chronic Obstructive Pulmonary Disease		Binary
Asthma		Binary
State	Demographic	Categorical
Sex		Binary
Age		Categorical
Race/Ethnicity		Categorical
BMI		Integer
Education	Socioeconomic	Categorical
Income		Categorical
Marital Status		Categorical
Employment		Categorical
Health Insurance		Categorical
Affordable acesss to care		Binary
Tobacco Use	Behavioral/Lifestyle	Binary
E-Cigarette Use		Binary
Alcohol Consumption		Integer
Physical Health Self-Assessment		Integer
Mental Health Self-Assessment		Integer
Sleep		Integer
Depression		Binary
Checkup Frequency		Integer

After this, numerical data was normalized to the [0,1] range for more fair comparison against binary variables. This was primarily done for the logistic regression model so that

feature importance could be inferred from the coefficients, but is not strictly necessary for decision trees.

Finally, sampling was done with respect to potential confounding variables from other classes (notably age) prior to splitting the dataset into train and test (using a 70:30 ratio). Control variables (ex. diabetes) were monitored to ensure the split datasets were comparable. It is worth noting that explicit rebalancing of the dataset in terms of CVD and non-CVD datapoints is not necessary as the library used to implement decision trees and other models vulnerable to imbalanced datasets have built-in support for rebalancing.

D. Logistic regression model (RQ1)

A logistic regression model was used to model demographic risk factors and CVD prevalence. This approach is highly flexible and accommodates for both categorical and numeric data. Furthermore, the output is a probability between [0,1] which is well suited for this binary classification problem. The demographic features in Table 1 were fit to a logistic regression model in scikit-learn with the requirement that variables were at least at the $p < 0.05$ significance threshold. Hyperparameters (regularization method, iterations) were experimented via grid search to identify suitable parameters.

Model evaluation was done on the test dataset and will be discussed further in section G (*Outputs and evaluation*). The logistic regression model in particular will be visualized through an odds-ratio plot to report the strength and direction of associations between demographic risk factors and CVD prevalence.

E. Random forest classifier (RQ2)

A random forest classifier was fit on the socioeconomic risk factors for CVD. This method is highly flexible, captures complicated relationships between variables, and avoids overfitting through the use of an ensemble of trees. It is also suitable for binary classification. Random forest classifiers output feature importance based on lower entropy, which measures the reduction in uncertainty when a feature is used to split a tree node. Lower entropy indicates the feature is able to effectively separate the classes and subsequently has high explanatory power.

Random forest classifiers are vulnerable to imbalanced datasets so it was important to experiment with the hyperparameters, including the rebalancing ratio in scikit-learn.

F. Decision tree classifier (RQ3)

A decision tree classifier was used to model the effect of behavioral and lifestyle factors on CVD prevalence. This method was selected as it is a very interpretable model which aligns with the goal of encouraging everyday people to make healthier lifestyle choices. This model carries many of the same benefits as random forest classifiers, including its robustness to multiple data types and suitability to binary classification problems. However, it may overfit as it is not an ensemble method, so careful hyperparameter tuning is required, including the max depth, samples per leaf, impurity decrease, and class weights in scikit-learn.

G. Outputs and evaluation

Models were trained on the training set (n=291082) and evaluated on the test set (n=72771). Model evaluation was done qualitatively via ROC curves, and quantitatively via AUC, precision, recall, and accuracy. Feature importance was ranked and reported for all 3 approaches (*D, E, F*). k-folds cross validation was considered but results were comparable to directly using train and test sets.

III. Results

Processed dataset

Of the original 445,132 responses and 328 features, the final dataset contained 271,802 datapoints and 24 features (Table I). 81,279 datapoints were removed either due to missing data, refusal to answer, or 'Unsure' responses. 36 features were initially identified but were reduced to 24 features through aggregation (notably BMI and CVD) and removal due to high correlation. The train-test split was done using a 70:30 ratio which resulted in 291,082 training datapoints and 72,771 testing datapoints; controlled for age and the physical conditions listed in Table I.

Data exploration

A Binomial-GLM was directly fitted on all explanatory and control variables to explore whether they were significant ($p < 0.05$) to CVD. Each of the categorical variables were one-hot-encoded which resulted in 126 comparisons. Of these, about 58 were statistically significant and are reported in appendix S6. Some groups that were not significantly associated with CVD are: adults aged 25-29, most states of residence (41/50), and married or separated individuals.

Demographics and CVD prevalence (RQ1)

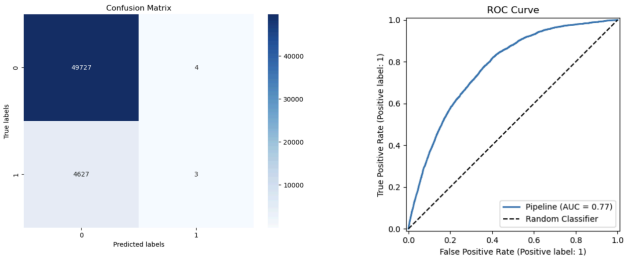


Fig. 2. Confusion matrix and ROC curve for the logistic regression model trained on demographic risk factors

The logistic regression model fits well to the data based on the high quantitative measures seen in table II and qualitative results in figure 2. However, many of the metrics used in the logistic regression did not meet the $p < 0.05$ significance threshold. These were primarily the one-hot encoded state names partially seen in figure 3, and fully described in appendix S7.

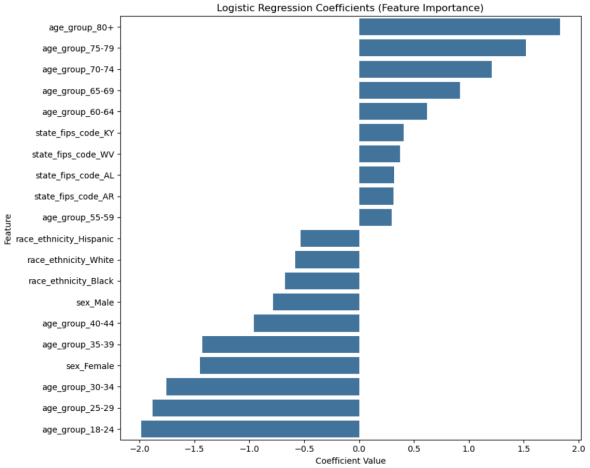


Fig. 3. Feature importance rankings according to coefficients of the logistic regression model

TABLE II
SUMMARY OF MODEL EVALUATIONS

Model	AUC	Precision	Recall	Accuracy
Logistic regression (RQ1)	0.77	0.91	1.00	0.91
Random forest classifier (RQ2)	0.74	0.97	0.60	0.62
Decision tree classifier (RQ3)	0.69	0.87	0.87	0.87

Figure 3 depicts the feature importance rankings based on the coefficients of the logistic regression model. The most important demographic factor is age where older individuals are more susceptible to CVD. Refer to appendix S9 for the odds-ratio plot which reinforces figure 3, but also depicts the sensitivity of risk factors, or in other words, how the odds change when a numerical risk factor is incremented by 1 unit (or toggled, if it is a binary variable).

Descriptive statistics were used in appendix 2 to investigate intra-group relations and this revealed that on average, males were more prone to CVD. A chi-squared test confirms significance ($p < 0.001$) but the low ranking on feature importance (figure 3) indicates a low effect size of sex on CVD prevalence.

Socioeconomic status and CVD prevalence (RQ2)

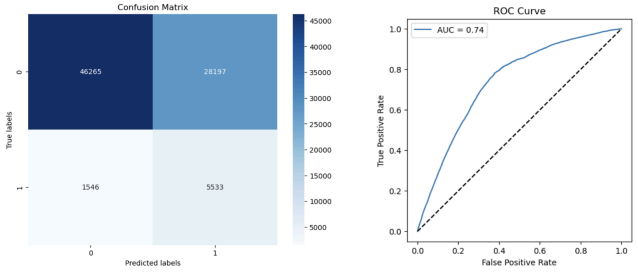


Fig. 4. Confusion matrix and ROC curve for the random forest classifier trained on socioeconomic risk factors

The random forest classifier with hyperparameters n_estimators=200, class_weight="balanced", max_depth=8, min_samples_split=5, and random_state=42 appeared to overfit to the CVD positive label as there were many false positives when ran on the test set (Figure 4). Interestingly, rebalancing the dataset had very little impact on the confusion matrix and corresponding metrics.

The most important features were employment status and primary health insurance (figure 5). However, employment status is heavily correlated with age due to the 'Retired' class (and similarly, marital status is correlated with age due to the 'Widowed' and 'Divorced' classes). The full enumeration of possible values are given in appendix S5. Instead of explicitly controlling for age, further investigation was done by disaggregating both employment status (figure 6) and primary health insurance (figure 7).

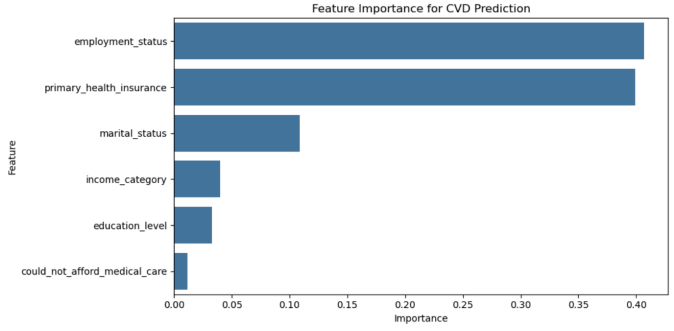


Fig. 5. Feature importance rankings according to the random forest classifier

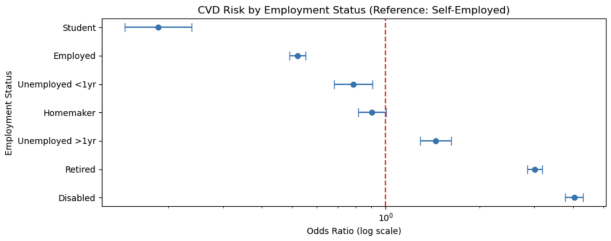


Fig. 6. Odds-ratio plot on the disaggregation of employment status versus cardiovascular disease prevalence (the reference group is self-employed)

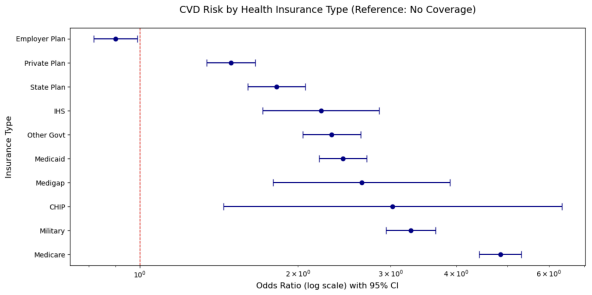


Fig. 7. Odds-ratio plot on the disaggregation of health insurance versus cardiovascular disease prevalence (the reference group is no coverage)

Figures 6 and 7 depict the results of the disaggregated exploration of how employment status and health insurer affect CVD prevalence. Logistic regression models were fit to both explanatory variables and all one-hot-encoded classes were found to be statistically significant at the $p < 0.03$ level, except for the homemaker occupation at $p = 0.053$. Refer to appendices S9 and S10 for the statistical significance of each class and model parameters.

Lifestyle factors and CVD prevalence (RQ3)

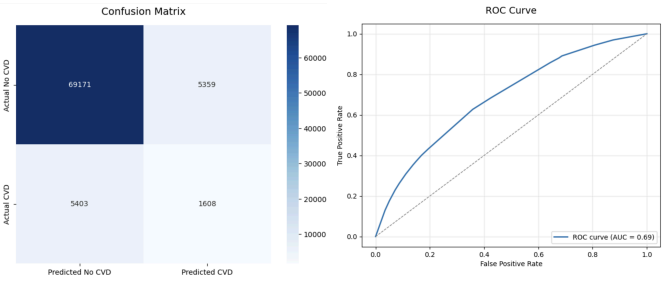


Fig. 8. Confusion matrix and ROC curve for the decision tree classifier trained on behavioral and lifestyle risk factors

The decision tree classifier performed very well in terms of the confusion matrix (figure 8), but the AUC score is relatively the lowest at about 68% (Table II). Together, these indicate the classifier is overfitting to the CVD negative label, potentially due to a class imbalance. The final proportion of CVD positive predictions was roughly 6.9%, as opposed to the 9.1% prevalence seen in the overall dataset. Rebalancing was done with respect to the CVD labels by altering the class weights, along with hyperparameter tuning via grid search.

The most important risk factors are given in Table III according to their relative importance. A score of '0' indicates the feature was not used by the decision tree - reasons for this will be discussed further in the discussion section. The final hyperparameters were: max depth = 5, min impurity decrease = 0.00001, min leaf sample = 0.01. Refer to appendix S10 for a visualization of the decision tree.

TABLE III
BEHAVIORAL/LIFESTYLE RISK FACTOR IMPORTANCE

Feature	Importance
physical_health_not_good_days	**0.617**
last_doctor_visit_<1yr	**0.224**
days_drinking_alcohol	**0.109**
mental_health_not_good_days	**0.051**
uses_tobacco_Never	0.000
average_sleep_hours	0.000
last_doctor_visit_Never	0.000
last_doctor_visit_5+ yrs	0.000
last_doctor_visit_2-5yrs	0.000
uses_tobacco_Daily	0.000
depression_diagnosis_Yes	0.000
depression_diagnosis_No	0.000
uses_e_cigarettes_Occasionally	0.000
uses_e_cigarettes_Never	0.000
uses_e_cigarettes_Former	0.000
uses_e_cigarettes_Daily	0.000
uses_tobacco_Occasionally	0.000
last_doctor_visit_1-2yrs	0.000

IV. DISCUSSION

This study investigated risk factors of cardiovascular disease with respect to demographics, socioeconomics, and behavioral/lifestyle factors (Table I). Various statistical and machine learning tools were employed to capture associations and identify the most important risk factors from each group. Quantitative assessment of each model is reported in Table II.

Demographics and CVD prevalence (RQ1)

RQ1) What is the prevalence of CVD across various age, sex, and ethnic groups?

Age is the most important demographic risk factor with older individuals, especially past 60 years of age. There were statistical differences in CVD prevalence between biological sex and racial/ethnic groups, but the effect was negligible compared to age.

Discuss, compare against clinical knowledge

Socioeconomic status and CVD prevalence (RQ2)

RQ2) Do socioeconomic factors such as income, education, and insurance status significantly impact CVD prevalence?

Employment and health coverage were the most important socioeconomic factors. Disaggregation revealed that retired and disabled workers are the most vulnerable, which reinforces the association between age and CVD prevalence. The effect of health coverage varied between providers, but in general, public health coverage such as Medicare yields higher CVD prevalence relative to private or employer plans.

Discuss, compare against clinical knowledge

Lifestyle factors and CVD prevalence (RQ3)

RQ3) How do behavioral factors like smoking, alcohol consumption, and physical activity influence CVD prevalence?

Self-assessed physical health is the most important behavioral/lifestyle risk factor, followed by the frequency of doctor visits, alcohol consumption, and self-assessed mental health. Findings were inconclusive on all other risk factors.

Discuss, compare against clinical knowledge

Limitations

Limitations but why results still valid

Future work

Future work (optional)

V. CONCLUSIONS

Overall, this study found that age presents as the highest risk factor for CVD, with individuals above the age of 60 being especially vulnerable. Based on this observation, medical practitioners should be more alert when addressing the healthcare needs of older patients who may be more susceptible to cardiovascular complications. Furthermore, this study found discrepancies between health coverage providers and CVD prevalence, with public programs such as Medicare being associated with increased CVD prevalence compared to private and employer-provided plans. This exposes socioeconomic disparities in healthcare and an opportunity for the government to further develop the public health system. Finally, self-assessed physical health is the most prominent variable of the behavioral/lifestyle risk factors - amongst a low frequency of doctor visits, increased alcohol consumption, and a low self-assessment of mental health. From this, people are encouraged to adopt more active lifestyle habits and reduce alcohol consumption in order to improve their cardiovascular health.

REFERENCES

[1] June 2021. Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). World Health Organization (WHO) Heart Disease Facts.

[2] April 2022. Cardiovascular disease. https://www.nhs.uk/conditions/cardiovascular-disease/. National Health Service (NHS)

[3] 2022 BRFSS Survey Data and Documentation. https://www.cdc.gov/brfss/annual_data/annual_2022.html. US Centers for Disease Control and Prevention (CDC)

SUPPLEMENTARY MATERIALS

Appendix S1: Sample excerpt from the Behavioral Risk Factors and Surveillance System (BRFSS) 2022 dataset. Left is the encoded raw data, and right is the corresponding data dictionary that includes the

	CVDINFR4	CVDCRHD4	_MICHD	_STATE	SEXVAR	_AGEG5YR	_RACEGR4	_EDUCAG	_INCOMG1
0	2.0	2.0	2.0	1.0	2.0	13.0	1.0	4.0	9.0
1	2.0	2.0	2.0	1.0	2.0	13.0	1.0	2.0	3.0
2	2.0	2.0	2.0	1.0	2.0	8.0	1.0	4.0	6.0
3	2.0	2.0	2.0	1.0	2.0	14.0	1.0	2.0	9.0
4	2.0	2.0	2.0	1.0	2.0	5.0	1.0	3.0	3.0
5	1.0	2.0	1.0	1.0	1.0	13.0	1.0	2.0	9.0
6	2.0	2.0	2.0	1.0	2.0	13.0	2.0	4.0	5.0
7	2.0	2.0	2.0	1.0	2.0	13.0	1.0	2.0	5.0
8	2.0	2.0	2.0	1.0	2.0	12.0	1.0	4.0	5.0
9	2.0	2.0	2.0	1.0	2.0	11.0	1.0	4.0	5.0
10	2.0	2.0	2.0	1.0	2.0	13.0	1.0	1.0	3.0
11	2.0	2.0	2.0	1.0	2.0	13.0	1.0	4.0	9.0
12	2.0	2.0	2.0	1.0	1.0	8.0	2.0	2.0	5.0
13	2.0	2.0	2.0	1.0	1.0	10.0	1.0	4.0	6.0

Label: Ever Diagnosed with Heart Attack
Section Name: Chronic Health Conditions
Core Section Number: 7
Question Number: 1
Column: 118
Type of Variable: Num
SAS Variable Name: CVDINFR4
Question Prologue:
Question: (Ever told) you had a heart attack, also called a myocardial infarction?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	25,108	5.44	4.38
2	No	416,959	93.67	94.83
7	Don't know/Not sure	2,731	0.61	0.69
9	Refused	330	0.07	0.09
BLANK	Not asked or Missing	4	.	.

Appendix S2: Descriptive statistics on demographic groups based on cardiovascular (CVD) prevalence. The numbers indicate mean measurements across the entire dataset.

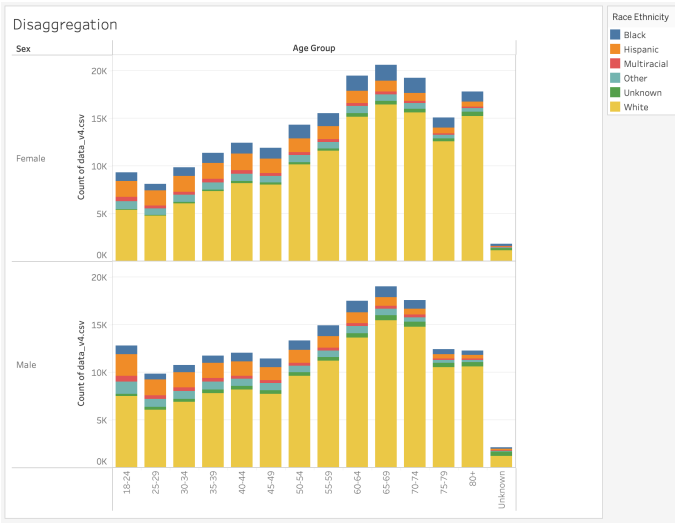
```
CVD prevalence by age_group:
age_group
18-24    0.006358
25-29    0.008385
30-34    0.011270
35-39    0.016004
40-44    0.023129
45-49    0.037013
50-54    0.054162
55-59    0.076258
60-64    0.097624
65-69    0.124387
70-74    0.156597
75-79    0.190166
80+      0.228070
```

```
CVD prevalence by sex:
sex
Female    0.07035
Male      0.11021

CVD prevalence by race_ethnicity:
race_ethnicity
Hispanic        0.056558
Other           0.065571
Black           0.075453
Multiracial     0.088730
White           0.096615
```

```
CVD prevalence by marital_status:
marital_status
Single              0.039784
Unmarried Partner   0.045047
Married             0.086197
Separated           0.097656
Divorced            0.114838
Widowed             0.177178
```

Appendix 3: Disaggregation done on age, biological sex, and race/ethnicity revealed a high imbalance toward respondents identifying as white (Tableau Public).



Appendix S4: Behavioral Risk Factors and Surveillance System (BRFSS) encoding for the question "During the past 30 days, how many days per week or per month did you have at least one drink of any alcoholic beverage?".

- 101-199: Days per week
- 201-299: Days in past 30 days
- 777: Don't know/Not sure *(removed during C.2)*
- 888: No drinks
- 999: Refused *(removed during C.2)*
- BLANK: Not asked or Missing *(removed during C.2)*

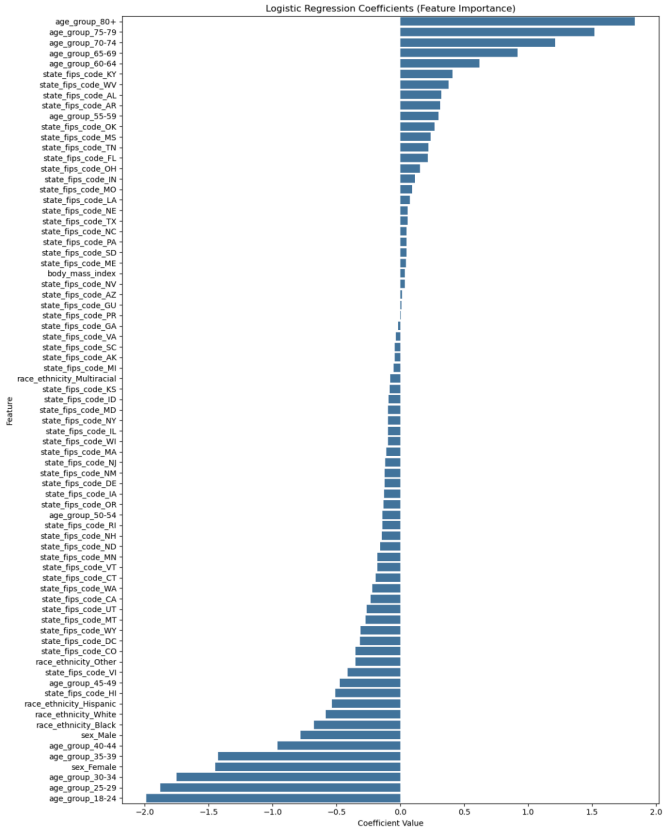
Appendix S5: Behavioral Risk Factors and Surveillance System (BRFSS) questionnaire and unique values for the selected features

Variable	Unique Values	Questionnaire
cvd	[0, 1]	-
state_fips_code	['AL', 'AK', 'AZ', 'AR', 'CA', 'CO', 'CT', 'DE', 'DC', 'FL', 'GA', 'HI', 'ID', 'IL', 'IN', 'IA', 'KS', 'KY', 'LA', 'ME', 'MD', 'MA', 'MI', 'MN', 'MS', 'MO', 'MT', 'NE', 'NV', 'NH', 'NJ', 'NM', 'NY', 'NC', 'ND', 'OH', 'OK', 'OR', 'PA', 'RI', 'SC', 'SD', 'TN', 'TX', 'UT', 'VT', 'VA', 'WA', 'WV', 'WI', 'WY', 'GU', 'PR',	State FIPS Code
sex	['Female', 'Male']	Sex of Respondent
age_group	['80+', '55-59', '40-44', '70-74', '65-69', '60-64', '75-79', '50-54', '45-49', '35-39', '30-34', '25-29', '18-24']	Fourteen-level age category
race_ethnicity	['White', 'Black', 'Multiracial', 'Other', 'Hispanic']	Five-level race/ethnicity category
education_level	['HS Grad', 'College Grad', 'Some College', 'No HS']	Level of education completed
income_category	['$25K-35K', '$100K-200K', '$50K-100K', '$35K-50K', '<$15K', '$15K-25K', '$200K+']	Income categories
marital_status	['Widowed', 'Married', 'Divorced', 'Separated', 'Unmarried Partner', 'Single']	Are you: (marital status)
employment_status	['Self-Employed', 'Retired', 'Homemaker', 'Disabled', 'Employed', 'Unemployed <1yr', 'Unemployed >1yr', 'Student']	Are you currently…?
uses_tobacco	['Never', 'Occasionally', 'Daily']	Do you currently use chewing tobacco, snuff, or snus every day, some days, or not at all?
uses_e_cigarettes	['Never', 'Former', 'Daily', 'Occasionally']	Would you say you have never used e-cigarettes or other electronic vaping products in your entire life or now use them every day, use them some days, or used them in the past but do not currently use them at all?
days_drinking_alcohol	*0 to 30*	During the past 30 days, how many days per week or per month did you have at least one drink of any alcoholic beverage?
average_sleep_hours	*1 to 24*	On average, how many hours of sleep do you get in a 24-hour period?
general_health	['Excellent', 'Very Good', 'Fair', 'Good', 'Poor']	Would you say that in general your health is:
physical_health_not_good_days	*0 to 30*	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?
mental_health_not_good_days	*0 to 30*	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?
diabetes_diagnosis	['No', 'Yes', 'Prediabetes', 'Gestational']	(Ever told) (you had) diabetes? (If 'Yes' and respondent is female, ask 'Was this only when you were pregnant?')
copd_diagnosis	['No', 'Yes']	(Ever told) (you had) C.O.P.D. (chronic obstructive pulmonary disease), emphysema or chronic bronchitis?
depression_diagnosis	['No', 'Yes']	(Ever told (you had) a depressive disorder (including depression, major depression, dysthymia, or minor depression)?
asthma_diagnosis	['No', 'Yes']	(Ever told) (you had) asthma?
body_mass_index	*12.05 to 97.65*	Body Mass Index (BMI)
primary_health_insurance	['Medicare', 'Employer Plan', 'Military', 'Private Plan', 'Medigap', 'Other Govt', 'Medicaid', 'No Coverage', 'State Plan', 'IHS', 'CHIP']	What is the current primary source of your health insurance?
could_not_afford_medical_care	['No', 'Yes']	Was there a time in the past 12 months when you needed to see a doctor but could not because you could not afford it?
last_doctor_visit	['Never', '<1yr', '1-2yrs', '2-5yrs', '5+ yrs']	About how long has it been since you last visited a doctor for a routine checkup?

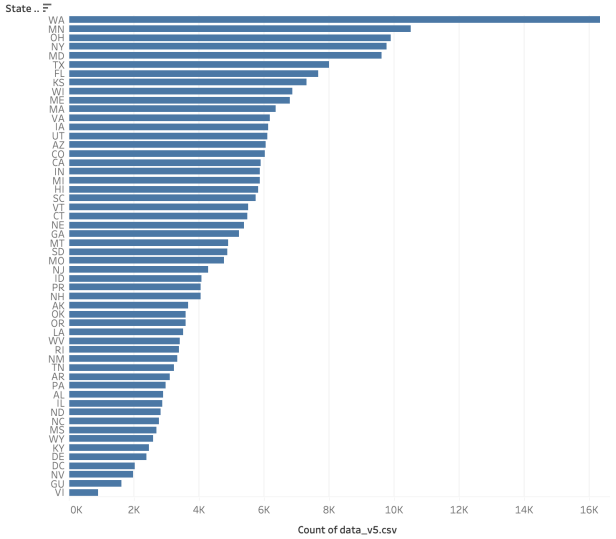
Appendix S6: Statistically significant ($p < 0.05$) explanatory variables with respect to CVD prevalence. The model is a generalized linear model based on the binomial function.

	coef	std err	P>|z|	[0.025	0.975]
state_fips_code[T.CA]	-0.221	0.088	0.012	-0.393	-0.049
state_fips_code[T.CO]	-0.2587	0.089	0.004	-0.434	-0.083
state_fips_code[T.HI]	-0.3763	0.088	0	-0.548	-0.204
state_fips_code[T.IA]	-0.1727	0.083	0.038	-0.336	-0.009
state_fips_code[T.MT]	-0.2386	0.087	0.006	-0.409	-0.068
state_fips_code[T.UT]	-0.2605	0.088	0.003	-0.433	-0.088
state_fips_code[T.VI]	-0.79	0.195	0	-1.172	-0.408
state_fips_code[T.WA]	-0.1774	0.073	0.015	-0.32	-0.035
state_fips_code[T.WY]	-0.2785	0.103	0.007	-0.48	-0.077
sex[T.Male]	0.812	0.017	0	0.779	0.845
age_group[T.30-34]	0.3645	0.137	0.008	0.095	0.634
age_group[T.35-39]	0.6717	0.129	0	0.418	0.925
age_group[T.40-44]	0.9391	0.125	0	0.694	1.185
age_group[T.45-49]	1.3585	0.122	0	1.119	1.598
age_group[T.50-54]	1.6269	0.12	0	1.391	1.863
age_group[T.55-59]	1.9192	0.119	0	1.686	2.153
age_group[T.60-64]	2.1548	0.119	0	1.922	2.387
age_group[T.65-69]	2.4224	0.119	0	2.188	2.656
age_group[T.70-74]	2.6704	0.12	0	2.435	2.906
age_group[T.75-79]	2.9423	0.121	0	2.706	3.179
age_group[T.80+]	3.2205	0.121	0	2.984	3.457
race_ethnicity[T.Multiracial]	0.5085	0.06	0	0.391	0.626
race_ethnicity[T.Other]	0.2636	0.053	0	0.16	0.367
race_ethnicity[T.White]	0.2871	0.032	0	0.224	0.351
education_level[T.No HS]	0.0842	0.035	0.017	0.015	0.153
education_level[T.Some College]	0.0951	0.019	0	0.057	0.133
income_category[T.$15K-25K]	0.0942	0.034	0.005	0.028	0.161
marital_status[T.Single]	-0.2128	0.032	0	-0.276	-0.149
marital_status[T.Widowed]	0.054	0.027	0.043	0.002	0.106
employment_status[T.Employed]	-0.4275	0.035	0	-0.495	-0.36
employment_status[T.Homemaker]	-0.2328	0.055	0	-0.341	-0.124
employment_status[T.Retired]	-0.2294	0.031	0	-0.29	-0.168
employment_status[T.Self-Employed]	-0.2556	0.04	0	-0.334	-0.177
employment_status[T.Student]	-0.2879	0.137	0.035	-0.556	-0.02
employment_status[T.Unemployed <1yr]	-0.3313	0.077	0	-0.482	-0.181
employment_status[T.Unemployed >1yr]	-0.201	0.061	0.001	-0.32	-0.082
general_health[T.Fair]	1.5664	0.038	0	1.491	1.642
general_health[T.Good]	1.0596	0.035	0	0.99	1.129
general_health[T.Poor]	1.8928	0.046	0	1.802	1.984
general_health[T.Very Good]	0.5043	0.036	0	0.433	0.575
diabetes_diagnosis[T.Yes]	0.3508	0.114	0.002	0.127	0.574
copd_diagnosis[T.Yes]	0.4762	0.021	0	0.435	0.518
depression_diagnosis[T.Yes]	0.1451	0.02	0	0.105	0.185
asthma_diagnosis[T.Yes]	0.128	0.021	0	0.087	0.169
primary_health_insurance[T.Employer Plan]	-0.9956	0.424	0.019	-1.827	-0.164
primary_health_insurance[T.Medicare]	-0.8439	0.424	0.046	-1.675	-0.013
primary_health_insurance[T.Medigap]	-1.0616	0.47	0.024	-1.983	-0.14
primary_health_insurance[T.Military]	-0.9499	0.425	0.025	-1.783	-0.117
primary_health_insurance[T.No Coverage]	-0.972	0.427	0.023	-1.808	-0.136
primary_health_insurance[T.Private Plan]	-0.9857	0.425	0.02	-1.818	-0.153
primary_health_insurance[T.State Plan]	-0.8984	0.426	0.035	-1.734	-0.063
could_not_afford_medical_care[T.Yes]	0.3422	0.03	0	0.283	0.402
last_doctor_visit[T.2-5yrs]	-0.1607	0.061	0.008	-0.279	-0.042
last_doctor_visit[T.5+ yrs]	-0.376	0.071	0	-0.515	-0.237
last_doctor_visit[T.<1yr]	0.3408	0.035	0	0.273	0.409
days_drinking_alcohol	-0.0055	0.001	0	-0.007	-0.004
average_sleep_hours	-0.0298	0.005	0	-0.039	-0.021
physical_health_not_good_days	0.005	0.001	0	0.003	0.007
body_mass_index	0.0053	0.001	0	0.003	0.008

Appendix S7: Logistic regression (RQ1) coefficients for feature importance



Appendix S8: Behavioral Risk Factors and Surveillance System (BRFSS) 2022 response counts by state



Appendix S9: Logistic regression model for the disaggregated employment status against cardiovascular disease prevalence

Logit Regression Results

Dep. Variable:	cvd	No. Observations:	271802
Model:	Logit	Df Residuals:	271794
Method:	MLE	Df Model:	7
Date:	Sat, 12 Apr 2025	Pseudo R-squ.:	0.08824
Time:	19:48:51	Log-Likelihood:	−73134.
converged:	True	LL-Null:	−80212.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	−2.7214	0.026	−103.077	0.000	−2.773	−2.670
emp_Disabled	1.3939	0.033	41.825	0.000	1.329	1.459
emp_Employed	−0.6530	0.031	−21.239	0.000	−0.713	−0.593
emp_Homemaker	−0.1004	0.052	−1.934	0.053	−0.202	0.001
emp_Retired	1.1025	0.028	39.348	0.000	1.048	1.157
emp_Student	−1.6835	0.126	−13.391	0.000	−1.930	−1.437
emp_Unemployed <1yr	−0.2393	0.073	−3.284	0.001	−0.382	−0.096
emp_Unemployed >1yr	0.3675	0.058	6.303	0.000	0.253	0.482

Appendix S10: Logistic regression model for the disaggregated health insurer against cardiovascular disease prevalence

Logit Regression Results

Dep. Variable:	cvd	No. Observations:	271802
Model:	Logit	Df Residuals:	271791
Method:	MLE	Df Model:	10
Date:	Sat, 12 Apr 2025	Pseudo R-squ.:	0.06786
Time:	19:47:52	Log-Likelihood:	−74768.
converged:	True	LL-Null:	−80212.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	−3.1870	0.046	−69.065	0.000	−3.277	−3.097
ins_CHIP	1.1076	0.378	2.931	0.003	0.367	1.848
ins_Employer Plan	−0.1063	0.049	−2.174	0.030	−0.202	−0.010
ins_IHS	0.7941	0.130	6.112	0.000	0.539	1.049
ins_Medicaid	0.8905	0.053	16.701	0.000	0.786	0.995
ins_Medicare	1.5793	0.047	33.549	0.000	1.487	1.672
ins_Medigap	0.9715	0.198	4.913	0.000	0.584	1.359
ins_Military	1.1880	0.055	21.471	0.000	1.080	1.296
ins_Other Govt	0.8406	0.065	12.963	0.000	0.713	0.968
ins_Private Plan	0.3996	0.054	7.379	0.000	0.293	0.506
ins_State Plan	0.5994	0.064	9.338	0.000	0.474	0.725

Appendix S10: Decision tree classifier that uses behavioral and lifestyle risk factors to predict cardiovascular disease risk.