Impact of Socioeconomic Factors on Heart Disease Incidence in U.S.

Elizabeth Choe, Yaodong Xin, Yijun Guo

1 Introduction

Heart disease is the leading cause of death in the U.S. Main risk factors of heart diseases among Americans including high blood pressure, high cholesterol, smoking, alcohol, and inactivity have been rigorously investigated. In recent years, aside from those physical factors, socioeconomic status(SES) has become a hot topic to determine the risk stratification of several chronic diseases.

SES is typically represented by income level, employment status, education level, and environmental factors. The discrepancy of SES not only leads to different risk levels of diseases adjusted for other factors but also influences the effect of interventions. For example, the efficacy of behavioral counseling on smoking cessation is more limited among low SES groups than high SES groups, and taking SES markers into account could enhance the precision of risk prediction systems[1]. Though it was shown that heart disease incidence and mortality are higher among middle and low SES groups, the impact of SES on the effects of different risk factors has not been rigorously demonstrated. The dynamic SES under social events (e.g.: COVID-19, economic downturn) also calls for renewal of research findings.

Therefore, in our project, we would like to investigate the influence of SES using 2021's research data and provide the latest evidence of SES's function on heart disease risk prediction, which might implicit possible prophylactic steps in different SES groups in America. The computing and visualization codes are in 625 project

2 Methods

2.1 Dataset

The dataset used in our project originates from the Behavioral Risk Factor Surveillance System (BRFSS) of CDC, which includes data of telephone health status surveys across America. The 2021 dataset consists of over 400,000 observations and 279 variables. 22 variables related to heart disease were extracted as subset for analysis in our project: 1) Outcome: the personal reported incidence of heart disease (CHD) or myocardial infarction (MI) was taken as the outcome; 2) SES factors: income level,education level, employment status, and environmental factors about insurance, healthcare provider, and frequency of body check; 3) Confounders: demographic information(age, sex, race), disease history(diabetes, kidney disease, skin disease), life habits(alcohol intake, smoking, exercise), general health indicators(physical/mental health score). The above variables are renamed for analysis convenience, See original BRFSS variable codes and new codes in coding table and **Table 1**.

2.2 Statistical analyses

Statistical analyses were conducted by RStudio, and result visualization were conducted by *ggplot*2, *gtsummary* and *dplyr* packages. To begin analysis, since the data was derived from social telephone survey, transformations of variable categories were firstly conducted. Values represent "Refuse","Don't know/Not Sure", "Not asked or Missing" were treated as missing values. Unordered categorical variables were treated as dummy variables during modeling.

After transformation, data visualizations were created to show associations between outcome and other variables by conducting logistic regression. By investigating the distribution of values in each categories, we also found that there exist unbalanced data in our data set, especially for outcomes (See detailed analysis in transformation and descriptive). Proportions of missing values for each variable are also displayed in **Table 1**, and **Figure 1** further demonstrated the nonrandom missing pattern for variables.

Table 1. Descriptive analyses of Variables

Variable Category	Name	Meaning	Type	Percentage of Missing Value(%)	$P^{a,b}$
Basic Factors	HEARTDISEASE	Having MI or CHD	Binary	0.00	<0.001***
	SEX	Sex	Binary	0.00	<0.001***
	EXERCISE	Exercise	Binary	0.18	0.193
	STROKE	Stroke	Binary	0.22	<0.001***
	SKINCANCR	Skin cancer	Binary	0.25	<0.001***
	KIDNEYDIS	Kidney disease	Binary	0.32	<0.001***
	DIABETES	Diabetes	Binary	0.15	<0.001***
	DECISION	Hard to make decision	Binary	4.50	<0.001***
	SMOKE100	Ever smoked more than 100 cigarettes	Binary	5.53	<0.001***
	ASTHMA	Asthma	Binary	0.72	<0.001***
	GENHLTH	General Health	Category	0.24	<0.001***
	RACE	Race	Category	2.44	<0.001***
	AGE	Age	Category	0.00	<0.001***
	PHYSHLTH	Days physical health impairment(past 30 days)	Continuous	2.10	<0.001***
	MENTHLTH	Days mental health impairment(past 30 days)	Continuous	1.75	0.276
SES Factors	ALCOHOL	Drinking days(past 30 days)	Continuous	6.92	<0.001***
	INSURANCE	Health insurance type	Category	3.89	<0.001***
	DOCTOR	Personal health provider	Category	0.81	<0.001***
	CHECKUPS	Last checkup time	Category	1.30	<0.001***
	EDUCATION	Education level	Category	0.52	<0.001***
	INCOME	Income level	Category	21.35	<0.001***
	EMPLOY	Employment status	Category	1.83	<0.001***
	MEDCOST	Can't afford medical cost in the past 12 months	Binary	0.27	<0.001***

a. Calculated by *t* test of coefficient in logistic regression

b. *** p <0.001

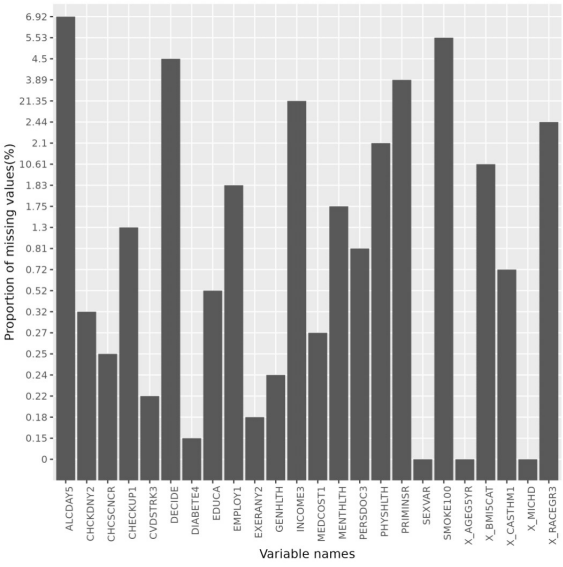

Figure 1 Missing Values

2.3 Data processing

2.3.1 Missing Values Imputation

The nonrandom missing pattern and missing abundance called for imputation.KNN method was applied to conduct imputation.

2.3.2 Unbalanced Labels

According the statistical analyses, there are 91.8% (398,735) normal observations and 8.2% (35,323) observations with heart disease in our dataset. Such unbalanced label pattern would introduce bias when classifying. To be specific, the classifier may pay more attention on the majority label rather than treat two labels equally. Several methods could be utilized to tackle this problem, e.g. assigning different weights to each class, sampling techniques, and changing threshold. Among these methods, a sampling method called Synthetic Minority Oversampling Technique (SMOTE) is broadly used for its simplicity and efficiency without loss of much information. SMOTE is a typical data augmentation technique combining the ideas of both oversampling and undersampling, automatically generates new samples from current dataset.

2.4 Models

2.4.1 Logistic Regression

Logistic regression (LR) is a type of generalized linear model, which means that it is a linear model that is used to predict a categorical response. It is called "logistic" because it uses the logistic function to model the probability of an event occurring (**Figure 2**). The logistic function is a sigmoid curve that maps any input value to a value between 0 and 1. This allows the model to predict the probability that an example belongs to a certain class.

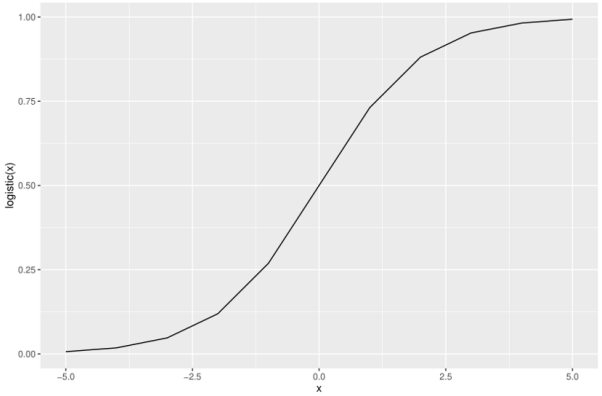

Figure 2 Logistic Function

2.4.2 XGBoost

XGBoost (eXtreme Gradient Boosting) is a powerful and widely-used gradient boosting algorithm in machine learning. Gradient boosting is a machine learning technique that creates a predictive model by adding a sequence of weak learners to an ensemble model. The base models are trained on the residual errors of the previous one, with the goal of minimizing the overall prediction error. XGBoost is an implementation of gradient boosting that is optimized for faster training and higher performance, using techniques such as multi-core CPU parallelization for big data problems. It has also provided a variety of hyperparameters that can be tuned to customize the model's learning process and improve its performance.

2.4.3 XGBoost + SparseLR

Typically, the LR model can only handle linear association between the outcome and predictors. It is always time consuming and sometimes subjective to extract features from the original dataset manually. Many researchers have explored various ways to perform an efficient feature engineering. GBDT+LR[2] was a combined machine learning model proposed to process and extract the information in the dataset automatically. To be specific, we firstly train a gradient boosting tree model. For each tree in the ensemble model, every data row is going to fall on one of the leaf nodes, forming a new feature vector. Suppose there are k trees in the model and each tree has l leaf nodes on average. The original $n \times m$ dataset will then be transformed into a $n \times kl$ dataset, with each row as an one-hot vector. Since the new dataset only contains 0 and 1 and most them are 0, we could apply sparse matrix algorithm to accelerate the latter Logistic Regression module. However, there is also some argument about whether the combination of GBDT and LR really outperforms a sole GBDT model. Therefore, we are also doing experiments [1]to compare the result of different models.

[1]For better implementation in R, we choose XGBoost instead of a vanilla GBDT model.

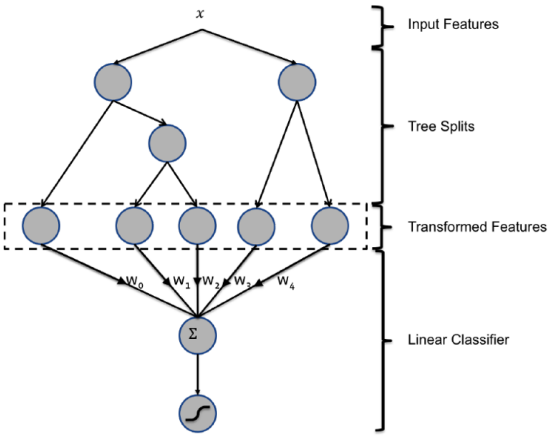

Figure 3 Overview of GBDT+LR [3]

2.5 Evaluation

2.5.1 Data Split

We firstly divide our dataset into the train set (80%) and the test set (20%). For all of our models mentioned above, we would use the train dataset to train the model and evaluate it on both sets but mainly focus on the result of the test set.

2.5.2 Hyperparameters Searching

Note that for SMOTE sampling method and XGBoost model, there are a set of hyperparameters to determine. For example, the number of neighbors chosen k in SMOTE, the number of trees and the nodes of each tree in XGBoost, etc. A feasible solution is to perform a 5-fold cross-validation on train dataset. To be specific, we firstly decide a rough range of the hyperparameters and divide the train dataset into 5 equal subsets. And then we use the four of them to try possible hyperparameters and select the best hyperparameters based on the performance of the rest one subset. The optimal hyperparameters used in our model are listed in table **Table 2**.

Table 2. Optimal hyperparameters used in models

Model	Hyperparameter	Value
SMOTE	#neighbor	5
SMOTE	Proportion	1:1
XGBoost	Learning rate	0.8
XGBoost	#trees	100
XGBoost	Max depth	3

3 Results

3.1 Model evaluation

1) LR Coefficients: For logistic regression model, we use the test set for inference and the full dataset to calculate model Coefficients, e.g. OR (Odds Ratio) and its corresponding p-value. The results are listed in **Table 3**.

Table 3. Logistic Regression Model Results

Variable Name	OR (95% CI)	*P* [a]
HEARTDISEASE (Intercept)	1.01 (0.77, 1.33)	0.921
SEX	2.16 (2.1, 2.23)	<0.001***
EXERCISE	1.02 (0.99, 1.06)	0.171
STROKE	2.58 (2.46, 2.71)	<0.001***
SKINCANCR	1.14 (1.09, 1.18)	<0.001***
KIDNEYDIS	1.67 (1.59, 1.75)	<0.001***
DIABETES	1.25 (1.23, 1.27)	<0.001***
DECISION	1.11 (1.06, 1.17)	<0.001***
SMOKE100	1.47 (1.43, 1.52)	<0.001***
ASTHMA	0.75 (0.72, 0.79)	<0.001***
GENHLTH	0.61 (0.6, 0.63)	<0.001***
RACE	1.05 (1.03, 1.06)	<0.001***
AGE	0.78 (0.78, 0.79)	<0.001***
PHYSHLTH	1.00 (1.00, 1.00)	<0.001***
MENTHLTH	1.00 (1.00, 1.00)	0.994
ALCOHOL	1.01 (1.01, 1.01)	<0.001***
INSURANCE	1.00 (1.00, 1.00)	0.012*
DOCTOR	0.78 (0.76, 0.8)	<0.001***
CHECKUPS	1.28 (1.24, 1.31)	<0.001***
EDUCATION	1.03 (1.01, 1.04)	0.002**
INCOME	1.02 (1.01, 1.03)	<0.001***
EMPLOY	0.96 (0.95, 0.96)	<0.001***
MEDCOST	1.31 (1.23, 1.39)	<0.001***

a. *** p <0.001, ** p <0.01, * p <0.05

2) XGBoost Vector Influence: Note that XGBoost, as a boosting ensemble learning strategy, is hard to interpret directly. However, we could calculate the influence value of each factor. In addition, this value would not

be affected by the scale of the data.

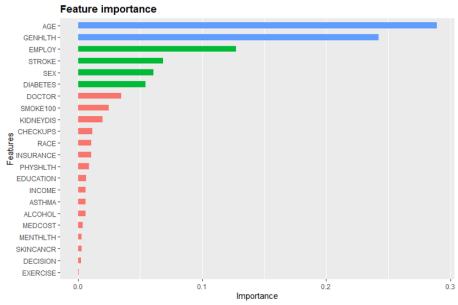

Figure 3 Feature Importance

3) Model Metrics: Based on the models and preprocessing methods mentioned above, we calculate different metrics for evaluation. The details are listed in the **Table 4**.

Table 4. Model evaluation and comparasion

		Delete+LR	Delete+XG	kNN+SMOTE +LR	kNN+SMOTE +XG	kNN+SMOTE +XG+LR
Train	AUC	0.84	0.86	0.85	0.90	0.91
	F1	0.33	0.35	0.78	0.81	0.82
	Specificity	0.74	0.76	0.76	0.81	0.81
	Precision	0.21	0.22	0.76	0.81	0.81
	Recall	0.80	0.81	0.79	0.81	0.82
	Accuracy	0.75	0.76	0.77	0.81	0.81
Test	AUC	0.84	0.85	0.85	0.85	0.85
	F1	0.35	0.41	0.34	0.36	0.36
	Specificity	0.74	0.76	0.75	0.81	0.80
	Precision	0.22	0.28	0.21	0.24	0.24
	Recall	0.79	0.79	0.79	0.72	0.73
	Accuracy	0.75	0.76	0.76	0.80	0.79
Time		<1s	15s	<1s	2s	1min

As shown in the table, the XGBoost based model always perform slightly better than LR models. In terms of SMOTE sampling, the F1 score increased drastically on train dataset after using SMOTE. However, the F1 score of test dataset seems to be similar. This may indicate that, sampling-based method is also essentially assigning different weights to two classes, as the distribution of train and test dataset changed as well as the threshold chosen. But there is an advantage of using SMOTE. After using sampling methods, the size of our dataset shrank by 80%, leading to much faster training speed, but remained almost the same result. This indicates that for some large dataset, we could use sampling methods to get a trained model more quickly. Also, note that the performance of XGBoost+LR almost has the same performance with a sole XGBoost model. A possible reason is because the XGBoost model has already learned enough information from the dataset and a LR does not really help to make it better. This result is also consistent to some research mentioned before.

3.2 Interpretation

As shown in **Table 3** , all of selected socioeconomic factors show significant associations with the heart disease incidence($p<0.05$), confounded by other risk factors. We also showed that employment status and existence of personal health care provider are most influential ones.

4 Conclusion

From the analysis results, we can conclude that SES factors, including income level, insurance coverage, employment status, education level, access to medical care in daily lives have significant impact on heart disease risk. We also showed that employment status and existence of personal health care provider have largest influence. In further study, subgroup analysis can be performed based on using those two variables as grouping criteria, to show how SES inequality might impact the effects of risk factors of heart disease.

References

[1] Schultz WM, Kelli HM, Lisko JC, Varghese T, Shen J, Sandesara P, Quyyumi AA, Taylor HA, Gulati M, Harold JG, Mieres JH, Ferdinand KC, Mensah GA, Sperling LS. Socioeconomic Status and Cardiovascular Outcomes: Challenges and Interventions. Circulation. 2018 May 15;137(20):2166-2178. doi: 10.1161/CIRCULATIONAHA.117.029652. PMID: 29760227; PMCID: PMC5958918.

[2] Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, and Joaquin Quiñonero Candela. 2014. Practical Lessons from Predicting Clicks on Ads at Facebook. In Proceedings of the Eighth International Workshop on Data Mining for Online Advertising (ADKDD'14). Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/2648584.2648589

Notes

See codes and original documents in Github https://github.com/elizabethjchoe/biostat625-group5-project