Predicting Myocardial Infarction Incidence Using Binary Logistic Model

Evan Donald[1], Dilli Bhatta[1]

ABSTRACT

Heart disease is the leading cause of death in the United States. Myocardial infarction is one of the various forms of heart disease. Determining the factors associated with the risk of heart attack is important to health professionals and the public. We investigated the prevalence of myocardial infarction incidence (1 = if myocardial infarction occurred, 0 = no myocardial infarction occurred) from NHANES III data and used a logistic regression model for the analysis of binary data. The model indicated that sex, age, diabetes, high cholesterol, congestive heart failure, and chest pain are the significant risk factors for heart attack with sex, congestive heart failure, and chest pain as the three most significant ones. The classification accuracy of the fitted logistic regression model was 91.28%. Because the prevalence of the incidence was fairly low, we also used Firth logistic regression model and compared it with the logistic regression model. The results were almost similar but comparing the AIC values, it was found that the firth logistic regression was slightly better. The estimated logistic regression equation is useful in predicting whether an individual is prone to the risk of heart attack based on his/her personal information, diet, smoking habits, and health.

Keywords: Cardiovascular Disease, Coronary Heart Disease, MLE, PMLE, AUC, ROC Curve

INTRODUCTION

Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States (Heart Disease 2020). A heart attack (or a myocardial infarction) is one of the various forms of heart disease, including arrhythmia, coronary artery disease, heart failure and many others. It occurs when blood flow to the heart is blocked. The blockage is often a buildup of fat, cholesterol, and other substances, which forms plaque in the coronary arteries. Once the plaque breaks away and forms a clot, the blood flow is stopped, and the heart muscle could be damaged or destroyed (Heart attack 2020). In the United States, someone experiences a heart attack every forty seconds and every year, about 805,000 Americans have a heart attack. Of those 805,000, almost 605,000 are first time heart attacks and 200,000 happen to people who have previously had a heart attack (Heart Disease 2020). About 1 in 5 heart attacks are silent-the damage is done, but the person is not aware of it (Heart Disease 2020).

Studies have shown that the number of heart disease deaths varies based on personal information, such as, race, sex, and age. Heart disease is the cause of 165.8, 205.7, 111.3, 82.6, 141.1 deaths per 100,000 populations in Whites, African Americans, Hispanic, Asian or Pacific Islanders and American Indian or Alaska Natives respectively (Health Status 2019). Similarly, for every 100,000 males, there are 204.8 deaths, and for every 100,000 females, there are 126.2 deaths due to heart disease (Health Status 2019). In the article by Mozaffarian et. al. (2015), chart 19-2: Prevalence of myocardial infarction by age and sex shows that the percentage of the United States population that has had a heart attack increases regardless of gender. Specifically, for males, 0.3% of death occurs in the 20-39 age group, 3.3% of death occurs in the 40-59 age group, 11.3% of death occurs in the 60-79 age group, and 17.3% in the age group above 80. Similarly, for female, 0.3% of death occurs in the 20-39 age group, 1.8% of death occurs in the 40-59 age group, 4.2% of death occurs in the 60-79 age group, and 8.9% in the age group above 80.

Studies also show that a person's diet can increase or decrease the risk of heart disease. A healthy diet is a major factor in reducing your risk for heart disease (Heart disease and diet 2020). A poor diet can drastically increase the chance of a person having a heart attack. Poor quality diets are high in refined grains and added sugars, salt, unhealthy fats and animal-source foods; and low in whole grains, fruits, vegetables, legumes, fish and nuts (Anand et al. 2015). A low-saturated fat, high-fiber, high plant food diet can substantially reduce the risk of developing heart disease (Heart disease and food 2020). Smoking is a major cause of cardiovascular disease (CVD). According to the American Heart Association, CVD accounts for about 800,000 U.S. deaths every year, making it the leading cause of all deaths in the United States. Of those, nearly 20 percent are due to cigarette smoking (How Smoking Affects 2020). One out of every five smoking-related deaths is caused by heart disease (Smoking 2021). Also, cigarette smokers are 2 to 4 times more likely to get heart disease than nonsmokers (Smoking and Cardiovascular Disease 2021). Other health concerns can also increase the risk of heart disease. High cholesterol increases the risk of other conditions, depending on which

[1]Division of Mathematics and Computer Science, University of South Carolina Upstate, 800 University Way, Spartanburg, SC 29303

Corresponding author: Dilli Bhatta, Division of Mathematics and Computer Science, University of South Carolina Upstate, 800 University Way, Spartanburg, SC 29303

How to cite this article: Evan Donald, Dilli Bhatta. Predicting myocardial infarction incidence using binary logistic model. International Journal of Contemporary Medical Research 2022;9(9):I1-I7.

blood vessels are narrowed or blocked. The main risk associated with high cholesterol is coronary heart disease (CHD) (Cholesterol: High Cholesterol Diseases 2020). About 38% of American adults have high cholesterol (total blood cholesterol ≥ 200 mg/dL) (Cholesterol 2020). High blood pressure increases the risk for heart disease and stroke, two leading causes of death for Americans (High Blood Pressure 2020). Similarly, diabetes and heart disease often go hand in hand. If someone has diabetes, one is twice as likely to have heart disease or a stroke than someone who doesn't have diabetes-and at a younger age. The longer one has diabetes, the more likely it is to have heart disease (Diabetes and Your Heart 2020). Cardiovascular disease (CVD) is the leading cause of death worldwide and a major public health concern, CVD prediction is one of the most effective measures for CVD control (Yang et.al. 2020).

In this study, we investigated the prevalence of myocardial infarction incidence in US adults and used a logistic regression model to estimate the probability that an individual is prone to the risk of heart attack based on various factors. Therefore, the objectives of this study are (i) to model the likelihood of myocardial infarction incidence in relation to a person's demographic information, diet, smoking habits and health and (ii) to identify the significant factors associated with the heart attack.

MATERIAL AND METHODS

Data and Variables

In this paper, we used data from the Third National Health and Nutrition Examination Survey (NHANES III). The survey was conducted between the years 1988 and 1994 and contains data for 33,994 people ages 2 months and older who participated in the survey. The National Center for Health Statistics of the Centers for Disease Control and Prevention collects, analyzes, and disseminates data on the health status of United States residents.

The data for the survey interview and examination components are found in four separate data files: (1) NHANES III Household Adult, (2) NHANES III Household Youth, (3) NHANES III Examination, (4) NHANES III Laboratory. We used the data from NHANES III Household Adult file only that contains 20050 observations. We selected some important variables from the data file and renamed them so that it would be convenient to present the results. These variables are listed in Table 1. Among the listed variables, we take HrtAtk which is the Myocardial Infarction Incidence (1 = myocardial infarction occurred, 0=no myocardial infarction occurred) as a dependent (response) variable and the rest as predictor variables. The responses for the dependent variable were self-reported by the respondents on the survey question stated as "Doctor ever told you had a heart attack". We classify the predictor variables as: (1) Personal information: race, sex, and age; (2) Smoking habits: SmkrAtHm, NSmkrAtHm, CurSmkrStat; (3) Diet: AmtCheese, AmtPrMt; (4) Health: HrtCongt, Hbp, HChol, ChstPain. The selection of variables was determined based on prior research into the causes of heart attacks. In order to prepare the data for the final statistical analysis, we removed the cases with "Blank but applicable", "Don't know" and "blank". This led to the significant reduction of the sample size, thereby leaving 4859 data rows at the end. In this sample 8.89% (= 432) of adults reported that they were told by their doctor that they had a heart attack. With regard to racial composition, 73.90% (= 3591) were White, 24.26% (= 1179) were African American and the remaining 1.84% were other races. Additionally, the data contains 55.57% males and 44.43% females. The mean age of the adults in the study was 670.3 months (= 55.9 yrs) with a standard deviation of 207.4 months (= 17.3 yrs). People who smoke make up 39.8% (= 1936) of the sample and those who do not smoke make up the remaining 60.2% (= 2923). About 44% of people interviewed live with a smoker. Of those who do live with a smoker, the most common number of smokers is one. In terms of diet, an average person consumes cheese about 10.3 (≈10) times on average in a month and processed meat about 7.8 (≈8). In the data, 6.3% (= 306) of people experienced congestive heart failure, 38.0% (= 1845) have high blood pressure, 35.3% (=1714) have high cholesterol, and 34.8% (=1691) experienced chest pain.

Statistical Analysis

The dependent variable in our study is a binary type (1 = myocardial infarction occurred, 0 = no myocardial infarction occurred). Therefore, we use logistic regression to model the probability (p) of the occurrence of event 1 of the dependent variable as a function of the predictor variables. Note that if the response variable has more than two categories, it would be multinomial logistic regression. Logistic regression is useful in predicting an outcome of the dependent variable based on one or more sets of independent variables. Given a set of k predictors (continuous or categorical) $x_1, x_2, \dots x_K$, we can write a logistic regression model as:

$$\log\left(\frac{p}{1\text{-}p}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\dots+\beta_kx_k \qquad (1)$$

where and $p=P(y=1|x_1,x_2,\dots,x_k)$ and $\frac{p}{1\text{-}p}$ is called the odds of success, a ratio of the probability of success (event = 1) to the probability of failure (event = 0). The ratio of odds of success to the odds of failure is known as an odds ratio (OR). The OR represents the constant effect of a predictor x, on the likelihood that one outcome will occur (Grace-Martin, n.d.). The coefficient β_i measures the expected changes in log-odds per unit change in x_i holding other predictors as constant. Simplifying equation (1), we get an expression for p as:

$$p=\left[1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\dots+\beta_kx_k)}\right]^{-1} \qquad (2)$$

Note that the models in (1) and (2) are equivalent to each other. In order to fit the logistic regression in (1), we obtain maximum likelihood estimates (MLE) of unknown parameters $\beta_0, \beta_1, \dots \beta_k$. We access the model fit with various goodness-of-fit tests e.g. the Likelihood Ratio (LR) Chi-Square test, Hosmer-Lemeshow test. The p-value for the Hosmer-Lemeshow goodness-of-fit test (HL test) tells us how well our data fits the model. Specifically, the HL test calculates if the observed event rates match the expected event rates in population subgroups. The test is only used for binary response variables (a variable with two outcomes). Small p-values mean that the model is a poor fit. But large p-values don't necessarily mean that your model is a good fit, just that there isn't enough evidence to say it's a poor fit (Glen, 2016). The fitted logistic regression model is useful in predicting the outcome of the dependent variable for the given set of predictor variables.

Because the dataset has a fairly low event rate (proportion of event='1') of the response variable, we also use Firth Logistic Regression which has been in common use for the analysis of binary data with a rare event. It uses penalized likelihood rather than the conventional maximum likelihood for logistic regression. The penalized maximum likelihood estimation (PMLE) was proposed by Firth (1993). Many researchers (Karabon, 2020; Puhr et. al., 2017; Allison, 2012; King, G. and Zeng, L., 2001a, 2001b) have discussed rare events problems in logistic regression. Rare Events and separation are both common analytical challenges encountered when working with a binary variable. Firth's Penalized Likelihood is a simplistic solution that can mitigate the bias caused by rare events in a data set (Karabon, 2020). If the sample size (*n*) is very small (*n*<200), covariates are discrete (preferably dichotomous) and the number of covariates is very small, Exact logistic regression is applicable (Leitgöb, 2013). We compare standard logistic regression and Firth logistic regression using various goodness-of-fit measures such as: AIC, ROC curve, and classification accuracy percentage. Ojha et. al. (2019) also used logistic regression and compared it with Firth logistic regression in an ecological application. In comparing the models, one with a lower AIC score is superior. Note that classification accuracy percentage is a basic diagnostic measure of prediction accuracy. It is a percentage of records correctly classified by the model. We used a cut-point probability value of z=0.50 to classify an observation as an event or nonevent observation. Note that z ϵ [0,1]. If the predicted event probability exceeds or equals 0.50, the observation is predicted to be an event observation; otherwise, it is predicted to be a nonevent observation. The ROC curve' graphically summarizes the tradeoff between true positives (the number of event observations that are correctly classified as events) and true negatives (the number of nonevent observations that are correctly classified as nonevents) for a rule or model that predicts a binary response variable (Wicklin, 2018). True positives are the number of event observations that are correctly classified as events and true negatives are the number of nonevent observations that are correctly classified as nonevents (SAS Documentation, 2019). In comparing two ROC curves, one that is closer to the upper-left corner, indicates better performance of the classification model. If the curve is closer to the 45-degree diagonal of the ROC space, the test is considered to be less accurate (Chan n.d.). The area under the ROC curve (AUC) is a way to assess classification model performance with higher AUC values indicating better test performance. The possible values of AUC range from 0.5 to 1.0 (NCSS Statistical Software n.d).

We split the dataset into two parts. The first part of the data, called the training set, contained 75% of the samples. The second data set, called the testing set, had the remaining 25% of the data. The training set was used to build the model and the testing data is used to assess the fitted model. We used SAS 9.4 for data management and entire statistical analysis. We used PROC logistic procedure to perform logistic regression and Firth logistic regression.

RESULTS

As mentioned in Section 2 (sub section 2.2), the first 75% (= 3644) of the data was used to construct the model and the remaining 25% (= 1215) was used to validate the fitted model. There are 13 predictor variables and one outcome variable. We used PROC logistic procedure with a stepwise variable selection method. This procedure retained six significant predictors (Sex, Age, Diabtc, HrtCongt, HChol, and ChstPain). The parameter estimates of the fitted logistic model along with odds ratio estimates using these six explanatory variables are presented in Table 2. If a variable has a positive coefficient, the likelihood of a heart attack increases due to that variable and if the coefficient is negative, the chance of heart attack decreases. For instance, the logistic regression coefficient 0.922 corresponding to the variable Sex compares the log-odds of heart attack for males with that of females given the other variables are held constant in the model. It indicates that the log-odds of heart attack is expected to be 0.922 units higher for males compared to females, while holding the other variables constant in the model. Alternatively, the odds of a heart attack in males is $e^{.922}$=2.514 times higher than the females which in turn is equivalent to having the odds ratio of 2.514.

The chi-square statistic of the likelihood ratio test is 795.618 (p-value <.0001). This shows that the fitted model is significant. Similarly, the chi-square statistics of the Hosmer-Lemeshow goodness-of-fit test is 13.6177 (p-value = 0.0923) indicating that there isn't enough evidence to say it's a poor fit. The measures for the goodness-of-fit of the model are AIC = 2165.164, SC = 2171.365 and -2logL = 2163.164. The ROC curve for the selected model is presented in Figure 1.

The AUC value is 0.9104 and the classification accuracy of the obtained model is 92.6%. The ROC curve is closer to the upper-left corner and the AUC value is also significantly high. Additionally, the classification accuracy value of 92.6% tells us that for the 3,644 observations used in the model, the model correctly predicted whether or not somebody has a heart attack 92.6% of the time. This seems to be a very good result. Note that this accuracy percentage was computed based on the same data that is used to fit the model. So, this is called in-sample prediction accuracy. We then used the obtained model in testing data which is not is used to create the model and calculate the prediction accuracy to access how well it performed. The ROC curve for the testing data is presented in Figure 2.

The AUC value is 0.8812 and the classification accuracy is 91.28%. Although these values are slightly

Variable Description	Coding	Rename
Has a doctor ever told you that you had a heart attack?”	1 = Yes, 2 = No, 8 = Blank but applicable, 9 = Don't know	HrtAtk
How many heart attacks have you had?	01, 02, 03, 04, 05, 06, 09, 10, 20, 25, 88 = Blank but applicable, 99 = Don’t know, “ ” = Blank.	NHrtAtk
Race	1 = White, 2 = Black, 3 = Other, 8 = Mexican-American of unknown race	Race
Sex	1 = Male, 2 = Female	Sex
Age in months at interview (screener)	0204 -1079 , 1080 = 1080+ months, 9999 = Don't know	Age
Does anyone who lives here smoke cigarettes in the home?	1 = Yes, 2 = No, 8 = Blank but applicable	SmkrAtHm
Total number of persons who smoke cigarettes in the home	00 = No smokers in the home, 01, 02, 03, 04, 05, 06, 88 = Blank but applicable	NSmkrAtHm
Did mother have heart attack?	1 = Yes, 2 = No, 8 = Blank but applicable, “ ” = Blank	MthrHrtAtk
Did father have heart attack?	1 = Yes, 2 = No, 8 = Blank but applicable, “ ” = Blank	FthrHrtAtk
Have you ever been told by a doctor that you have diabetes or sugar diabetes?	1 = Yes, 2 = No, 8 = Blank but applicable, 9 = Don't know	Diabtc
Has a doctor ever told you that you had congestive heart failure?	1 = Yes, 2 = No, 8 = Blank but applicable, 9 = Don't know	HrtCongt
Have you ever been told by a doctor or other health professional that you had hypertension, also called high blood pressure?	1 = Yes, 2 = No, 8 = Blank but applicable, 9 = Don't know, “ ” = Blank	Hbp
Have you ever been told by a doctor or other health professional that your blood cholesterol level was high?	1 = Yes, 2 = No, 8 = Blank but applicable, 9 = Don't know, “ ” = Blank	HChol
Have you ever had any pain or discomfort in your chest?	1 = Yes, 2 = No, 8 = Blank but applicable, 9 = Don't know	ChstPain
Cheese, all types - times/month	000 = Never, 001-182, 888 = Blank but applicable, 999 = Don’t know	AmtCheese
Bacon/sausage/processed meats - times/mo	0000 = Never, 0001-0608, 8888 = Blank but applicable, 9999 = Don't know	AmtPrMt
Do you smoke cigarettes now?	1 = Yes, 2 = No, 8 = Blank but applicable, “ ” = Blank	CurSmkrStat

Table-1: List of selected variables from NHANES III Household Adult file

lower in comparison to the in-sample case they both indicate the model performed fairly well in prediction and classification in testing data. Because the percentage of individuals in the data set who had a heart attack is about 8.8%, we use the Firth logistic regression for rare events. To implement the Firth method in SAS, we use the FIRTH option in PROC LOGISTIC. We used the same significant variables retained from the stepwise selection method in logistic regression. The result of Firth logistic regression is presented in Table 3.

The chi-square statistic of the likelihood ratio test is 793.645 (p-value <.0001). This shows that the fitted model is significant. The measures for the goodness-of-fit of the model are AIC = 2124.380, SC = 2130.581 and -2logL = 2122.380. The ROC curve, AUC value, and classification accuracy of the fitted model are similar to the logistic regression model. When the model is applied to testing data the ROC curve, AUC value, and classification accuracy are similar to when the fitted logistic regression model was applied to testing data. These results indicate that the Firth logistic regression model predicted the probability of a heart attack at the same accuracy as the logistic regression model for both in-sample and out-of-sample. However, comparing the goodness-of-fit measure like AIC (2165.164 for logistic regression vs 2124.380 for Firth logistic regression), it seems that Firth logistic regression has slightly a better fit in comparison to logistic regression. Moreover, the estimated standard errors of the regression coefficient are slightly smaller than the logistic regression model standard error.

DISCUSSION

For years, researchers have worked to develop prediction models for heart disease. Wilson et. al. (1998) developed a coronary heart disease (CHD) prediction model based on the blood pressure, cholesterol, and LDL-C categories proposed by the Joint National Committee (JNC-V) and National Cholesterol Education Program (NCEP) and some other risk factors using a Framingham Heart Study sample. It was a prospective, single-center study in the setting of a community-based cohort where the patients were 2489 men and 2856 women 30 to 74 years old at baseline with 12 years of follow-up. Sex-specific prediction models were also formulated to predict CHD risk according to age, diabetes, smoking, JNC-V blood pressure categories, and NCEP total cholesterol and LDL cholesterol categories under the various sets of independent variables. The c-statistic (which is equal to the area under the receiver operating characteristic curve, AUC), a commonly used measure for quantifying the predictability of working models for these models were in the range 0.68 to 0.77. The AUC for our logistic model when was 0.8812. Yang et. al. (2020) used a Random Forest algorithm

Analysis of Maximum Likelihood Estimates					Odds Ratio Estimates		
Parameter	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr> Chi-Square**	**Point Estimate**	**95% Wald Confidence Limits**	
Intercept	-8.276	0.4439	347.538	<.0001			
Sex Male	0.922	0.1615	32.587	<.0001	2.514 (Male vs Female)	1.832	3.449
Age	0.005	0.0005	104.454	<.0001	1.005	1.004	1.006
Diabtc Yes	0.595	0.1779	11.198	0.0008	1.814 (Yes vs No)	1.280	2.570
HrtCongt Yes	2.465	0.1744	199.770	<.0001	11.760 (Yes vs No)	8.355	16.551
HChol Yes	0.465	0.1470	9.995	0.0016	1.592 (Yes vs No)	1.193	2.123
ChstPain Yes	1.946	0.1606	146.810	<.0001	6.998 (Yes vs No)	5.108	9.586

Table 2: Maximum likelihood estimates (MLE) and odds ratio estimates of logistic regression

Analysis of Penalized Maximum Likelihood Estimates					Odds Ratio Estimates		
Parameter	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr> Chi-Square**	**Point Estimate**	**95% Wald Confidence Limits**	
Intercept	-8.227	0.4406	348.699	<.0001			
Sex Male	0.914	0.1606	32.408	<.0001	2.495 (Male vs Female)	1.821	3.419
Age	0.005	0.0005	104.420	<.0001	1.005	1.004	1.006
Diabtc Yes	0.596	0.1772	11.301	0.0008	1.815 (Yes vs No)	1.282	2.568
HrtCongt Yes	2.450	0.1740	198.178	<.0001	11.583 (Yes vs No)	8.236	16.291
HChol Yes	0.463	0.1464	9.999	0.0016	1.589 (Yes vs No)	1.192	2.117
ChstPain Yes	1.934	0.1596	146.776	<.0001	6.916 (Yes vs No)	5.058	9.457

Table-3: Penalized maximum likelihood estimates (PMLE) and odds ratio estimates of Firth logistic regression

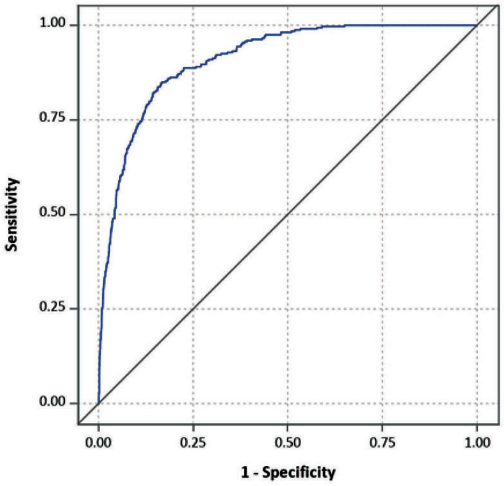

Figure-1: ROC curve in training data. Area under the curve = 0.9104

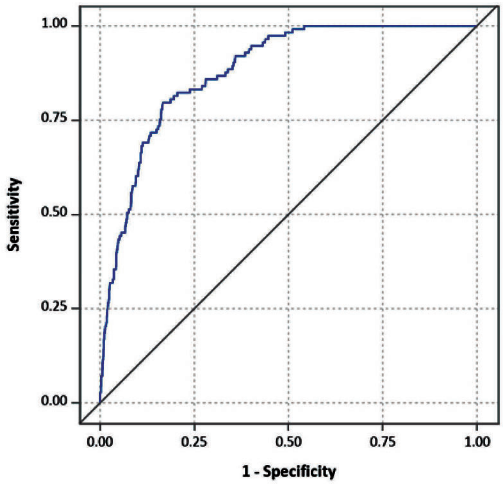

Figure-2: ROC curve in testing data. Area under the curve = 0.8812

to provide a prediction model for the risk of cardiovascular disease (CVD). This study focused on predicting the three-year risk of CVD for high-risk subjects in eastern China (Yang, 2020). In order for the comparison purpose, they also used several other methods to build a prediction model including multivariate regression model, classification and regression tree (CART), Naïve Bayes, Bagged trees, Ada Boost setting multivariate regression model as a benchmark for performance evaluation. They used nearly 30 predictors related to the risk of CVD in building the prediction models. These predictors were screened through logistic regression analysis and the set included sex, age, family income, smoking, drinking, obesity, excessive waist circumference, abnormal cholesterol, abnormal low-density lipoprotein, abnormal fasting blood glucose, and many more. The AUC was used to evaluate the prediction ability of each model. The study resulted in a multivariate regression model with an AUC of 0.7143 and a Random Forest model was superior to other models with an AUC of 0.787. While the Random Forest model performed quite well, our model performed slightly better if AUC is used as a performance gauge. A key difference in our study and Yang et. al. (2020) is the region that the data was collected. Yang et. al. (2020) collected a sample from a population in eastern China with a known high risk of CVD. Our study used data collected by the CDC using the National Health and Nutrition Examination Survey. Since our data was not collected from a specific area within the United States, the model created can be used nationwide to predict heart attack risk.

Many prediction models for the risk of cardiovascular disease (CVD) have been developed and several studies were conducted from time to time to review these published models. A systematic review of 212 articles by Damen et.

al. (2016) showed 363 models were developed regarding CVD in the general population. Most prediction models were developed using Cox proportional hazards regression, accelerated failure time analysis, or logistic regression, and the majority of the models were sex-specific (Damen et. al. 2016). Different types of predictive performance measures were reported for developed models. For models that used AUC were in the range of 0.60 to 1.00. Our model differed from most models in a few key categories. Our study is not sex-specific, includes participants with existing CVD, and uses a logistic regression model. In comparison to other models using AUC, our model performs at a high level and can be used nationwide. With the six variables Sex, Age, Diabtc, HrtCongt, HChol, and ChstPain, the risk of heart attack can be predicted for any person within the United States thereby providing a simplified prediction model for heart disease.

CONCLUSION

In this study, we investigated the prevalence of myocardial infarction incidence in US adults and used a logistic regression model to estimate the probability that an individual is prone to the risk of heart attack based on various factors. The model showed that of the thirteen independent variables, six are significant in predicting heart attacks namely sex, age, diabetes, high cholesterol, congestive heart failure, and chest pain. The three most significant variables are sex, congestive heart failure, and chest pain. The result showed that males are more likely to have a heart attack. Also, as the person gets older, the probability of heart attack increases. Similarly, people who have congestive heart failure and people who experience chest pain are also more likely to have had a heart attack. Lastly, people with diabetic problems and high cholesterol are also at-risk of a heart attack. The obtained model can be used to predict the likelihood of a person having a heart attack in the future and predict the probability that a person had a previous heart attack and did not know it. The classification accuracy for the fitted logistic model was 91.28% (88.89% true positive and 2.39% true negative) when applied to the training data. Similar accuracy was also obtained for the Firth logistic regression. This shows that the model was more accurate in predicting the probability of nonevent observation (no heart attack).

Acknowledgments

We would like to thank the Centers for Disease Control and Prevention in making all NHANES data available for public use.

Declaration of Interest

The author declares that there is no conflict of interest.

REFERENCES

1. Allison, P.D., 2012. Logistic Regression for Rare Events. Stat. Horiz. Accessed December 29 2021. https://statisticalhorizons.com/logistic-regression-for-rare-events
2. Chan, C., What is a ROC Curve and How to Interpret It. https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it
3. Cholesterol (last updated Sept. 8, 2020), accessed July 23 2021. https://www.cdc.gov/cholesterol/index.htm
4. Cholesterol: High Cholesterol Diseases (last updated Dec. 28, 2020). Accessed July 23, 2021. https://my.clevelandclinic.org/health/articles/11918-cholesterol-high-cholesterol-diseases
5. Damen, J. A., Hooft, L., Schuit, E., Debray, T. P., Collins, G. S., Tzoulaki, I., Lassale, C. M. et. al. (2016). Prediction models for cardiovascular disease risk in the general population: Systematic review. BMJ, 353: i2416. https://doi.org/10.1136/bmj.i2416
6. Diabetes and Your Heart (last updated May 07 2020). Accessed August 10 2021. https://www.cdc.gov/diabetes/library/features/diabetes-and-heart.html
7. Firth, D. (1993): Bias reduction of maximum likelihood estimates. In: Biometrika 80: 27-38.
8. Fryar CD, Chen T-C, Li X. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010 pdf icon[PDF-494K]. NCHS data brief, no. 103. Hyattsville, MD: National Center for Health Statistics; 2012. Accessed May 9 2019.
9. Grace-Martin, K., Why use Odds Ratios in Logistic Regression? https://www.theanalysisfactor.com/why-use-odds-ratios/
10. Heron, M. Deaths: Leading causes for 2017 pdf icon[PDF – 3 M]. National Vital Statistics Reports; 68(6). Accessed November 19, 2019.
11. Heart attack (last updated June 16, 2020). Accessed July 24 2021. https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106
12. Heart Attack Facts & Statistics. (n.d.). Accessed April 15 2019. https://www.cdc.gov/heartdisease/heart_attack.htm
13. Heart disease and diet (last updated July 30, 2020). Accessed July 31 2021. https://medlineplus.gov/ency/article/002436.htm
14. Heart disease and food. (August 20, 2020). Accessed July 31 2021. https://www.betterhealth.vic.gov.au/health/ConditionsAndTreatments/heart-disease-and-food
15. Heart Disease (last updated Sept 8, 2020). Accessed July 23 2021. https://www.cdc.gov/heartdisease/facts.htm
16. Health Status (2019). Accessed July 23 2021. https://www.kff.org/state-category/health-status/heart-disease/
17. How Smoking Affects Heart Health (last updated May 04 2020). Accessed July 23 2021.https://www.fda.gov/tobacco-products/health-information/how-smoking-affects-heart-health
18. High Blood Pressure (last updated Oct. 22, 2020). Accessed July 23 2021. https://www.cdc.gov/bloodpressure/index.htm
19. King, G. and Zeng, L. (2001a): Logistic Regression in Rare Events Data. In: Political Analysis 9: 137-163.
20. King, G. and Zeng, L. (2001b): Explaining Rare Events in international Relations. In: International Organization 55: 693-715.
21. Karabon, P (2020). Rare Events or Non-Convergence with a Binary Outcome? The Power of Firth Regression in PROC LOGISTIC. SAS Global

Forum 2020. Paper 4654-2020. https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2020/4654-2020.pdf

22. Leitgöb, H. (2013). The Problem of Modeling Rare Events in ML-based Logistic Regression - Assessing Potential Remedies via MC Simulations. European Survey Research Association. https://www.europeansurveyresearch.org/conf/uploads/494/678/167/PresentationLeitg_b.pdf
23. Mozaffarian, D., Benjamin, E.J., Go, A.S., et al. (2015) Heart Disease and Stroke Statistics—2015 Update: A Report from the American Heart Association. Circulation, 131, e29-e322. https://doi.org/10.1161/CIR.0000000000000152 https://www.ahajournals.org/doi/epub/10.1161/CIR.0000000000000152
24. NCSS Statistical Software, Comparing Two ROC Curves – Paired Design (Chapter-547). https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Comparing_Two_ROC_Curves-Paired_Design.pdf
25. Puhr, R., Heinze, G, Nold, M, Lusa, L, Geroldinger, A. (2017). Firth's logistic regression with rare events: accurate effect estimates and predictions? Stat Med. 36(14), 2302-2317. https://doi.org/10.1002/sim.7273
26. Prevalence of coronary heart disease by age and sex. (2015). Retrieved April 15, 2019, from https://www.heart.org/idc/groups/heart-public/@wcm/@sop/@smd/documents/downloadable/ucm_449846.pdf
27. SAS Documentation (2019). The LOGISTIC Procedure. https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_logistic_details32.htm
28. Smoking and Cardiovascular Disease (2021). Accessed July 23 2021. https://www.hopkinsmedicine.org/health/conditions-and-diseases/smoking-and-cardiovascular-disease
29. Stephanie Glen (2016). "Hosmer-Lemeshow Test: Definition" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/hosmer-lemeshow-test/
30. Sonia S. Anand, Corinna Hawkes, Russell J. de Souza, Andrew Mente, Mahshid Dehghan, Rachel Nugent, Michael A. Zulyniak et.al. (2015). Food Consumption and its impact on Cardiovascular Disease: Importance of Solutions focused on the globalized food system: A Report From the Workshop Convened by the World Heart Federation. Journal of American College of Cardiology, 66(14): 1590–1614. https://doi.org/10.1016/j.jacc.2015.07.050
31. Stephanie Glen (2016). "Hosmer-Lemeshow Test: Definition" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/hosmer-lemeshow-test/
32. Wicklin, R. (2018). Create and compare ROC curves for any predictive model. SAS Blogs. https://blogs.sas.com/content/iml/2018/11/14/compare-roc-curves-sas.html
33. Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. Circulation, 97(18), 1837–1847. https://doi.org/10.1161/01.cir.97.18.1837
34. Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., Yu, W., and Yan, J (2020). Study of cardiovascular disease prediction model based on random forest in eastern China. Scientific Reports, 10:5245. https://doi.org/10.1038/s41598-020-62133-5

Source of Support: Nil; **Conflict of Interest:** None

Submitted: 26-07-2022; **Accepted:** 28-08-2022; **Published:** 30-09-2022