Dean&Francis
ISSN 2959-6157

Associative Study of Smoking, Alcohol Consumption Frequency and Age on the Risk of Heart Disease

Zitong Zheng

School of statistics, University of International Business and Economics, Beijing, China

*Corresponding author: 202292025@uibe.edu.cn

Abstract:

Nowadays, the impact of heart disease on global human health is profound and severe. Cardiovascular disease (CVD) remains to be the leading cause of death worldwide. According to the latest research from the Institute for Health Metrics and Evaluation, CVD continues to be the primary cause of death globally, urgently necessitating the establishment of public health strategies by countries to prevent CVDs. It is possible to create a world free of heart disease and stroke, yet millions of lives are prematurely ended by heart disease each year. The article studies on the main factors supposed to have relationship with heart disease suffering, such as aging, smoking status, alcohol intake status and family history. By using Chi-square Test/Fisher Test as well as Monte Carlo Simulation and logistic regression, the article receives expected results that smoking, alcohol intake and aging both have positive correlation with heart disease suffering with smoking owning a greater impact on heart disease. The significance of this paper is to verify the positive correlation between each factor and the risk of heart disease, and to compare the biggest influence factors on the risk of heart disease.

Keywords: Cardiovascular disease; Smoking; Chi-square Test; Logistic regression.

1. Introduction

Under the global background that heart disease still being a main disease responsible for people's short lifespan [1,2], many studies and research have been done as heart disease being the main cause for people's death [3]. About statistics on heart disease and Stroke, the American Heart Association (AHA) collaborates with the National Institutes of Health (NIH) to report the latest statistics on heart disease, stroke, and cardiovascular risk factors annually. These data encompass core health behaviors such as smoking, physical activity, nutrition, sleep, and obesity, as well as health factors that contribute to cardiovascular health, such as cholesterol, blood pressure, blood sugar control, and metabolic syndrome. And one global Heart and Brain Health report highlights the top ten factors affecting heart and brain health, which are of significant importance for the prevention and treatment of cardiovascular diseases [4].

Regarding the global risk factors for heart disease, such as smoking, alcohol consumption, and aging, the following is significant research background. For Smoking and Heart Disease, according to data from the World Health Organization, smoking and secondhand smoke exposure are harmful to the cardiovascular system. Approximately 1.9 million avoidable deaths from coronary heart disease (accounting for about 21% of all global deaths from coronary heart disease) are attributed to smoking and secondhand smoke exposure each year [5]. The health benefits of quitting smoking are significant, starting to show just a few hours after quitting and can last for several years. After 15 years of quitting smoking, the excess risk of coronary heart disease is reduced to the level of those who have never smoked [6]. For Alcohol Intake and Heart Disease, alcohol intake mainly leads to the increase of HDL and LDL (both are types of cholesterol) which supposed to have strong relationship with heart disease suffering. A study has found that consuming 23 grams of alcohol per day was associated with a 31% reduction in the risk of ischemic heart disease (IHD). However, no significant association has been found between drinking and the risk of IHD when going further and it indicates that moderate alcohol consumption may have a positive impact on certain biomarkers and be associated with a reduced risk of IHD. Nevertheless, the study emphasized that intermittent heavy drinking may have an adverse effect on IHD [7]. Research from Massachusetts General Hospital has found that light to moderate alcohol consumption may be associated with a lower risk of heart disease, but this does not mean that alcohol should be used to reduce the risk of heart attacks or strokes, as alcohol has other concerning impacts on health [8]. Furthermore, a study from the University of Calgary in Canada synthesized the results of dozens of studies over 30 years and found that moderate drinkers had a lower risk of heart disease than non-drinkers. Moderate alcohol consumption can significantly increase the level of “good” cholesterol (HDL) in the body, effectively preventing heart disease. However, excessive drinking is detrimental to health [9]. For Aging and Heart disease, though the incidence of heart failure in high-income countries has stabilized or decreased over the past decade, the prevalence continues to rise due to population aging, an increase in risk factors, the effectiveness of new therapies, and improved survival rates [10]. This rise in prevalence is becoming increasingly prevalent among younger individuals and is accompanied by an increase in heart failure with preserved ejection fraction.

The article is structured into four key sections: methods, function, results, and conclusion. It commences with a detailed description of the methods, including database selection and statistical analyses for correlation testing. The variables chosen for these tests are outlined in the subsequent section. The function section delves into the correlation tests, illustrating the processes and their respective functions. The results section presents the study’s findings and procedures, followed by a summarizing conclusion that encapsulates the entire study’s insights.

2. Basic Information of the Text and used Methods

2.1 Basic Information

The database chosen for this study is sourced from Kaggle, specifically the “Heart Disease Prediction” data set, which was last updated approximately three months prior and boasts a high official usability rating of 10. This comprehensive data set is comprised of detailed information on various individuals. It is presumed to be linked to the presence of heart disease, along with a range of risk factors that may contribute to its development. The data set includes 16 columns of factors, encompassing demographic details such as age and gender, an individual’s medical history, lifestyle choices, and symptoms that are commonly associated with heart disease. The target variable within this data set clearly indicates whether a particular individual has been diagnosed with heart disease.

The primary objective of this study is to elucidate the correlations between the habits of smoking and alcohol intake and the likelihood of suffering from heart disease. Furthermore, the research aims to conduct a single factor analysis to examine the impact of smoking and alcohol consumption on the progression of heart disease. By isolating these variables, the study seeks to provide a clearer understanding of how these lifestyle choices might independently influence the risk of developing heart disease, thereby contributing valuable insights to the field of cardiovascular health research.

This study aims to establish a significant link between smoking and alcohol consumption habits and the prevalence of heart disease. With a sample of 1000 participants, it examines the impact of varying levels of alcohol and tobacco use, categorized as Never/Former/Current for smoking and None/Moderate/Heavy for alcohol. The research also considers family history as a contributing factor.

2.2 Statistical Analysis

Data from 1000 individuals, stratified by sex, age (25-34, 35-64, 65-79), smoking, and alcohol intake, were analyzed. To explore the relationship between these factors and heart disease, a two-factor hypothesis test was conducted. The Chi-square Test was applied where expected frequencies were above 5, and the Fisher test was used for frequencies below this threshold. A p-value of <0.05 was deemed statistically significant.

Logistic regression was employed to determine the regression coefficients of these factors. By comparing coefficients, the study seeks to identify the greatest contributor to heart disease risk between smoking and alcohol intake. Notably, the high incidence of heart disease among non-smokers and non-drinkers warrants further correlation tests to understand underlying causes. The analysis also compares heart disease rates across nine groups (Heavy & Current, Heavy & Former, Heavy & Never, Moderate & Current, Moderate & Former, Moderate & Never, None & Current, None & Former, None & Never) and investigates exceptional cases of heart disease in non-smokers and non-drinkers without a family history. A line chart will visually represent the percentage of heart disease cases within each group. All statistical analyses were conducted using R version 4.4.1.

3. Correlation Tests and Logistic Regression

3.1 Different Correlation Tests

The Chi-Square test is used to test the independence between two categorical variables. Its basic equation is:

$$X^2 = \sum \frac{(O-E)^2}{E} \tag{1}$$

Where X^2 is the Chi-Square statistic, O is the observed frequency, and E is the expected frequency, typically calculated based on the marginal totals.

The Fisher's exact test is used for testing the independence of two categorical variables in small sample sizes (if any of the expected frequency is smaller than 5). It calculates the probability of each observed frequency based on the hyper geometric distribution. There is no single equation to describe Fisher's test. It involves the following steps. 1) Calculate the probability of the observed configuration. 2) Identify all configurations extreme or more extreme than the observed configuration. 3) Sum the probabilities of all these configurations to obtain the P-value. For a $2x2$ contingency table, the "equation" for Fisher's test can be represented by the formula for calculating the hyper geometric probability:

$$P(X=x) = \binom{R}{x}\binom{N-R}{n-x} / \binom{N}{n} \tag{2}$$

where $P(X=x)$ is the probability of the specific frequency observed in the given contingency table. In addition, N is the total sample size, n is the size of a subset of the sample, R is the total number of the category of interest in the population, and x is the number of the category of interest observed.

Monte Carlo simulation is a method based on repeated random sampling to calculate the probabilities of possible outcomes. It does not have a fixed equation but proceeds through the following steps. Define the parameters of the base model at first. Then drawing random samples from the base distribution. Conducting statistical analysis on the drawn samples. And repeating steps 2 and 3 multiple times to obtain the distribution of results.

3.2 Logistic Regression

Logistic regression is used to predict the probability of a binary dependent variable. Its basic equation is:

$$log(\frac{P(Y=1)}{1-P(Y=1)}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \tag{3}$$

Or, in a more common form

$$log(\frac{P}{1-P}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X \tag{4}$$

where P is the probability of the dependent variable $Y=1$. $X_1, X_2, \cdots, X_n$ are the independent variables. β_0 is the intercept term, and $\beta_1, \beta_2, \cdots, \beta_n$ are the regression coefficients. The above equations typically estimate the regression coefficients through Maximum Likelihood Estimation.

4. Results

4.1 Study Sample

The study encompasses 1000 comprehensive data points, primarily focusing on smoking and drinking habits. It also integrates personal demographics such as age and sex, along with health metrics like cholesterol levels, blood pressure, and heart rate. The data set is balanced for gender and consists of a 35%:65% ratio of young adults (25-44) to older adults (45-79). Table 1 details the baseline characteristics stratified by alcohol consumption, smoking status, and sex across different age groups.

Table 1. The data from individuals under Smoking status groups

Smoking							
Gender	Age	Men			Women		
Habit		Current	Former	Never	Current	Former	Never
N	The Young, 25-34 (N=163)	22	28	25	28	31	29
Heart disease (%)		0	0	0	0	0	0
Cholesterol-mean(mg/dL)		246.2	248.4	258.5	246.8	263.1	246.2
Blood Pressure-mean(mmHg)		138.2	128	132.6	139.9	139.6	132.7
Heart Rate-mean (in 1 minute)		76	77.4	82.1	77.1	76.3	78.4
N	The Middle-aged, 35-64 (N=564)	100	95	87	89	92	101
Heart disease (%)		33	30.5	40.2	33.7	28.3	36.6
Cholesterol-mean (mg/dL)		250.3	242.1	255.5	248.4	247.3	251.7
Blood Pressure-mean (mmHg)		137.9	134	135.7	135.3	136.2	133.5
Heart Rate-mean (in 1 minute)		79.9	79.5	79.5	80.1	78.4	80.6
N	The Elder, 65-79 (N=273)	47	44	49	50	36	47
Heart disease (%)		83	62.2	77.6	80	66.7	70.2
Cholesterol-mean (mg/dL)		252.1	237.5	254.5	261.8	238.9	253.3
Blood Pressure-mean (mmHg)		131.6	131.5	136.4	134.8	138.4	137.9
Heart Rate-mean (in 1 minute)		79.5	82.1	76.7	77.7	79.3	79.4

Table 2. The data from individuals under Alcohol intake groups

Alcohol Intake							
Gender	Age	Men			Women		
Habit		Heavy	Moderate	None	Heavy	Moderate	None
N	The Young, 25-34 (N=163)	29	23	23	30	22	36
Heart disease (%)		0	0	0	0	0	0
Cholesterol-mean (mg/dL)		256	239.3	256.8	240	258.1	259.2
Blood Pressure-mean (mmHg)		130.9	132.4	134.8	135.8	142.9	135.5
Heart Rate-mean (in 1 minute)		80.5	74.8	79.8	78.1	75.9	77.4
N	The Middle-aged, 35-64 (N=564)	100	95	87	89	92	101
Heart disease (%)		33	30.5	40.2	33.7	28.3	36.6
Cholesterol-mean (mg/dL)		250.3	242.1	255.5	248.4	247.3	251.7
Blood Pressure-mean (mmHg)		137.9	134	135.7	135.3	136.2	133.5
Heart Rate-mean (in 1 minute)		79.9	79.5	79.5	80.1	78.4	80.6
N	The Elder, 65-79 (N=273)	47	44	49	50	36	47
Heart disease (%)		83	62.2	77.6	80	66.7	70.2
Cholesterol-mean (mg/dL)		252.1	237.5	254.5	261.8	238.9	253.3
Blood Pressure-mean (mmHg)		131.6	131.5	136.4	134.8	138.4	137.9
Heart Rate-mean (in 1 minute)		79.5	82.1	76.7	77.7	79.3	79.4

In Table 1 and Table 2, the “Heart disease” row reflects the prevalence of heart disease within each category. For instance, under the Smoking Status section, “33.0” (line 2, seventh from the left) signifies that 33% of middle-aged men (35-64) who currently smoke have heart disease. The data also indicate that smoking impacts heart rate and cholesterol levels in individuals, with these factors trending higher across most groups. Additionally, the quantity of alcohol intake is observed to affect heart rate, particularly in the elderly population.

4.2 Smoking & Drinking & Family history and Heart Disease

From conclusions in the former studies, the article selects: H_0: Smoking and Alcohol intake are *independent* from getting heart disease; H_1: Smoking and Alcohol intake are *dependent* from getting heart disease. During the hypothesis test, data collected from former smokers are picked out. The data used for the test as is shown in Table 3. Table 4 below shows the expected frequency. The calculated p-value is 0.6516, which is bigger than 0.05. Thus, smoking and alcohol are associated with heart disease.

Table 3. The data base used for double factor hypothesis test

N	Alcohol Intake					
	Heavy		Moderate		None	
Smoking Status	Current	Never	Current	Never	Current	Never
Heart Disease	44	49	52	46	46	48
Non - Heart Disease	70	67	53	63	71	65

Table 4. The data base used for double factor hypothesis test

E_{ij}	Alcohol Intake					
	Heavy		Moderate		None	
Smoking Status	Current	Never	Current	Never	Current	Never
Heart Disease	48	44	49	49	46	48
Non - Heart Disease	66	61	68	67	63	65

*Every E_{ij} is bigger than 5, thus *Chi-square Test* is suitable for the hypothesis test.

Single factor correlation tests for further studying on which of them has a stronger link with heart disease are conducted. Assuming that p_1 stands for the p-value of smoking and heart disease test, p_2 stands for the p-value of alcohol intake and p_3 stands for the p-value of family history.

Table 5. The result for both single factor hypothesis test and logistic regression model

Values Factors	Estimate	Std. Error	z value	Pr(>|z|)	OR
Smoking	0.19	0.14	1.39	0.17	1.21
Alcohol Intake	0.06	0.14	0.46	0.64	1.07
Family History	0.12	0.13	0.91	0.36	1.13

The estimate value E represents the change in the logarithmic odds ratio of the probability of disease occurrence with each unit change of the independent variable, holding other variables constant. Through the test shown in Table 5, $p_1 = 0.19$ is much bigger than 0.05 which shows a strong relation with the disease and given with $p_2 = 0.06(withoutformersmokers)$, with $p_3 = 0.12$. In addition, the article conducts a further study in calculating the regression coefficient of the two factors. Eventually, the odds ratio results are 1.21(for smoking), 1.07(for alcohol intake), 1.13(for family history). It is obviously that smoking owns a much stronger relationship with heart disease suffering. In addition, giving up smoking is promised to decrease the rate of suffering from heart disease.

Meanwhile, over-viewing the Table 1 and Table 2, it shows that the rate of people suffering from heart disease appears to be high in those people neither smoke nor taking in alcohol as individuals become older. Despite the age, family history becomes the third factor the study focused on as with the probability of carrying CVD gene people tend to live a healthier life. The result shows a strong relationship between family history and heart disease suffering, with a p-value of 0.5145.

4.3 Aging and Heart Disease

Given that a significant portion of participants are elderly, the article also investigates whether advancing age has a pronounced impact on individuals with healthy habits (non-smokers and non-drinkers) and no family history of heart disease. Due to the limited sample size, Fisher's test and Monte Carlo simulation are employed to obtain a more accurate outcome. For this analysis, the article considers the following: H_0: Aging are dependent from getting heart disease; H_1: Aging are independent from getting heart disease.

The analysis yields exceptionally low p-values (5×10^{-4} for Fisher's test and 0.001 for Monte Carlo simulation), refuting the notion that aging is unrelated to heart disease risk. This anticipated outcome prompts a deeper examination of the age factor across all individuals. The two-sided test yields a similar p-value, suggesting a link between aging and heart disease.

As can be seen from the trend shown in Fig. 1, while young individuals exhibit a heart disease rate of 0%, this rate escalates significantly with age, indicating a positive correlation between aging and heart disease incidence. Previous research identifies aging as a key heart disease risk factor. However, this study's population skews older (mean age 52.293), with a 7:13 ratio of younger (25-44) to older (45-79) individuals. This age distribution, biased towards middle and older ages, may impact results. A balanced representation of age groups would enhance the study's reliability.

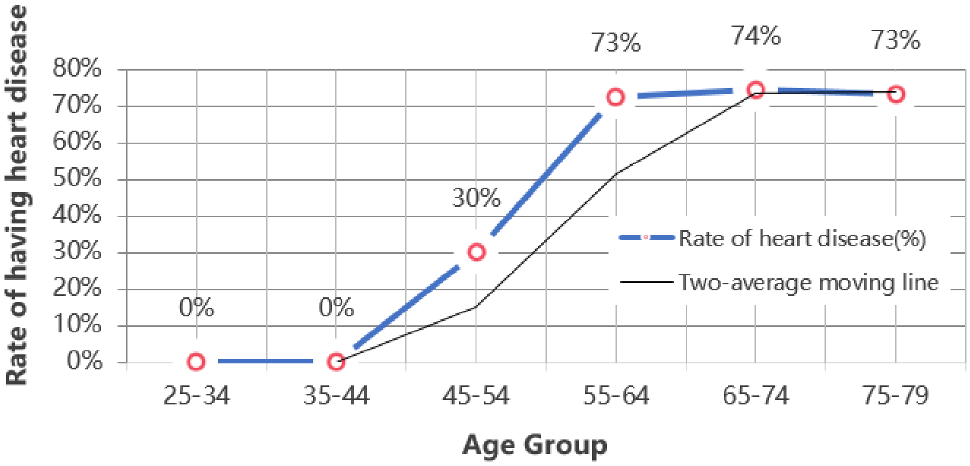

Fig. 1 Line chart for trends in aging and heart disease rates

5. Conclusion

The article deftly unpacks the complex relationship between aging, alcohol intake, and smoking in the onset of heart disease, assigning them a hierarchy of influence with aging second, alcohol consumption third, and smoking at the forefront. It highlights a precipitous increase in the incidence of heart disease as individuals progress in age. Smoking and aging are strongly correlated with the incidence of heart disease, while the connection with alcohol intake appears less pronounced, hinting that moderate alcohol consumption may be associated with a lower risk of heart disease. Despite these insights, the research points to several areas requiring additional investigation. An upcoming study will compare the incidence of heart disease among moderate drinkers against those who abstain to determine if there is a protective effect from moderate alcohol consumption. Furthermore, the study will examine the role of HDL cholesterol as a key indicator to elucidate the underlying reasons for any potential reduction in heart disease rates among moderate drinkers. The current findings are based on a relatively small data set; hence, the anticipation that a larger sample size will yield more robust and generalized conclusions. The variables under consideration demand further scrutiny, and the article is poised to deliver a more sophisticated analysis of these elements, enhancing people's understanding of the multifaceted factors contributing to heart disease.

References

[1] Global Education Monitoring Report, 2023. https://gem-report-2023.unesco.org/.

[2] Zhang, Bei, et al. Global Burden of Cardiovascular Disease from 1990 to 2019 Attributable to Dietary Factors. The Journal of Nutrition, 2023, 153(6): 1730-1741.

[3] Archibald, Carla L., and Nathalie Butt. Using Google Search Data to Inform Global Climate Change Adaptation Policy. Climatic Change, 2018, 150(3-4): 447-456.

[4] Martin, Seth S., et al. 2024 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. Circulation, 2024, 149(8).

[5] World Health Organization. Inaugural World Health Organization Partners Forum Report: Stockholm. Sweden, 2019. World Health Organization, 2020.

[6] Kempton, Hannah R., et al. Using Coronary Artery Calcium Scoring as a Preventative Health Tool to Reduce the High Burden of Cardiovascular Disease in Indigenous Australians. Heart, Lung and Circulation, 2020, 29(6): 835-839.

[7] Carr, Sinclair, et al. A Burden of Proof Study on Alcohol Consumption and Ischemic Heart Disease. Nature Communications, 2024, 15(1): 4082.

[8] Mezue, Kenechukwu, et al. Reduced Stress-Related Neural Network Activity Mediates the Effect of Alcohol on Cardiovascular Risk. Journal of the American College of Cardiology, 2023, 81(24): 2315–25.

[9] Brien, Susan E., et al. Effect of Alcohol Consumption on Biological Markers Associated with Risk of Coronary Heart Disease: Systematic Review and Meta-Analysis of Interventional Studies. BMJ, 2011, 342: d636.

[10] Khan, Muhammad Shahzeb, et al. Global Epidemiology of Heart Failure. Nature Reviews Cardiology, 2024, 21(10): 717–34.