Analysis of Risk Factors for Heart Disease Using CDC Dataset: An Inferential Statistical Approach*

Niracha Chaiwong
School of Computer Science
University of Lincoln
26683869@students.lincoln.ac.uk

***Abstract*—The purpose of this study was to use statistical inference to analyse data from the Behavioural Risk Factor Surveillance System (BRFSS), which is a health survey that collects data on risk factors associated with heart disease from over 400,000 each year. Several statistical tests, including independent t-tests, Chi-square tests, and correlation analyses, were used to test the research hypotheses and investigate the significant associations between risk factors and heart disease. The findings suggest that all risk factors used in this study are positively related to heart disease, except alcohol consumption, which showed a negative relationship.**

***Index Terms*—Independent t-test, Chi-square, correlation**

I. Introduction

Nowadays, heart disease is considered as one of the most causes of death in the world [1]. According to World Health Organization(WHO), over 17 million people died from cardiovascular disease. One of the biggest problems that medical centres face is that the knowledge and skill of doctors and specialists are not at the same level, which leads to negative outcomes in treating their patients and could result in death. Many research methods are proposed to solve the mentioned problem by applying several Machine Learning algorithms to assist humans in terms of disease classification and prediction [2]. There are many factors related to heart disease. For example, UCI Machine Learning Repository in the Cleveland Heart Disease dataset contains 14 heart disease risk factors. This data set is meant to predict heart disease based on clinical and demographic factors [3]. Another dataset from the Centers for Disease Control and Prevention (CDC) which provides information on ten risk factors for heart disease based on population studies and clinical trials [4]. This research aims to apply inferential statistics to draw conclusions about the population from the sample data. So, the CDC has been used in this task since it is intended to provide samples that are representative of a large population, indicating that the data are likely to be representative of the target population. Moreover, the factors are easier to understand and monitor. Inferential statistics will be used to analyse the sample data and drive the conclusions. The proposed method will be mentioned in section III.

II. Literature Review

A. Centers for Disease Control and Prevention (CDC)

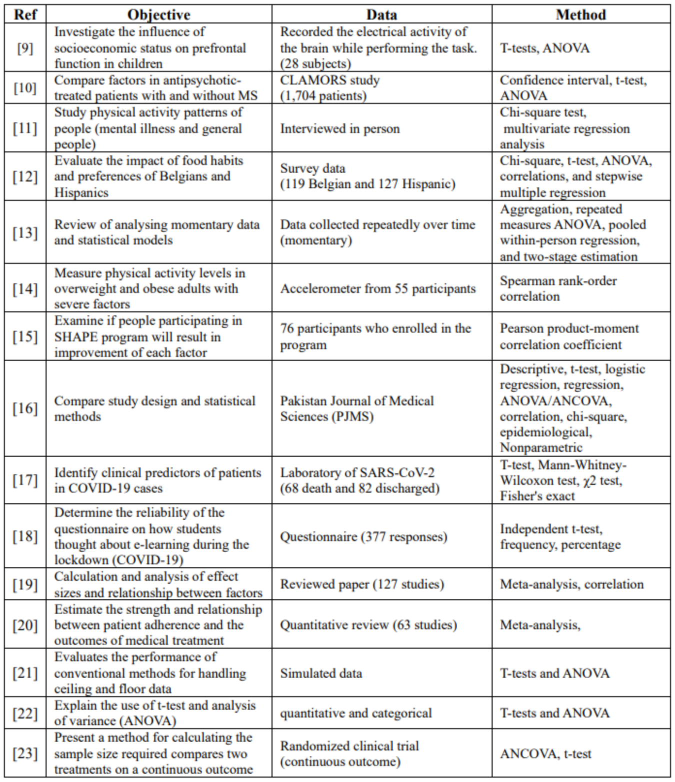
CDC is the United States national public health institute that has collected and analysed data on health-related topics through various methods such as population-based surveys, clinical trials, registries etc. The goal is to understand the impact of the disease and find effective ways to avoid and manage the disease. CDC also provides heart disease risk factors datasets through the National Health and Nutrition Examination Survey (NHANES) and the Behavioral Risk Factor Surveillance System (BRFSS). NHANES is a survey that evaluates the health and nutritional status of American children and adults. The data was collected through interviews, physical examinations, and laboratory testing. BRFSS is a state-based telephone survey which provides essential data to monitor health behaviour trends and inform the development of public health policy and programmes. There are many risk factors based on CDC. For example, high blood pressure, high cholesterol, diabetes, obesity, physical inactivity, poor diet, smoking, etc. [5].

B. Related Work

In this section, several relevant research papers have been reviewed to find a suitable method to analyse and adapt with the heart disease dataset used in this research.

According to Table I, there are many kinds of research about health-related surveys and techniques that have been used to analyse data that collected from surveys, questionnaires, interviews, measurements, simulations, repositories, open source data, etc. Mostly, the objective of the research is to compare factors that may have effects on each other. The methods used are based on what they want to evaluate and the type of dataset. For example, t-tests and ANOVA can deal with interval and category data in terms of mean comparison between two factors. The p-value is used to evaluate if there is a significant relationship between each factor with a significant level of 0.05 [22]. There are several tools that can help to analyse data which are SAS, R and SPSS. SAS can deal with complex and large datasets, while R is known for being flexible and open-source, and SPSS is easy to use. Finally, presenting demographics in frequencies and percentages graphs are easier to observe the data distribution [18].

Table I
Related work summary table



Ref	Objective	Data	Method
[9]	Investigate the influence of socioeconomic status on prefrontal function in children	Recorded the electrical activity of the brain while performing the task. (28 subjects)	T-tests, ANOVA
[10]	Compare factors in antipsychotic-treated patients with and without MS	CLAMORS study (1,704 patients)	Confidence interval, t-test, ANOVA
[11]	Study physical activity patterns of people (mental illness and general people)	Interviewed in person	Chi-square test, multivariate regression analysis
[12]	Evaluate the impact of food habits and preferences of Belgians and Hispanics	Survey data (119 Belgian and 127 Hispanic)	Chi-square, t-test, ANOVA, correlations, and stepwise multiple regression
[13]	Review of analysing momentary data and statistical models	Data collected repeatedly over time (momentary)	Aggregation, repeated measures ANOVA, pooled within-person regression, and two-stage estimation
[14]	Measure physical activity levels in overweight and obese adults with severe factors	Accelerometer from 55 participants	Spearman rank-order correlation
[15]	Examine if people participating in SHAPE program will result in improvement of each factor	76 participants who enrolled in the program	Pearson product-moment correlation coefficient
[16]	Compare study design and statistical methods	Pakistan Journal of Medical Sciences (PJMS)	Descriptive, t-test, logistic regression, regression, ANOVA/ANCOVA, correlation, chi-square, epidemiological, Nonparametric
[17]	Identify clinical predictors of patients in COVID-19 cases	Laboratory of SARS-CoV-2 (68 death and 82 discharged)	T-test, Mann-Whitney-Wilcoxon test, $\chi 2$ test, Fisher's exact
[18]	Determine the reliability of the questionnaire on how students thought about e-learning during the lockdown (COVID-19)	Questionnaire (377 responses)	Independent t-test, frequency, percentage
[19]	Calculation and analysis of effect sizes and relationship between factors	Reviewed paper (127 studies)	Meta-analysis, correlation
[20]	Estimate the strength and relationship between patient adherence and the outcomes of medical treatment	Quantitative review (63 studies)	Meta-analysis,
[21]	Evaluates the performance of conventional methods for handling ceiling and floor data	Simulated data	T-tests and ANOVA
[22]	Explain the use of t-test and analysis of variance (ANOVA)	quantitative and categorical	T-tests and ANOVA
[23]	Present a method for calculating the sample size required compares two treatments on a continuous outcome	Randomized clinical trial (continuous outcome)	ANCOVA, t-test

III. Methodology

A. Research Questions

The main research questions of this paper are to determine what risk factors that can cause heart disease and find a correlation between them. In this part, the hypothesises been created as follows:

1) There is a significant association between cholesterol, smoking status, alcohol drinking, diabetes, general health, age, gender and the risk of heart disease.
2) There is a significant difference in the mean BMI, mental health between individuals with and without heart disease.
3) There is a significant correlation between smoking status and alcohol drinking among individuals with heart disease.

B. Data Acquisition

To answer the research questions from the above section, Heart Disease Health Indicators Dataset has been used, which is the cleaned data from Behavioral Risk Factor Surveillance System (BRFSS) 2015 which has 441,455 responses and 330 factors [6]. The raw data can be accessed from the CDC website [7]. The dataset used in this research contains 253,680 samples and 22 risk factors which is suitable to be used for testing the research hypothesis.

C. Data Prepossessing

From the Heart Disease Health Indicators dataset, there are many column factors which not related to the research questions. So, some irrelevant factors are removed. The factors that have been used in this research are cholesterol, BMI, smoking status, alcohol drinking, diabetes, general health, age, gender, and mental health. Mostly, the factors are encoded. To understand more about the data, the CDC codebook needs to be compared with the data in order to define the label. For example, 0 = No, 1 = Yes, Female = 0, Male = 1, General health = excellent 1 to poor 5, etc [8].

D. Data Analysis

To analyze the dataset, IBM SPSS Statistics Software has been used to test the research hypothesises.

1) Chi-square test: is used to find the difference between the factor's frequency, which can be used in nominal and ordinal data. According to the first research question, the null hypothesis (H_0) is there is no significant association between risk factors and heart disease, while the alternative hypothesis (H_a) is there is a significant association. The same method has been processed with nominal and ordinal factors, which are cholesterol, smoking status, alcohol drinking, diabetes, general health, age and gender.

2) Independent-sample t-test: is used to compare the mean difference between two factors which can analyse quantity data. This method is used to test if BMI and frequent mental health have a significant association with heart disease or not. According to the second research question, the null hypothesis (H_0) is there is no significant association (mean difference) between risk factors and heart disease, while the alternative hypothesis (H_a) is there is a significant association.

3) Correlation: is used to find the direction of the relationship between smoking status and alcohol drinking among heart disease. According to the third research hypothesis, H_0 is there is no significant correlation between smoking status and alcohol drinking, H_a is there is a significant correlation.

E. Data Visualization

After doing the data analysis, all relevant factors will be summarized by using appropriate graphs or tables. Using data visualization can simplify the complexes and make it easier to write the conclusion by using PowerBi.

IV. Result

This section will show some examples of SPSS results in the summarized table and interpret the results. The full SPSS output results are available in supporting material.

A. Data Analysis

1) Chi-square test: Table II summarizes the output from the SPSS program. The results show that Pearson's Chi-Square value is less than a significant value of 0.05. So, it can be concluded that the null hypothesis is rejected or heart disease has a relationship with gender. For cholesterol, smoking status, alcohol drinking, diabetes, general health, age and gender, the results show that all (H_0) are rejected, which can be concluded that all factors have a significant association with heart disease.

Table II
Chi-Square test between gender and heart disease

Gender	Heart disease		
	No	Yes	
Female	131769 (57.3%)	10205(42.7%)	Pearson Chi-square P < 0.001
Male	98018 (42.7%)	13688(57.3%)	
Total	229787(100%)	23893(100%)	

2) Independent-sample t-test: Table III summarizes the output from the SPSS program. The P-value is less than a significant value of 0.05. So, it can be concluded that the null hypothesis is rejected or people with heart disease have a different BMI. The same method has been processed with mental health. The results show that all (H_0) are rejected, which can be concluded that it has a significant association with heart disease.

Table III
Independent-sample t-test between BMI and heart disease

Heart disease	N	Mean	Std. deviation	t	P
No	229,787	28.27	6.585	26.181	<.001
Yes	23,893	29.47	6.741		

3) Correlation: Table IV summarises the output from the SPSS program. The correlation coefficient ranges from -1 to 1, the negative value means a negative correlation, zero means no correlation, and the positive value means a positive correlation. The results show that all factors have a positive direction on heart disease except alcohol consumption which has a negative value. This indicates that as alcohol consumption increases, heart disease tends to decrease.

Table IV
Correlation between smoking status and alcohol drinking

	Pearson Correlation							
	Heart disease	Diabetes	Smoker	Cholesterol	General health	Mental health	BMI	Alcohol
Heart disease	1	.180**	.114**	.181**	.258**	.065**	.053**	-.029**
Diabetes	.180**	1	.063**	.209**	.303**	.074**	.224**	-.058**
Smoker	.114**	.063**	1	.091**	.163**	.092**	.014**	.102**
Cholesterol	.181**	.209**	.091**	1	.208**	.062**	.107**	-.012**
General health	.258**	.303**	.163**	.208**	1	.302**	.239**	-.037**
Mental health	.065**	.074**	.092**	.062**	.302**	1	.085**	.025**
BMI	.053**	.224**	.014**	.107**	.239**	.085**	1	-.049**
Alcohol	-.029**	-.058**	.102**	-.012**	-.037**	.025**	-.049**	1

**. Correlation is significant at the 0.01 level (2-tailed).

B. Data Visualization

After performing statistical analysis, the data was downloaded to Power Bi software to transform data into graphs. Figure 1 shows bar charts, line charts and pie charts of the sample data used in this research. As you can see, some data are not balanced. For example, heart disease has more responses in no heart disease than having heart disease. It is because of the data collected based on the survey. Normally, people with no disease are more than people who have heart disease in our society. The bar chart can show the range and amount of age. Moreover, a line chart can show the number of samples and the most likely survey sample. For example, this data was collected mostly in ages 60-64 and less in 18-24.

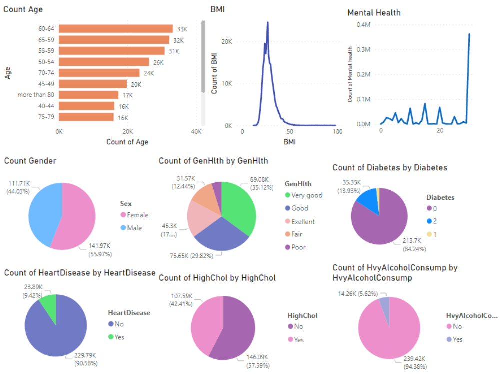


Fig. 1. Contribution of data.

Figure 2 shows some examples between two factors. In heart disease and smoking status, the bar chart shows that people who smoke and have heart disease are more than people who don't smoke and have heart disease. People who don't drink alcohol and have heart disease are more than those drinking alcohol. People who have high cholesterol and heart disease are more than those who don't. Lastly, older people are more likely to have heart disease than younger age.

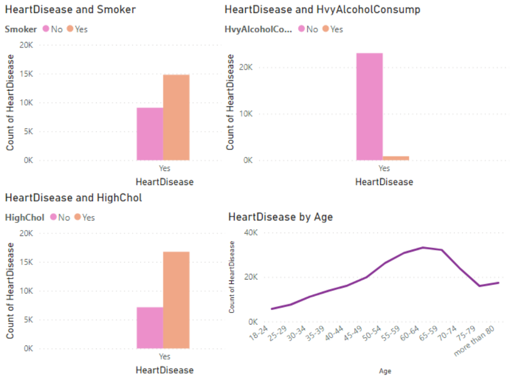


Fig. 2. Example of data visualization between factors.

V. Conclusion

In this research, the data is acquired from the CDC heart disease dataset based on BRFSS. The risk factors that cause heart disease that used to analyse and answer the research questions by setting three hypotheses and using statistical inference. This research focuses on finding the relationship between each risk factor and whether they have a significant association with heart disease or not. The data were analysed by using the SPSS program, and which results showed that every factor has a significant association with heart disease. In the correlation method, results show that all of the factors have a positive direction on heart disease except alcohol consumption which has a negative value. The limitation of this research is that the data that have been used in this research are randomly collected, and most of the responses are without heart disease. This may affect the analysis and conclusion of this research. Increasing the number of samples with heart disease may cause the model more accurate and reliable. Otherwise, this research ensures that all risk factors that have been used truly affect heart disease and provide directions. Moreover, it can be used in heart disease prediction by using Machine Learning from these risk factors.

References

[1] M. Sultana, A. Haider, and M. S. Uddin, “Analysis of data mining techniques for heart disease prediction,” 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Sep. 2016.

[2] M. Sultana, A. Haider, and M. S. Uddin, ”Analysis of data mining techniques for heart disease prediction,” presented at the 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, Bangladesh, Sep. 2016.

[3] “UCI Machine Learning Repository,”Uci.edu, 2018. http://archive.ics.uci.edu/ml/index.php [accessed Apr. 22, 2023].

[4] CDC, “Heart Disease — cdc.gov,” Centers for Disease Control and Prevention, Jan. 19, 2021. http://www.cdc.gov/heartdisease [accessed Apr. 22, 2023].

[5] CDC, “About CDC,” Centers for Disease Control and Prevention, Apr. 29, 2022. https://www.cdc.gov/about/index.html [accessed Apr. 22, 2023].

[6] “Heart Disease Health Indicators Dataset,” www.kaggle.com. https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset [accessed Apr. 23, 2023].

[7] “NHANES 2013-2014: Glycohemoglobin Data Documentation, Codebook, and Frequencies,”wwwn.cdc.gov.https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014.htm [accessed Apr. 23, 2023].

[8] “Behavioral Risk Factor Surveillance System 2015 Codebook Report Land-Line and Cell-Phone data,” 2016. Accessed: Jun. 27, 2021. [Online]. Available: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

[9] M. M. Kishiyama, W. T. Boyce, A. M. Jimenez, L. M. Perry, and R. T. Knight, “Socioeconomic Disparities Affect Prefrontal Function in Children,” Journal of Cognitive Neuroscience, vol. 21, no. 6, pp. 1106–1115, Jun. 2009

[10] C. Arango et al., “A comparison of schizophrenia outpatients treated with antipsychotics with and without metabolic syndrome: findings from the CLAMORS study,” Schizophrenia Research, vol. 104, no. 1–3, pp. 1–12, Sep. 2008

[11] G. L. Daumit et al., “Physical Activity Patterns in Adults With Severe Mental Illness,” The Journal of Nervous and Mental Disease, vol. 193, no. 10, pp. 641–646, Oct. 2005

[12] W. Verbeke and G. Poquiviqui López, “Ethnic food attitudes and behaviour among Belgians and Hispanics living in Belgium,” British Food Journal, vol. 107, no. 11, pp. 823–840, Dec. 2005

[13] J. E. Schwartz and A. A. Stone, “Strategies for analyzing ecological momentary assessment data.,” Health Psychology, vol. 17, no. 1, pp. 6–16, Jan. 1998

[14] G. J. Jerome et al., “Physical activity levels of persons with mental illness attending psychiatric rehabilitation programs,” Schizophrenia Research, vol. 108, no. 1–3, pp. 252–257, Mar. 2009, doi: https://doi.org/10.1016/j.schres.2008.12.006

[15] A. D. Van Citters et al., “A Pilot Evaluation of the In SHAPE Individualized Health Promotion Intervention for Adults with Mental Illness,” Community Mental Health Journal, vol. 46, no. 6, pp. 540–552, Dec. 2009, doi: https://doi.org/10.1007/s10597-009-9272-x.

[16] S. Akhtar, S. W. Ali shah, M. Rafiq, and A. Khan, “Research design and statistical methods in Pakistan Journal of Medical Sciences (PJMS),” Pakistan Journal of Medical Sciences, vol. 32, no. 1, Dec. 1969, doi: https://doi.org/10.12669/pjms.321.9033.

[17] Q. Ruan, K. Yang, W. Wang, L. Jiang, and J. Song, “Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China,” Intensive Care Medicine, pp. 1–3, Mar. 2020, doi: https://doi.org/10.1007/s00134-020-05991-x.

[18] S. Abbasi, T. Ayoob, A. Malik, and S. I. Memon, “Perceptions of students regarding E-learning during Covid-19 at a private medical college,” Pakistan Journal of Medical Sciences, vol. 36, no. COVID19-S4, May 2020, doi: https://doi.org/10.12669/pjms.36.covid19-s4.2766.

[19] K. B. Haskard Zolnierek and M. R. DiMatteo, “Physician Communication and Patient Adherence to Treatment,” Medical Care, vol. 47, no. 8, pp. 826–834, Aug. 2009, doi: https://doi.org/10.1097/mlr.0b013e31819a5acc.

[20] M. Robin DiMatteo, P. J. Giordani, H. S. Lepper, and T. W. Croghan, “Patient Adherence and Medical Treatment Outcomes,” Medical Care, vol. 40, no. 9, pp. 794–811, Sep. 2002, doi: https://doi.org/10.1097/00005650-200209000-00009.

[21] Q. Liu and L. Wang, “t-Test and ANOVA for data with ceiling and/or floor effects,” Behavior Research Methods, Jul. 2020, doi: https://doi.org/10.3758/s13428-020-01407-2.

[22] Hun Sik Park, “Comparing Group Means: T-tests and One-way ANOVA Using Stata, SAS, R, and SPSS,” Jan. 2009.

[23] G. F. Borm, J. Fransen, and W. A. J. G. Lemmens, “A simple sample size formula for analysis of covariance in randomized clinical trials,” Journal of Clinical Epidemiology, vol. 60, no. 12, pp. 1234–1238, Dec. 2007, doi: https://doi.org/10.1016/j.jclinepi.2007.02.006.