Dean&Francis

ISSN 2959-6157

Research on the Analysis of Heart Disease based on Logistic Regression

Xidan Zhang[1, *]

Junyi Zhu[2]

[1]Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom
[2]Department of Mathematics, University of Warwick, Coventry, CV4 7AL, United Kingdom
*Corresponding author: Xidan. Zhang.1@warwick.ac.uk

Abstract:

Heart disease is a general term for heart diseases, including rheumatic heart disease, congenital heart disease, hypertensive heart disease, coronary heart disease, myocarditis and other heart diseases. As a major cause of death worldwide, it is important to further improve the prediction, monitoring, and effective prevention and treatment of heart disease based on the current level of science, technology, and medical care. This study is based on thirteen quantifiable factors influencing heart disease and conducts both overall logistic regression and paired logistic regression analyses. In the process of regression analysis, this paper highlights the statistical significance of each variable, and the results of the regression analysis are presented as a more accurate heart disease prediction model, which can be updated regularly to improve the timeliness of the heart disease prediction model, taking into account the influence of time factors. In addition to this, the paper also asserts the dilemmas and challenges faced by this research from other perspectives, such as genetics, lifestyle, etc., and realizes that in real life, researchers should consider the multidimensional influences of heart disease more comprehensively.

Keywords: Logistic regression; correlation analysis; heart disease.

1. Introduction

Nowadays, heart disease has become the leading cause of death worldwide, causing millions of deaths every year [1]. According to available data, 1.2 million people had coronary heart attacks in the United States in 2005, and nearly half of them died from the disease [2]. And it is evident that heart disease is the leading cause of death worldwide [3]. Therefore, there is an urgent need to increase researchers' focus on prevention and treatment methods for heart disease.

With the rapid development of science and technology, treatments for heart disease are constantly being updated, from the early methods such as drug eluting stents, coronary-artery bypass-graft surgery and anti-thrombosis to today's innovative treatments like Biopharmaceutical-based therapies, the medical community is constantly working to develop new treatments and drugs to reduce the mortality caused by heart disease [4]. However, it is also crucial to analyze the factors that influence the incidence of heart disease. An in-depth study of the multiple risk factors behind this prevalent and highly lethal disease can help doctors achieve early surveillance and develop effective preventive measures [5].

Notably, advances in early screening technologies have played a significant role in reducing mortality from heart disease, such as the electrocardiography method, which provides a more viable pathway for early detection of heart disease [6]. Although the risk factors for heart disease are now well understood by the medical and scientific communities, most research has focused primarily on the impact of underlying diseases (e.g., hypertension and diabetes) on heart disease risk [7]. The interactions between different body markers and these risk factors have not been fully resolved to some extent.

Available data suggest that genetic factors such as congenital heart disease, lifestyle habits, and living environment are important risk factors for heart disease. For example, unhealthy lifestyle, dietary habits such as smoking and physical inactivity can seriously affect the body's indicators, thereby increasing the risk of heart disease [8]. In addition, the prevalence of heart disease is closely related to geographic location and the level of development of a country, with relatively high prevalence in developing countries and low prevalence in developed countries such as Europe [9]. Other studies have also shown that poor lifestyle habits such as lack of daily physical activity are important factors influencing the risk of heart disease [10]. However, these non-quantitative influences make it difficult to precisely describe their interactions and impact on the incidence of heart disease, making it impossible to perform a precise comparative analysis of the impact of each factor. Therefore, this study attempts to match non-quantitative factors such as lifestyle habits and underlying diseases with corresponding physiological indicators in order to provide a more in-depth data analysis of heart disease incidence.

This study aims to provide a comprehensive assessment and data analysis of the risk factors for heart disease incidence from a statistical point of view through multivariate regression analysis and binary logistic regression analysis and to further explore the complex relationship between these factors and the incidence of heart disease at the biomedical level. In addition, building on previous studies, this article pays special attention to the role of quantitative indicators in the analysis of heart disease risk, thus providing a new perspective for understanding the complexity of heart disease incidence. Meanwhile, predicting future trends in heart disease incidence has become one of the topics of extensive research [11].

This paper used R Studio, Python, and SPSS for data analysis in order to visualize the data and provide clear and intuitive graphical presentations. Through this method, this research hopes to provide a more scientific basis for the prevention, diagnosis, and treatment of heart disease from a statistical perspective, and to reveal the multiple influencing factors of heart disease incidence, thus paving the way for the development of effective treatment and prevention strategies.

2. Methods

2.1 Data Sources and Processing

The study used data (from Kaggle) from 271 independent patients with potential heart disease. The data included indicators of the physiologic and medical dimensions of each sample and whether or not those samples had heart disease. The information for each sample contained: age, gender, type of chest pain, blood pressure, cholesterol level, fasting glucose, electrocardiogram results, maximum heart rate, exercise angina, ST-segment depression, ST-segment slope, number of blood vessels visible on fluoroscopy, and thallium scan results. This research ensured the completeness of these data and labeled some binary data as 0 and 1 to facilitate logistic regression analyses based on these data.

2.2 Method Introduction

This analysis used multivariate logistic regression models to analyze the relationship between these variables and the occurrence of heart disease. The independent variables included age, sex, type of chest pain, blood pressure, cholesterol, fasting blood glucose level over 120 mg/dL, electrocardiogram findings, maximum heart rate, exercise-induced angina, ST-segment depression, ST-segment slope, number of visible blood vessels, and thallium scan findings. The corresponding dependent variable was the presence of heart disease.

The data were initially divided into a training set (70%) and a test set (30%). Data were processed using the Python Pandas library, and logistic regression models were built and trained using the Sklearn library. The outputs relate to the confusion matrix, precision, recall and F1 score, which this paper can use to assess the value of the model, while the regression coefficients are calculated to assess the importance of each independent variable in predicting the occurrence of heart disease.

ISSN 2959-6157

3. Results and Discussion

3.1 Model Result Evaluation

The confusion matrix visualizes the classification effect of the model, which accurately predicted 46 non-diseased and 21 diseased individuals, 3 non-diseased individuals were misclassified, and 11 diseased individuals were missed. The overall accuracy of the model is 0.83, the macro-mean accuracy is 0.84, and the weighted mean accuracy is 0.83.

Firstly, these metrics show that for predicting non-diseased individuals (category 0), the model has a very high recall (0.94), indicating that the model recognizes the vast majority of non-diseased individuals, but the precision is relatively low (0.81). Secondly, for the prediction of diseased individuals (category 1), the precision is high (0.88) but the recall is low (0.66), indicating some degree of underreporting (i.e., failing to correctly identify cardiac patients). Thirdly, with an accuracy of 83%, the model had a balanced performance in terms of precision and recall and performed particularly well in predicting not disease cases. This analysis utilized logistic regression modeling to assess the impact of various physiological and medical predictors on the likelihood of heart disease. The model equation is expressed as:

$$logit\left(p\right)=-0.0076\times Age+1.1092\times Sex+\ldots+0.3222\times Thallium \quad (1)$$

In this equation, the presence of heart disease (binary outcome variable: 0 for absence and 1 for presence) is the outcome variable, and the predictor variables include age, sex, type of chest pain, blood pressure, cholesterol level, fasting glucose over 120 mg/dL, electrocardiogram findings, maximum heart rate, exercise-induced angina, ST-segment depression, ST-segment slope, number of blood vessels visible by fluoroscopy, and thallium scanning results.

Each coefficient in the model quantifies the effect of a one-unit increase in one of the predictor variables on the prevalence of heart disease while controlling for other variables to remain constant. For example, the coefficient for sex is 1.109213, which suggests that transitioning from the reference group (usually female) to the other group (usually male) increases the log odds of developing heart disease by approximately 1.109213, holding other factors constant. In addition, negative coefficients, such as -0.541668 for fasting glucose above 120 mg/dL, suggest that the higher the predictive value of this predictor, the lower the probability.

The validity of the model was assessed using statistical metrics including confusion matrix, accuracy, recall, and F1 score, indicating a relatively high level of accuracy and reliability in the prediction of heart disease.

3.2 Paired Regression Analysis

In this section, the paper used binary logistic regression modeling to assess the direct relationship between multiple heart disease risk factors and heart disease incidence. By analyzing the data-complete sample, all risk factors were significantly associated with heart disease incidence, except fasting blood sugar over 120 mg/dL (FBS over 120) and electrocardiogram results (EKG results). Specifically, age, gender, type of chest pain, blood pressure, cholesterol level, maximum heart rate, exercise-induced angina, ST-segment depression, ST-segment slope, number of visualized blood vessels (number of blood vessels fluro), and thallium test results were all significantly associated with the incidence of heart disease. Logistic regression analysis showed statistically significant regression coefficients and narrow confidence intervals for these variables, with high predictive value.

The results of this study demonstrate significant differences in the importance of individual factors in assessing and managing heart disease risk. While most of the factors examined were significantly correlated with the development of heart disease, fasting blood glucose and electrocardiogram results did not show similar associations, suggesting the need for further research into the value of these factors in the assessment of heart disease risk. In addition, the findings provide valuable information for clinicians to identify and predict patients with heart disease.

This section presents a visualization of the relationship between various influencing factors and the incidence of heart disease, utilizing Python for analysis. The resulting graphs depict the associations between several key variables and the probability of heart disease incidence, as determined through binary logistic regression analysis. In each graph, the blue shaded area represents the 95% confidence interval, which serves as an indicator of the reliability of the model predictions. The horizontal axis of the plots delineates the factors influencing heart disease, while the vertical axis reflects the probability of heart disease onset.

To facilitate a comprehensive examination of these relationships, the correlation plots are categorized into three distinct groups. The first category consists of plots characterized by narrower confidence intervals, which indicate high accuracy in model predictions and hold significant clinical relevance. These plots assist healthcare professionals in better understanding and assessing the risk factors associated with heart disease. The second category encompasses plots with moderate confidence intervals, which, while still meaningful, suggest a greater degree of uncertainty. Finally, the third category includes plots with wider confidence intervals, indicating poorer predictive performance. This systematic approach to categorizing the plots allows for a nuanced interpretation of the data, thereby enhancing researchers' understanding of the multifaceted nature of heart disease risk factors. Initially, this part will analyze the six plots characterized by narrow confidence intervals in this section.

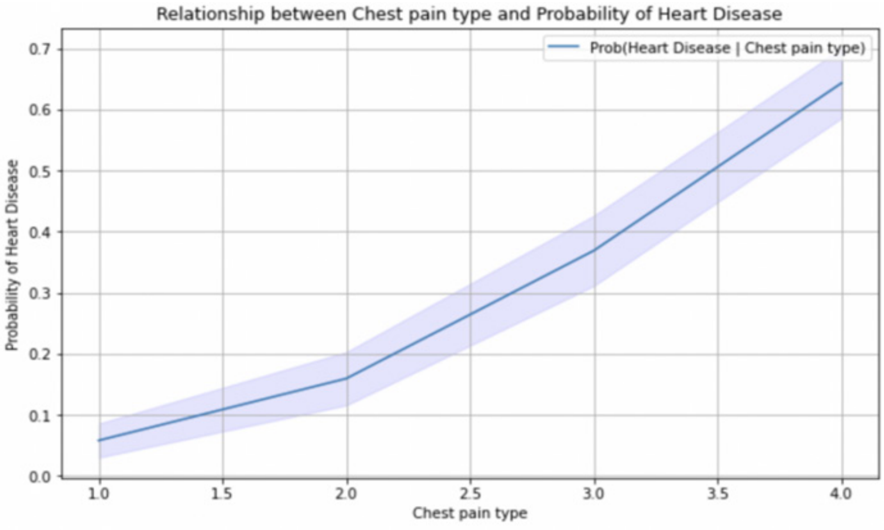


Fig. 1 Chest pain type versus Probability of heart disease

Figure 1 illustrates a positive relationship between the severity of chest pain and the probability of heart disease. Coronary artery stenosis or blockage is a common manifestation of heart disease, leading to insufficient oxygen supply to the myocardial tissue, which subsequently induces chest pain. This finding underscores the significance of chest pain type as a critical predictor of heart disease risk; more severe chest pain is often indicative of a higher likelihood of underlying cardiac pathology.

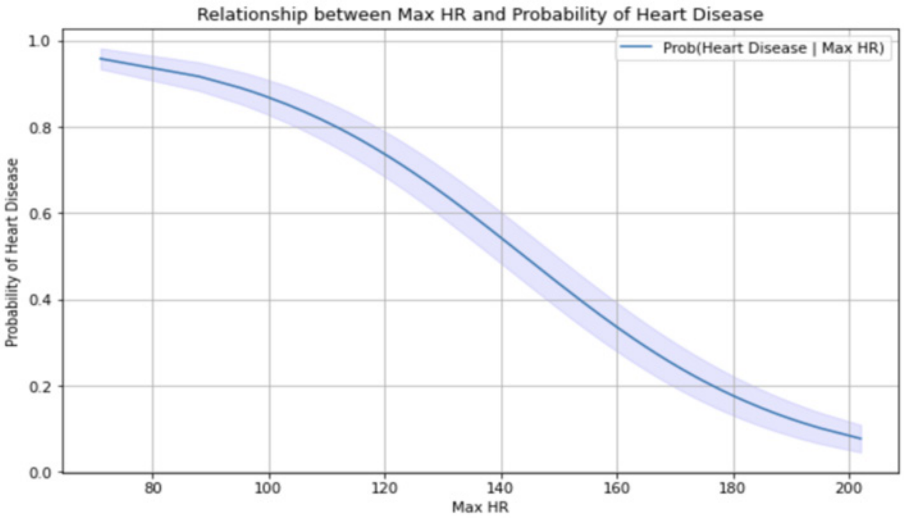


Fig. 2 Max HR versus Probability of heart disease

Figure 2 depicting maximum heart rate reveals a non-linear relationship with the probability of heart disease. It shows that as the maximum heart rate increases, the probability of heart disease decreases, indicating an overall negative correlation. A higher maximum heart rate typically signifies better myocardial function and contractility, suggesting that individuals with higher maximum heart rates exhibit a lower incidence of heart disease.

ISSN 2959-6157

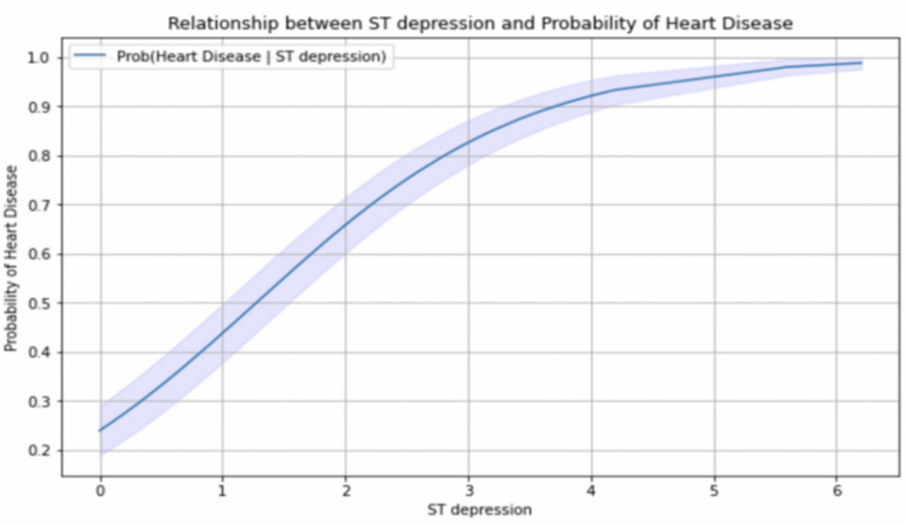


Fig. 3 ST depression versus Probability of heart disease

Figure 3 shows that ST-segment depression is positively associated with heart disease, as it often reflects myocardial ischemia. The greater the degree of ST-segment depression, the more pronounced the myocardial ischemia, which increases the cardiac workload and elevates the risk of cardiovascular events. Consequently, a higher degree of ST-segment depression correlates with an increased risk of myocardial damage and, thus, a greater incidence of heart disease.

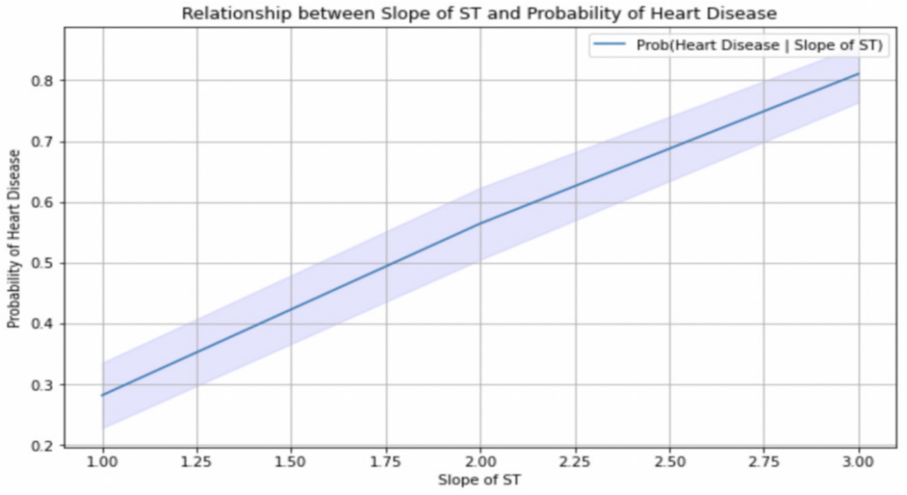


Fig. 4 Slope of ST versus Probability of heart disease

Figure 4 shows ST slope with the probability of heart disease which also demonstrates a positive correlation. The ST slope serves a similar diagnostic function to ST-segment depression, providing further differentiation of myocardial ischemia characteristics. A steeper ST slope indicates a higher likelihood of ischemic events, reinforcing the association between increased slope and heart disease risk.

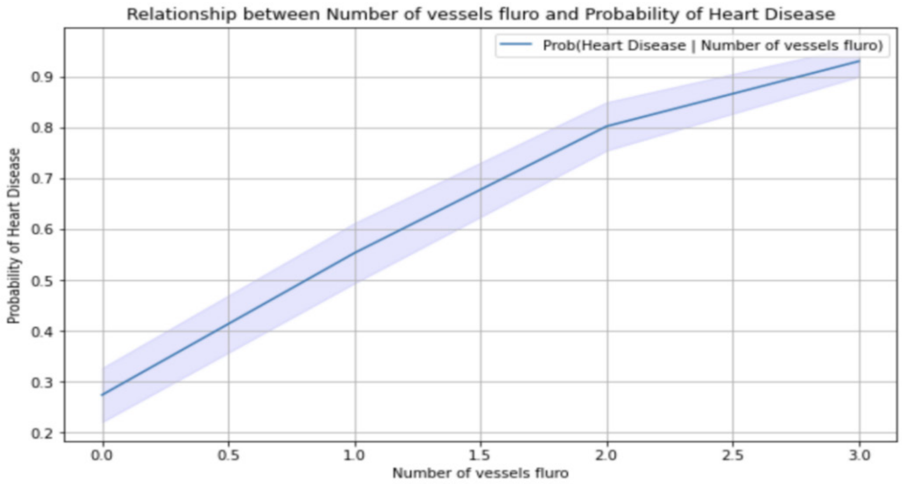


Fig. 5 Number of vessels fluro versus Probability of heart disease

Figure 5 renders the analysis of visible blood vessels indicating a positive correlation with heart disease probability. The term visible vessels typically refer to the number of diseased arteries identified through imaging techniques, such as coronary angiography. An increase in the number of affected vessels correlates with a heightened risk of heart disease, signifying that as the count of visibly diseased blood vessels rises, so does the incidence of heart disease.

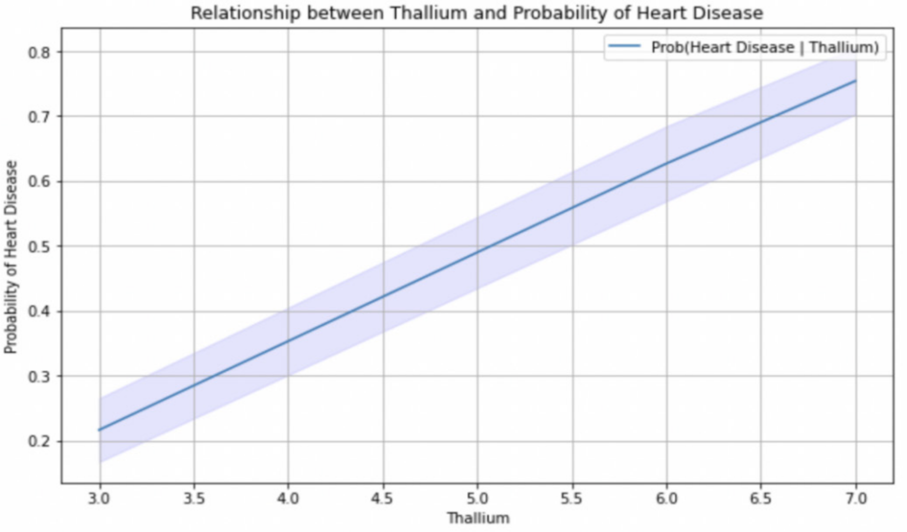


Fig. 6 Thallium versus Probability of heart disease

Figure 6 demonstrates the thallium test exhibits a positive correlation with the likelihood of heart disease. This nuclear medicine test assesses myocardial perfusion status; more significant defects in myocardial perfusion indicate inadequate blood supply to the heart. Therefore, a greater degree of abnormality in thallium test results corresponds to an elevated probability of heart disease. Next, this paper analyzed four graphs characterized by moderate confidence intervals.

ISSN 2959-6157

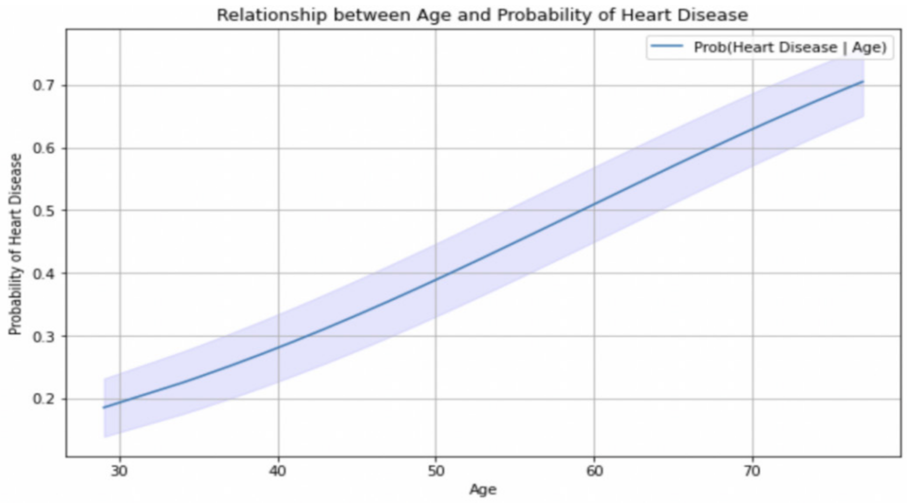


Fig. 7 Age versus Probability of heart disease

Figure 7 demonstrates a positive correlation between age and the incidence of heart disease. As individuals age, vascular aging becomes more pronounced, accompanied by an increased prevalence of conditions such as hypertension and diabetes. These complex factors contribute to a heightened likelihood of developing heart disease as one age.

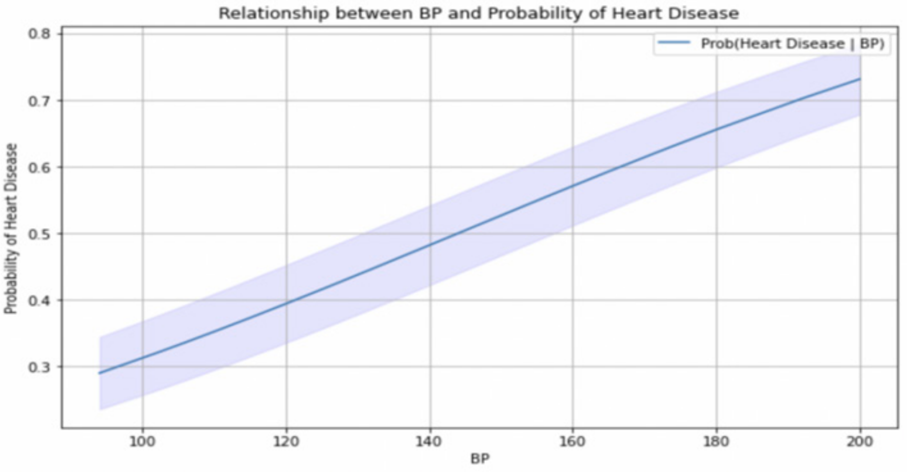


Fig. 8 BP versus Probability of heart disease

Figure 8 reveals a positive correlation between blood pressure and the probability of heart disease. Elevated blood pressure exerts additional strain on the heart and blood vessels, thereby increasing the risk of cardiovascular complications. Consequently, as blood pressure rises, the likelihood of developing heart disease also escalates.

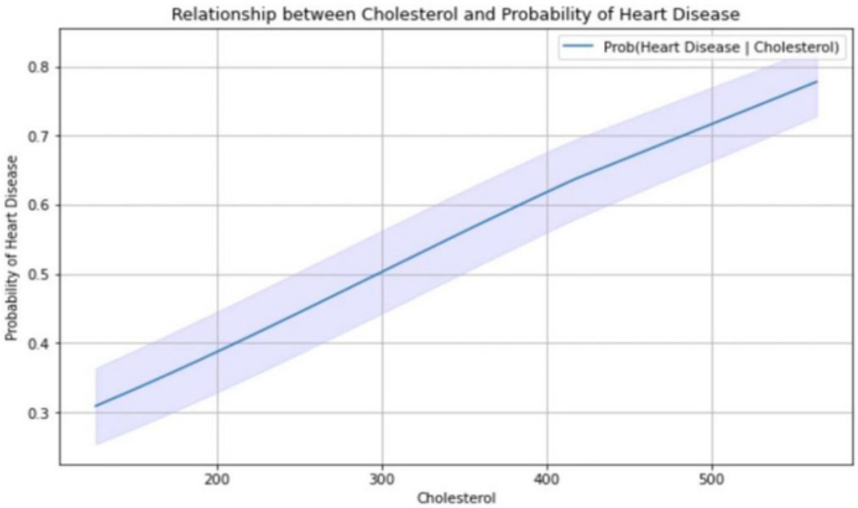


Fig. 9 Cholesterol versus Probability of heart disease

Figure 9 indicates a positive correlation between cholesterol levels and the incidence of heart disease. Elevated cholesterol levels, particularly low-density lipoprotein cholesterol (LDL-C), contribute to the narrowing and hardening of arterial walls, subsequently increasing the incidence of heart disease. As cholesterol levels rise, the probability of developing heart disease tends to increase significantly.

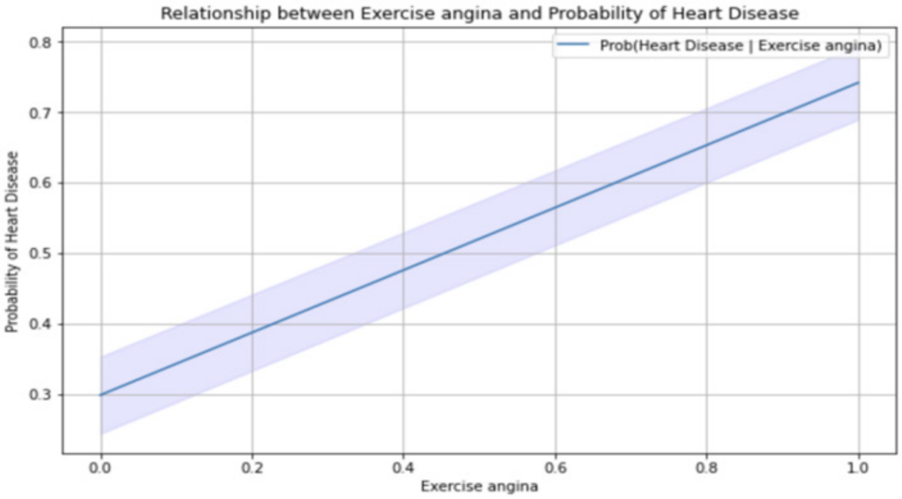


Fig. 10 Exercise angina versus Probability of heart disease

Figure 10 illustrates a positive correlation between exercise-induced angina and the risk of heart disease. Exercise angina typically manifests as myocardial ischemia resulting from the narrowing or blockage of coronary arteries, indicating that the heart is unable to receive adequate oxygen and blood supply during physical activity. Thus, as the severity of exercise angina increases, so too does the likelihood of developing heart disease.

Finally, this paper examined three graphs characterized by wider confidence intervals. While these wider intervals reflect a higher degree of uncertainty in the predictions, they still retain clinical relevance.

ISSN 2959-6157

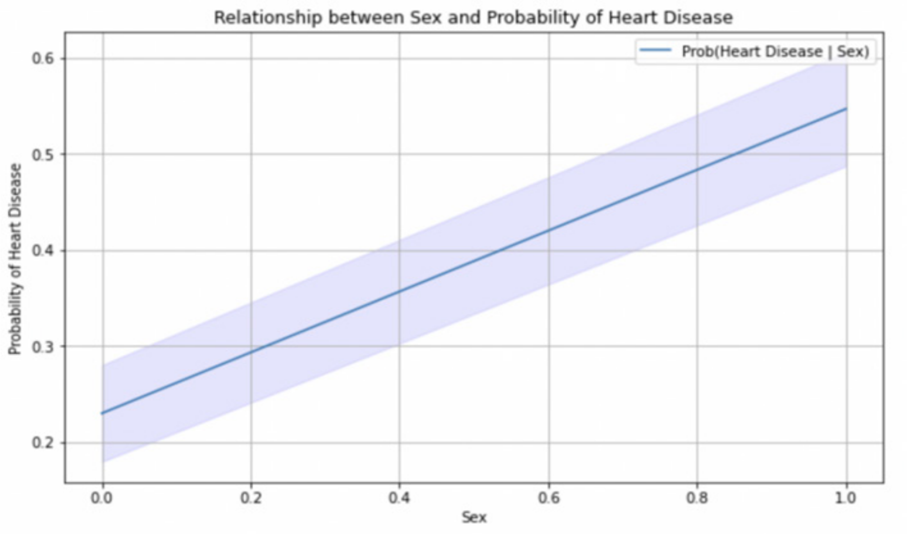


Fig. 11 Sex versus Probability of heart disease

Figure 11 reveals a positive correlation between sex and the incidence of heart disease (coded as 0 for women and 1 for men), illustrating that men exhibit a higher risk of heart disease compared to women. It is noteworthy that the relatively wide confidence intervals suggest some uncertainty in the heart disease risk estimates.

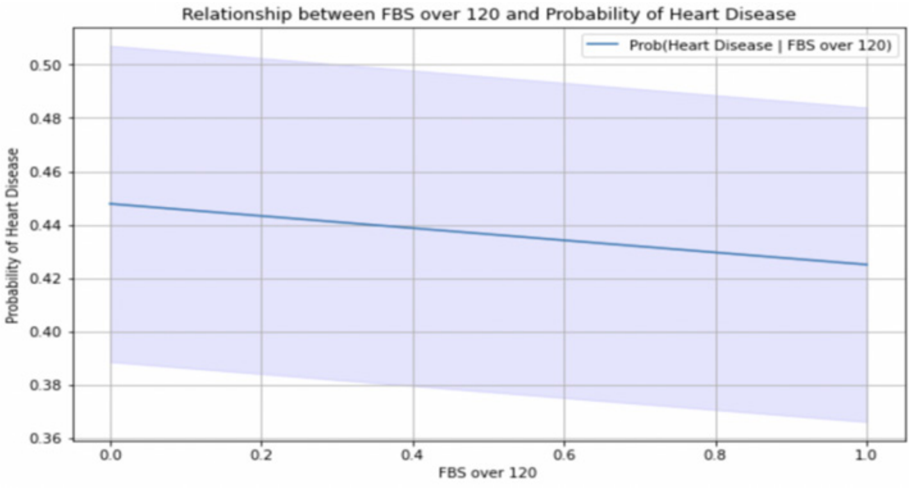


Fig. 12 FBS over 120 versus Probability of heart disease

Figure 12 shows that fasting blood glucose levels exceeding 120 mg/dL are inversely associated with the occurrence of heart disease, reflecting a slight decreasing trend. This indicates minimal changes in the likelihood of developing heart disease among individuals with fasting blood glucose levels above this threshold.

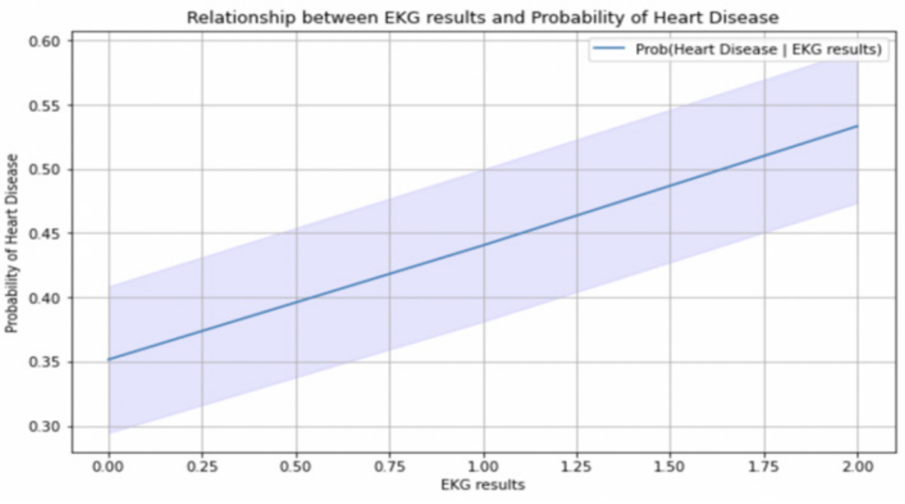


Fig. 13 EKG results versus Probability of heart disease

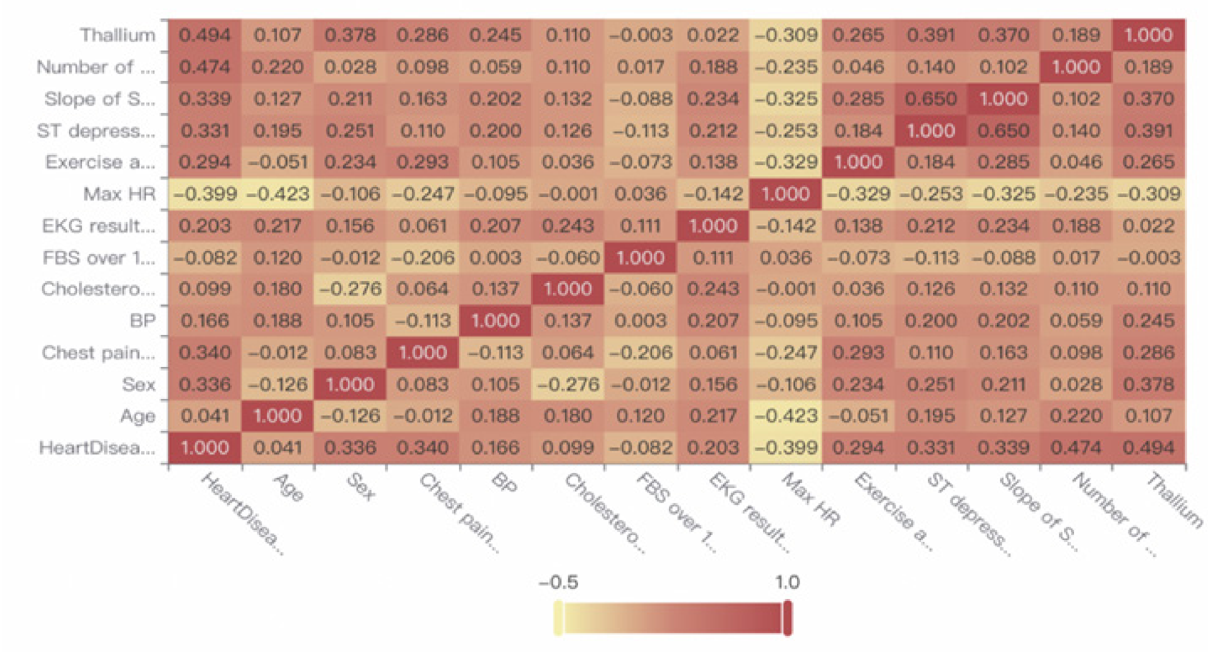
Finally, Figure 13 describes the relationship between EKG results and the probability of heart disease is positive, suggesting that abnormal EKG findings are associated with an increased risk of heart disease.

3.3 Discussion

In the model, this research included 13 prevailing risk factors for heart disease as independent variables and the presence of heart disease as the dependent variable to complete a multivariate logistic regression analysis. This paper then compared these variables to the incidence of heart disease to assess the value of each variable. Overall, all variables showed great value and the results of the multivariate regression analysis showed a high degree of goodness of fit.

The expectation for this research paper was to construct a methodology for predicting the incidence of heart disease based on the most reasonable pool of variables. Therefore, the most challenging part of the research process was the selection and optimization of variables. The variables initially selected are widely recognized. However, further optimization could certainly improve the overall value of the model. The following are the aspects to be optimized:

Correlation: This research requires each variable in the model to contribute independently to the risk of heart disease while maintaining the simplicity of the model. Therefore, correlation analysis is crucial.



	HeartDisea...	Age	Sex	Chest pain...	BP	Cholestero...	FBS over 1...	EKG result...	Max HR	Exercise a...	ST depress...	Slope of S...	Number of ...	Thallium
Thallium	0.494	0.107	0.378	0.286	0.245	0.110	−0.003	0.022	−0.309	0.265	0.391	0.370	0.189	1.000
Number of ...	0.474	0.220	0.028	0.098	0.059	0.110	0.017	0.188	−0.235	0.046	0.140	0.102	1.000	0.189
Slope of S...	0.339	0.127	0.211	0.163	0.202	0.132	−0.088	0.234	−0.325	0.285	0.650	1.000	0.102	0.370
ST depress...	0.331	0.195	0.251	0.110	0.200	0.126	−0.113	0.212	−0.253	0.184	1.000	0.650	0.140	0.391
Exercise a...	0.294	−0.051	0.234	0.293	0.105	0.036	−0.073	0.138	−0.329	1.000	0.184	0.285	0.046	0.265
Max HR	−0.399	−0.423	−0.106	−0.247	−0.095	−0.001	0.036	−0.142	1.000	−0.329	−0.253	−0.325	−0.235	−0.309
EKG result...	0.203	0.217	0.156	0.061	0.207	0.243	0.111	1.000	−0.142	0.138	0.212	0.234	0.188	0.022
FBS over 1...	−0.082	0.120	−0.012	−0.206	0.003	−0.060	1.000	0.111	0.036	−0.073	−0.113	−0.088	0.017	−0.003
Cholestero...	0.099	0.180	−0.276	0.064	0.137	1.000	−0.060	0.243	−0.001	0.036	0.126	0.132	0.110	0.110
BP	0.166	0.188	0.105	−0.113	1.000	0.137	0.003	0.207	−0.095	0.105	0.200	0.202	0.059	0.245
Chest pain...	0.340	−0.012	0.083	1.000	−0.113	0.064	−0.206	0.061	−0.247	0.293	0.110	0.163	0.098	0.286
Sex	0.336	−0.126	1.000	0.083	0.105	−0.276	−0.012	0.156	−0.106	0.234	0.251	0.211	0.028	0.378
Age	0.041	1.000	−0.126	−0.012	0.188	0.180	0.120	0.217	−0.423	−0.051	0.195	0.127	0.220	0.107
HeartDisea...	1.000	0.041	0.336	0.340	0.166	0.099	−0.082	0.203	−0.399	0.294	0.331	0.339	0.474	0.494

Fig. 14 Correlation Analysis

The results of Figure 14 which conduct the correlation analysis clearly show that there was no excessive correlation between these variables, indicating that the variable pool does not suffer from redundancy.

Quantification of variables: To improve the applicability and accuracy of the model, this paper quantified all categorical variables. This included translating qualitative variables such as the presence of certain symptoms and medical history into numerical indicators suitable for data analysis. However, there are still many other influences, such as genetic factors, dietary habits, and ethnicity type, that have a significant impact on heart health but are difficult to quantify and therefore difficult to analyze.

Timeliness of the model: Clinical medicine is a rapidly evolving field, and new research and data will continue to change the understanding of risk factors. Therefore, this paper plans to regularly review and update the models to incorporate the latest clinical data and research findings and to identify new cardiac influences, thus ensuring that the models are based on the most current medical knowledge and practice.

Model Diversity: While the paper has developed a comprehensive model to predict the incidence of heart disease, this paper recognized that different populations (e.g., different ages, genders, races, etc.) may require different models. Future studies will explore the possibility of creating multiple sub-models to more accurately serve specific groups.

Through these efforts, this research hopes to provide more accurate and practical tools for heart disease prediction and incidence risk control. This will help improve prevention strategies for heart disease and reduce morbidity and mortality due to heart disease.

4. Conclusion

The study evaluated the predictive power of each variable by integrating 13 recognized influences on the prevalence of heart disease into a multivariate logistic regression model to predict the incidence of heart disease. Meanwhile, conducting paired regression analysis to identify the independent significance of each variable. The independent variables included a wide range of physiologic and medical indicators that provided a rigorous and comprehensive assessment of an individual’s heart health. The results showed a high degree of model fit, suggesting that these factors collectively provide strong support for the prediction of heart disease. Notably, completed correlation analyses examined the independence of the variables, confirming that the model was not redundant and therefore each variable was essential to the completeness and accuracy of the predictive model.

Despite the high predictive power of the established variables, challenges remain in quantifying complex multifactorial influences, such as genetic predisposition, dietary patterns, and socioeconomic factors, which are critical but difficult to quantify. With these issues in mind, the models are regularly updated and incorporate the latest research findings and clinical data to ensure that they reflect current scientific understanding and clinical practice, thereby improving their timeliness. In addition, given the diversity of the population, future research will involve sub-models for specific groups to improve the accuracy and applicability of predictions. These models will be able to cater to risk profiles associated with different ages, genders, and ethnicities, thus refining disease prediction efficiency.

In summary, this study not only reinforces the importance of established cardiac risk factors but also emphasizes the dynamic nature of model development in cardiology - a field that requires constant adaptation and enhancement. By continually refining the models, these analyses are committed to improving heart disease prevention strategies and thereby significantly reducing its morbidity and mortality.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Nowbar A N, Gitto M, Howard J P, et al. Mortality from ischemic heart disease: Analysis of data from the World Health Organization and coronary artery disease risk factors From NCD Risk Factor Collaboration. Circulation: cardiovascular quality and outcomes, 2019, 12(6): e005375.

[2] Mensah G A, Brown D W, Croft J B, et al. Major coronary risk factors and death from coronary heart disease: baseline and follow-up mortality data from the Second National Health and Nutrition Examination Survey (NHANES II). American journal of preventive medicine, 2005, 29(5): 68-74.

[3] Lüscher T F. Prevention: some important steps forward, but many unmet needs in a world with cardiovascular disease as the leading cause of death. European Heart Journal, 2016, 37(42): 3179-3181.

[4] Choi D, Hwang K C, Lee K Y, et al. Ischemic heart diseases: current treatments and future. Journal of controlled release, 2009, 140(3): 194-202.

[5] Chaurasia V, Pal S. Early prediction of heart diseases using data mining techniques. Caribbean Journal of Sciences and Technology, 2013, 1(1): 208-217.

[6] Moyer V A, US Preventive Services Task Force. Screening for coronary heart disease with electrocardiography: US Preventive Services Task Force recommendation statement. Annals of internal medicine, 2012, 157(7): 512-518.

[7] Grossman E, Messerli F H. Diabetic and hypertensive heart disease. Annals of internal medicine, 1996, 125(4): 304-310.

[8] Tikkanen E, Gustafsson S, Ingelsson E. Associations of fitness, physical activity, strength, and genetic risk with cardiovascular disease: longitudinal analyses in the UK Biobank Study. Circulation, 2018, 137(24): 2583-2591.

[9] Wu W, He J, Shao X. Incidence and mortality trend of congenital heart disease at the global, regional, and national level, 1990–2017. Medicine, 2020, 99(23): e20593.

[10] Williams P T. Physical fitness and activity as separate heart disease risk factors: a meta-analysis. Medicine and science in sports and exercise, 2001, 33(5): 754.

[11] Diwakar M, Tripathi A, Joshi K, et al. Latest trends on heart disease prediction using machine learning and image fusion. Materials today: proceedings, 2021, 37: 3213-3218.