Dean&Francis

ISSN 2959-6157

Research on the Relationship between Prevalence of Cardiovascular Disease and Behavioral Risk Factors in Adults

Zirui Yao[1, *]

[1]College of Pharmacy, Shenzhen Technology University, Shenzhen, 515100, China

*Corresponding author: 202201101119@stumail.sztu.edu.cn

Abstract:

The cardiovascular system is an significant system which transports various metabolic essential substances to the body and excretes respiratory metabolites from the body to maintain the basic needs of the human. Cardiovascular disease is a widespread disease with a high mortality rate, which poses a great threat to the health of the population. This study analyzed data from the Behavioral Risk Factor Monitoring System (BRFSS) to build a predictive model to assess which behavioral factors are strongly associated with cardiovascular disease. The data included 13,210 clinical samples and 11 variables such as Exercise, Smoking-History. The target variable is whether the patient has been diagnosed with heart disease. In this study, a logistic regression model was used to process the data of BRFSS. It was concluded that the incidence of cardiovascular disease was positively associated with age, history of smoking, having other cancers and having depression, while fruit-consumption and exercise had an inhibitory effect on the incidence of cardiovascular disease. The accuracy of the training set was about 95.95% and the prediction results of the model were found to be acceptable after fitting. In order to better predict the association of heart disease with behavioral factors, more comprehensive clinical data and more advanced analytical techniques are needed. This model provided assistance for studying the induction of heart disease by behavioral factors.

Keywords: Cardiovascular disease; logistic regression; prediction model; behavioral risk factor.

1. Introduction

The cardiovascular system is a circulatory system composed of the heart and blood vessels, which transports various metabolic essential substances to the body and excretes respiratory metabolites from the body [1]. Because the cardiovascular system is crucial to preserving the human body's regular op-

eration and blood flow, researchers have looked at it in great detail. The most recent statistics available from the World Health Organization indicated that cardiovascular diseases (CVD), a group of illnesses that widely distributed around the world, were the world's leading cause of death, with an estimated 17.9 million fatalities each year [2]. In other related studies, Predicting cardiovascular disease (coronary heart disease and stroke) in adults aged 35-84 years from 2010 to 2030 using the Markov Computer Model-China. With the same level of risk factors, a 50% increase in the cardiovascular events per year was projected between 2010 and 2030 based solely on population ageing and growth [3]. Cardiovascular diseases are widely distributed worldwide, with high morbidity and mortality, and are an urgent public health problem that affects everyone. Therefore, finding the risk factors for the onset of cardiovascular disease and discovering the degree of effect of each factor on cardiovascular disease is an effective measure to help the public prevent the disease.

A study conducted by Aljefree indicated that there were two groups of risk factors for coronary heart disease and stroke, behavioral risk factors (diet, physical activities, and smoking) and metabolic risk factors (diabetes, hypertension, obesity, and dyslipidemia) [4]. For the metabolic risk factors, the researchers found that hypertension was the most significant risk factor for all stroke subgroups [5]. Yusuf conducted a structured questionnaire survey, which utilized logistic regression analysis on the risk factors (hypertension, diabetes) and social-psychological factors (depression, control points, perceived stress and life events) in the case and control groups. A factor for intracerebral hemorrhagic stroke has more potent risk than ischemic stroke and particularly significant in people of 45 years or younger and Social-Psychological factors, abdominal obesity, diabetes, were the equal leading hazard factors in people [6].

In terms of the cognition and prevention of cardiovascular diseases, most people pay more attention to various metabolic indicators such as hypertension, obesity, diabetes and dyslipidemia. However, people's behavioral factors (diet, smoking, physical activity) are more likely to determine the trend of various metabolic indicators in daily life. Diet is one of the most widely differentiated behavioral factors. The correlations between intake of a variety of vegetarian dishes and the prevalence of cardiovascular disease events and death were examined using Cox frailty models with random factors. According to Miller's age and sex adjusted analyses, more vegetarian dishes intake was linked to a decreased prevalence of morbidity and mortality in major cardiovascular diseases, as well as in various branches of this type of disease [7]. In addition to behavioral factors such as diet, Smoking is a behavioral factor that is widely considered profitless to the body, and researchers have found a relationship between smoking behavior and cardiovascular disease. Cigarette smoke exposure (CSE) might show up as a clinical symptom anywhere in the range of CVD. The oxygen demand of cardiac muscle is increased by CSE, which delays the onset of angina. Cigarette smokers have multiple times the chance of having Myocardial ischemia and hypoxia as compared to nonsmokers. Both past and present smoking can raise the probability of heart abnormal pumping by up to 33% to 93% when compared to lifetime nonsmokers [8]. One of the major global causes of acute myocardial infarction (AMI), particularly among men, is tobacco abuse. Reducing tobacco use can be an effective prevention of these diseases [9]. Physical activity is a distinguished behavioral factor in the daily life. According to new research and recommendations, developing an effective exercise program and putting it into practice can be effective in preventing all types of CVD [10].

This article utilized the correlation between behavioral factors (diet, smoking, physical activity) and the prevalence of cardiovascular diseases. On this basis, this article utilized logistic regression analysis to find out the preventive and induced behaviors among behavioral factors. This can improve the public's awareness of the correlation between behavioral factors and cardiovascular diseases, promote individual behaviors for effective prevention.

2. Methods

2.1 Data Source

The data for this article came from the Kaggle website, compiled by Alphiree in The Behavioral Risk Factor Surveillance System (BRFSS), which selected 19 behavioral factor variables that may contribute to cardiovascular disease from 304 unique variables, and was updated and published in 2023. There are no missing values in this database.

2.2 Variable Selection

The database in this article had a sample of 308856 individuals, and 78% of the participants reported having undergone a health screening within the past 1 year. This article used data from the population who have undergone health check-ups in the past year, with a total of 239371 people, including people with and without cardiovascular disease, 109443 males and 129928 females. The data excluded people with diabetes, skin cancer, and Arthritis. In this paper, a random sample of six age groups was conducted, and the final sample size was 13210. The sample

ISSN 2959-6157

population ranged in age from 18 to 80 years old. The data consisted of 11 variables (Exercise, Gender, Age-Category, Smoking-History, Alcohol-Consumption, Fruit-Consumption, Green-Vegetables-Consumption, Other-Cancer, Fried Potato-Consumption, Depression, General-Health). In this article, the gender variable was treated as a dummy variable, and the other variables were shown in Table 1.

Table 1. Different types of variables

Logogram	Term	Type	Range
x_1	Age-Category	Categorical	0- 18~29, 1- 30~39, 2- 40~49, 3- 50~59, 4- 60~69, 5- 70~79
x_2	Gender	Categorical	Female , male
x_3	General-Health	Categorical	0- Excellent, 1- Very Good, 2- Good, 3- Fair, 4- Poor
x_4	Smoking-History	Categorical	0- false, 1- true
x_5	Alcohol-Consumption	Numeric	0 to 30 oz.av/day
x_6	Fruit-Consumption	Numeric	0 to 120 g/day
x_7	Fried Potato-Consumption	Numeric	0 to 128 g/day
x_8	Green-Vegetables-Consumption	Numeric	0 to 128 g/day
x_9	Exercise	Categorical	0- false, 1- true
x_{10}	Other-Cancer	Categorical	0- false, 1- true
x_{11}	Depression	Categorical	0-false, 1- true
Y	Heart-Disease	Categorical	0- false, 1- true

2.3 Research Protocol

The relationship between the effects of X and Y, and the association of 10 factors on cardiovascular disease, was examined in this article using a logistic regression model. The dependent variable (Y) in this model was the presence or absence of heart disease, while the independent variable (X) was the 10 factors. A logistic regression model that predicted the results of the diagnostic process was fitted using each variable. The expression for a logistic regression model is:

$$P = \frac{exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_m x_m\right)}{1 + exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_m x_m\right)} \quad (1)$$

In this formula, β_0 is constant term, $\beta_1, \beta_2, \cdots, \beta_m$ are partial regression coefficients. Perform a logit transformation on $f(x) = \frac{1}{1+e^{-x}}$, then $L(p) = ln\frac{p}{1-p}$. So, the logistic regression model can be registered in the following linear form:

$$ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m \quad (2)$$

3. Results and Discussion

3.1 Descriptive Analysis

In order to visually understand their numerical variable, this article used a histogram to describe 10 behavioral factors that may affect the prevalence of cardiovascular disease, of which 4 factors were Numeric variables with very different distributions.

The figure 1 was the histogram of the Alcohol-Consumption(x_5) level of the sample. x_5 was most frequent in the 0-5 oz.av/day range. High blood pressure was common in patients. So this was a skewed distribution.

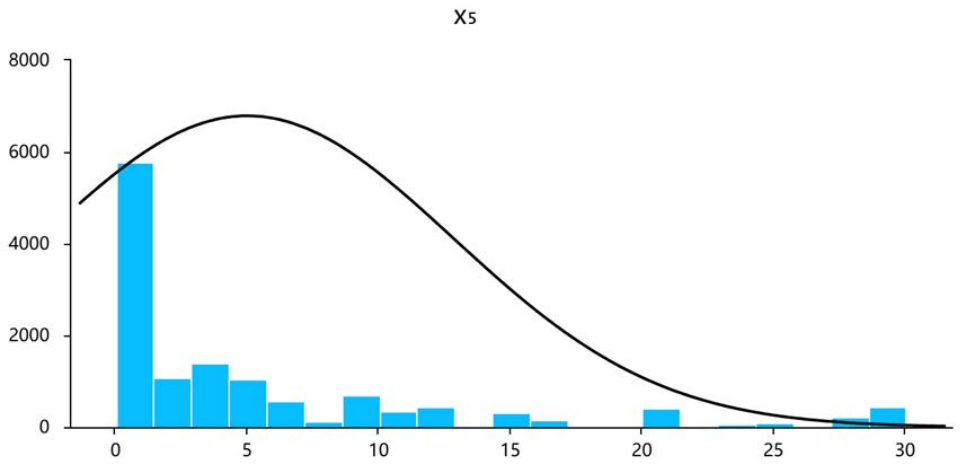

Fig. 1 The histogram of Alcohol-Consumption (x_5)

The figure 2 was the histogram of the Fruit-Consumption (x_6) level of the sample. Most of the intake levels of x_6 were in the range of 0 to 40 g/day, and there was a high frequency of 60 g/day.

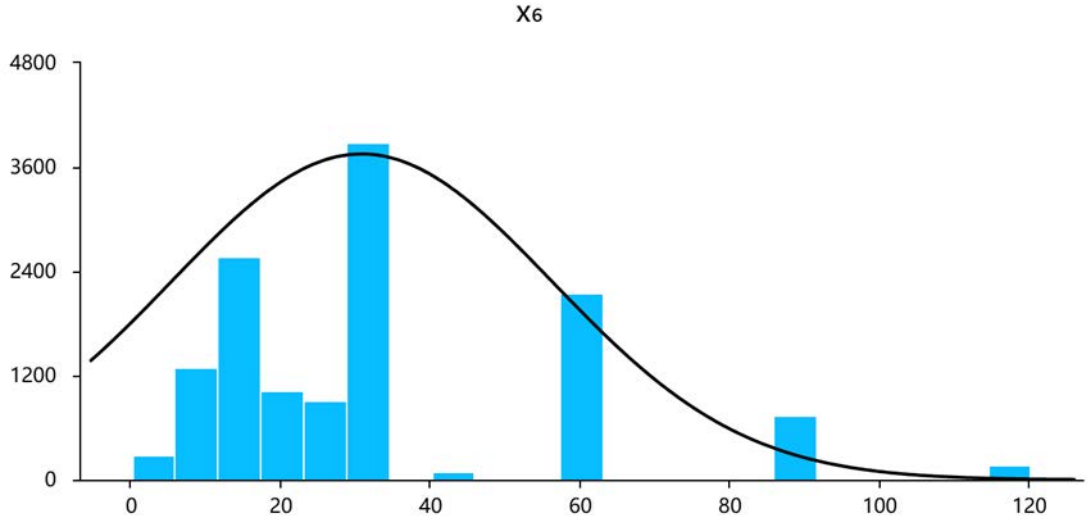

Fig. 2 The histogram of Fruit-Consumption(x_6)

The figure 3 was the histogram of the Fried Potato-Consumption (x_7) level of the sample. For the x_7, most of the samples were located at 0-20g/day, which was at a low consumption, and the data was approximately normally distributed.

ISSN 2959-6157

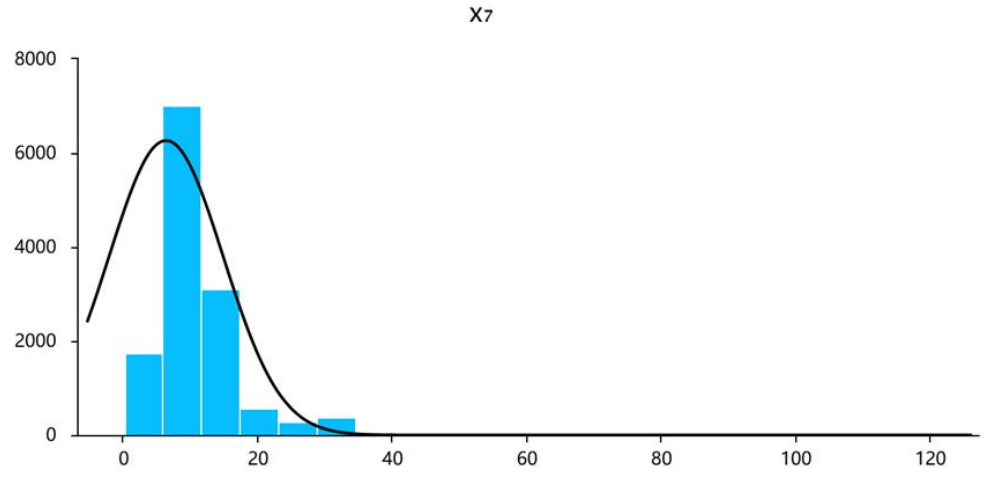

Fig. 3 The histogram of Fried Potato-Consumption (x_7)

The figure 4 was the histogram of the Green-Vegetables-Consumption (x_8) level of the sample. Most of the data for X were located at 0-20g/day, indicating that the consumption of vegetables in the sample was relatively balanced.

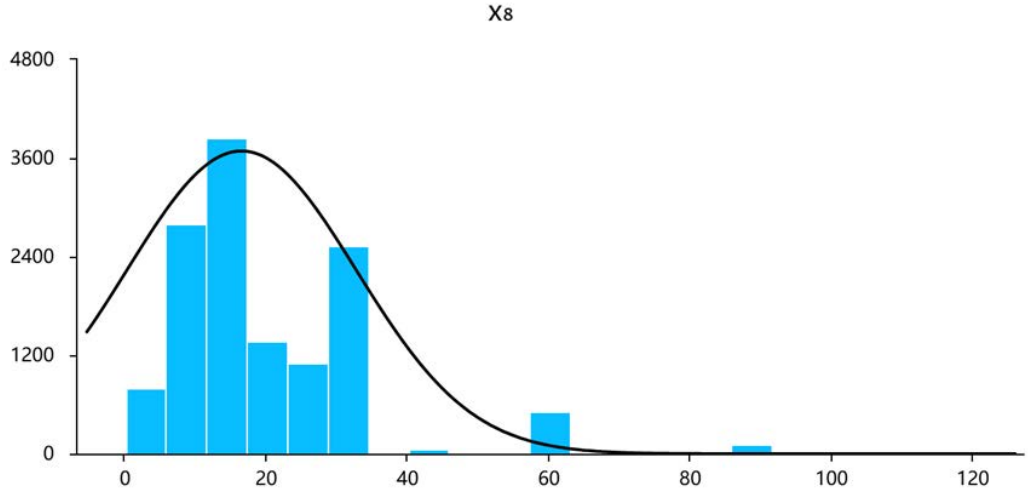

Fig. 4 The histogram of Green-Vegetables-Consumption (x_8)

3.2 Exploratory Analysis

As shown in Table 1, there were 4 continuous predictors and 7 categorical predictors in the database. Equilibrium was reached after random sampling. In this database, the four continuous variables ($x_5 x_6 x_7 x_8$) were all behavioral factors such as diet. In order to prevent the four variables of the same type of behavioral factors from having a strong correlation and causing a large bias to the research results, it was necessary to use the Pearson correlation to determine the linear correlation degree of these variables before fitting the model.

The definition of Pearson correlation is the following: the covariance (X, Y) between two continuous variables divided by the product of their respective standard deviations ($\sigma X, \sigma Y$) yields the Pearson correlation coefficient $\rho_{X,Y}$ of those variables. The coefficient's value is always in the range of -1.0 to 1.0, with variables around 0 being thought to be uncorrelated and those near 1 or -1 to be significantly correlated. The following is its formula:

$$\rho_{X,Y} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{N}\right)}} \quad (3)$$

In this paper, the above four variables were visualized by using the heat map, and the correlation strength of the

variables was judged by combining the value range of Pearson correlation. Thus Figure 5 and Table 2 were obtained.

Table 2. Linear magnitude relationship of Pearson correlation

Strength of Association	Coefficient, r	
	Positive	Negative
Small	0.1 to 0.3	-0.1 to -0.3
Medium	0.3 to 0.5	-0.3 to -0.5
Large	0.5 to 1.0	-0.5 to -1.0

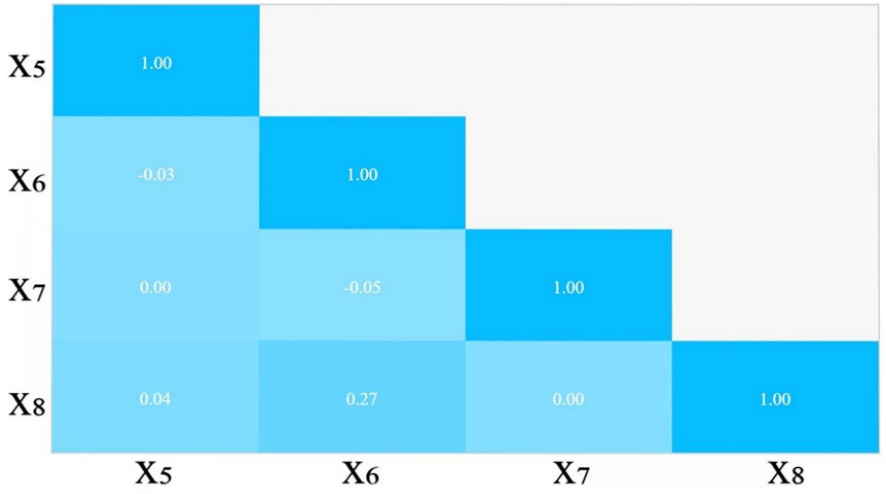

Fig. 5 Correlation plot among continuous variables

As shown in Figure 5 and Table 2, there was no strong positive correlation between these continuous variables, suggesting that there was no collinearity between these four variables. Therefore, these four variables did not interact with each other to produce large biases and can appear in the same model.

3.3 Logistic Regression Results

For binary Logit regression analysis, the independent variables were $x_1, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}$ and the dependent variable was Y. The independent variable screening method used was a stepwise, step-by-step approach. Once the model was automatically identified, the final remaining $x_1, x_4, x_6, x_9, x_{10}, x_{11}$ were a total of 6 items in the model.

As the above table shows, 13210 samples were included in the analysis, and no missing data was found. Since the p-value in Table 3 was less than 0.05, the model was considered viable.

Table 3. The results of likelihood ratio test of the binary Logit regression model

Model	-2x log-likelihood	Chi-square	df	p	AIC value	BIC value
Intercept only	4478.949					
Final model	3916.088	562.861	6	0.000	3930.088	3982.509

Table 4. Summary of binary Logit regression analysis results

item	Regression coefficients	standard error	Z value	Wald $\chi 2$	p value	OR value	OR value 95% CI
x_1	0.658	0.038	17.152	294.189	0.000	1.930	1.790 ~ 2.081
x_4	0.520	0.092	5.628	31.670	0.000	1.683	1.404 ~ 2.017

ISSN 2959-6157

item	Regression coefficients	standard error	Z value	Wald χ2	p value	OR value	OR value 95% CI
x_6	-0.006	0.002	-3.132	9.811	0.002	0.994	0.990 ~ 0.998
x_9	-0.520	0.101	-5.122	26.239	0.000	0.595	0.487 ~ 0.725
x_{10}	0.284	0.137	2.064	4.259	0.039	1.328	1.014 ~ 1.739
x_{11}	0.391	0.116	3.363	11.311	0.001	1.479	1.177 ~ 1.858
Intercept	-5.704	0.216	-26.376	695.703	0.000	0.003	0.002 ~ 0.005
Note: Dependent variable = Y							
McFadden R^2 = 0.126							

From the table 4, it can be seen that the formula of the model was:

$$ln\frac{p}{1-p} = -5.704 + 0.658x_1 + 0.52x_4 - 0.006x_6 - 0.52x_9 + 0.284x_{10} + 0.391x_{11} \quad (4)$$

The results show that $x_1x_4x_{10}x_{11}$ have a significant positive impact on Y, and x_6x_9 have a significant negative impact on Y. In table 5, as shown, this training set model was about 95.95% accuracy rate. There was a problem of sample difference in true value, where the number of people with heart disease and the number of people without heart disease exceed the 10% range. As a result, the sample with heart disease has 0.00% predictive accuracy.

Table 5. Training set model evaluation results

0		Predicted value		Prediction accuracy	Prediction error rate
		1			
True value	0	12675	0	100.00%	0.00%
	1	535	0	0.00%	100.00%
Summary				95.95%	4.05%

The fitted model underwent testing. The Hosmer-Lemeshow fit test was utilized in table 6 to assess the model's goodness of fit. The model test initial presumption was that the observed value and the model's fit would coincide. Table 6 showed that he p-value was higher than 0.05 (Chi=6.016, p=0.645>0.05), indicating that the first hypothesis-that was, the model good goodness of fit and passing the HL test was accepted.

Table 6. Testing set model evaluation results

Hosmer-Lemeshow fit test		
χ2	Degree of Freedom df	p value
6.016	8	0.645

4. Conclusion

In this article, the step-by-step method was used to select several behavioral factors that have a greater impact on cardiovascular disease as predictors, and the logistic regression was used to fit the statistical model. This study shows that the Increasing Age, Smoking-History, Other-Cancer, Depression have a positive stimulating effect on cardiovascular disease, and Fruit-Consumption, Exercise have an inhibitory effect. The accuracy of both the test set and the training set was above 90%. Enhance public awareness of the behavioral factors that predispose to cardiovascular disease, and increase fruit intake and exercise time in daily life. This can better help the public to prevent or improve the harm caused by cardiovascular diseases.

However, there are some limitations to this study. Data imbalances that occur during data acquisition can affect the reliability of the results. In the future, research needs to collect more clinical data and include more behavioral factors in the analysis. It will also use more advanced models to analyze potential interactions and nonlinear effects between variables.

References

[1] Oliver Michael Francis, Entman, Mark L, Jacob Stanley W. human cardiovascular system. Encyclopedia Britannica, 2024,

[2] World Health Organization. Cardiovascular diseases (CVDs). June 11, 2021. Retrieved on August 8, 2024. Retrieved from: Cardiovascular diseases (CVDs) (who.int)

[3] Moran A, et al. Future cardiovascular disease in China Markov model and risk factor scenario projections from the coronary heart disease Policy Model-China. Circulation: Cardiovascular Quality and Outcomes, 2010, 243-252.

[4] Aljefree Najlaa, Ahmed Faruk. Prevalence of Cardiovascular Disease and Associated Risk Factors among Adult Population in the Gulf Region: A Systematic Review, Advances in Public Health, 2015, 235101.

[5] O'Donnell, Martin J et al. Risk factors for ischemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. Lancet (London, England), 2010, 3: 112-123.

[6] Yusuf Salim et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet (London, England), 2004, 937-952.

[7] Miller Victoria et al. Fruit, vegetable, and legume intake, and cardiovascular disease and deaths in 18 countries (PURE): a prospective cohort study. Lancet (London, England), 2017, 2037-2049.

[8] Pamela Morris, et al. Cardiovascular Effects of Exposure to Cigarette Smoke and Electronic Cigarettes. Journal of the American College of Cardiology, 2015, 66: 1378-1391.

[9] Teo Koon K et al. Tobacco use and risk of myocardial infarction in 52 countries in the INTERHEART study: a case-control study. Lancet (London, England), 2006, 647-658.

[10] Vazquez-Guajardo Mauricio, et al. Exercise as a Therapeutic Tool in Age-Related Frailty and Cardiovascular Disease: Challenges and Strategies. The Canadian journal of cardiology, 2024, 1458-1467.