Application of Statistical Learning Techniques for Cardiovascular Disease Prediction

S.Gandhiya Vendhan and K.Elsa

Department of Statistics, Bharathiar University, Coimbatore, India.

DOI: https://doi.org/10.51584/IJRIAS.2026.11100106

Received: 31 October 2026; Accepted: 05 November 2026; Published: 17 November 2026

ABSTRACT

Cardiovascular disease and heart disease are most often used interchangeably. The human heart is only the size of a fist, but it is the strongest muscle in the human body. The Societies with cardiovascular disease or who are at high cardiovascular risk due to the presence of one or more risk factors with statistical applications.

Keywords: Cardiovascular Disease, Data Visualization, Biostatistics, Data Integration and Health Informatics.

INTRODUCTION

Disease, any harmful deviation from the normal structural or functional state of an organism, generally associated with certain signs and symptoms and differing in nature from physical injury. A diseased organism commonly exhibits signs or symptoms indicative of its abnormal state. Thus, the normal condition of an organism must be understood in order to recognize the hallmarks of disease. Nevertheless, a sharp demarcation between disease and health is not always apparent. Cardiovascular disease (CVD) is a class of diseases that involve the heart or blood vessels. CVD includes coronary artery diseases (CAD) such as angina and myocardial infarction (commonly known as a heart attack). Other CVDs include stroke, heart failure, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, abnormal heart rhythms, congenital heart disease, valvular heart disease, carditis, aortic aneurysms, peripheral heart disease, thromboembotic disease, and venous thrombosis. The underlying mechanisms vary depending on the disease. Coronary artery disease, stroke, and peripheral artery disease involve atherosclerrosis. This may be caused by high blood pressure, smoking, diabetes mellitus, lack of exercise, obesity, high blood cholesterol, poor diet, and excessive alcohol consumption, among others. High blood pressure is estimated to account for approximately 13% of CVD deaths, while tobacco accounts for 9%, diabetes 6%, lack of exercise 6% and obesity 5%. Rheumatic heart disease may follow untreated strep throat. It is estimated that up to 90% of CVD may be preventable. Prevention of CVD involves improving risk factors through: healthy eating, exercise, avoidance of tobacco smoke and limiting alcohol intake. Treating risk factors, such as high blood pressure, blood lipids and diabetes is also beneficial. Treating people who have strep throat with antibiotics can decrease the risk of rheumatic heart disease. The use of aspirin in people, who are otherwise healthy, is of unclear benefit. Heart attacks and strokes are usually acute events and are mainly caused by a blockage that prevents blood from flowing to the heart or brain. The most common reason for this is a build-up of fatty deposits on the inner walls of the blood vessels that supply the heart or brain. Strokes can also be caused by bleeding from a blood vessel in the brain or from blood clots. The cause of heart attacks and strokes are usually the presence of a combination of risk factors, such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol, hypertension, diabetes and hyperlipidaemia.

Risk Factors

The most important behavioural risk factors of heart disease and stroke are unhealthy diet, physical

inactivity, tobacco use and harmful use of alcohol. The effects of behavioural risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. These intermediate risks factors can be measured in primary care facilities and indicate an increased risk of developing a heart attack, stroke, heart failure and other complications. Cessation of tobacco use, reduction of salt in the diet, consuming fruits and vegetables, regular physical activity and avoiding harmful use of alcohol have been shown to reduce the risk of cardiovascular disease. In addition, drug treatment of diabetes, hypertension and high blood lipids may be necessary to reduce cardiovascular risk and prevent heart attacks and strokes. Health policies that create conducive environments for making healthy choices affordable and available are essential for motivating people to adopt and sustain healthy behaviour. There are also a number of underlying determinants of CVDs or the causes of the causes. These are a reflection of the major forces driving social, economic and cultural change globalization, urbanization and population ageing. Other determinants of CVDs include poverty, stress and hereditary factors. The epidemiological transition in India in the past 2 decades has been dramatic in a short timeframe, the predominant epidemiological characteristics have transitioned from infectious diseases, diseases of undernutrition, and maternal and childhood diseases to noncommunicable diseases (NCDs). The disease burden attributable to maternal disorders, measles, protein-energy malnutrition, and diarrheal diseases decreased >50% in the past 2 decades, whereas life expectancy at birth increased from 58.3 to 65.2 years, resulting in the ageing of the population during the same period. Consequently, the NCD burden increased rapidly in India, with a proportional rise in burden attributable to CVD. Nearly two-thirds of the burden of NCD mortality in India is currently contributed by CVD-related conditions. The disease transition in India in the past 2 decades resembles the accelerated epidemiological transition model with a rapid shift to the age of delayed chronic diseases.

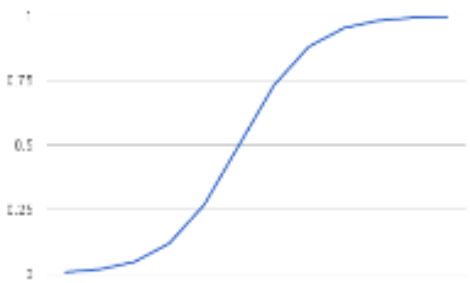
The current study focuses on heart disease among people. The project begins with the information about the cardiovascular disease and various factors that influencing the heart disease. This project is able to predict whether a person has heart disease. Through this project you will learn how heart disease affects a person's lifestyle. Finally, this study provides an understanding of how to prevent cardiovascular disease. The following are the objectives of the study are, To create awareness about cardiovascular disease, To study the influence of life style of people towards cardiovascular disease, To analyse various factors that influencing heart disease, To understand the correlation between age and the chance of increasing heart disease, To study the association between cholesterol and heart disease, To understand the relation between alcohol consumption and cardiovascular disease among people, Suggest a model that can classify and make prediction on whether an individual is susceptible to cardiovascular disease, To understand the association between the smoking habits and heart disease among people, To analyse the association between systolic blood pressure and cardiovascular disease. Facts in CVDs are the number 1 cause of death globally: more people die annually from CVDs than from any other cause. An estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. Of these deaths, 85% are due to heart attack and stroke. Over three quarters of CVD deaths take place in low- and middle-income countries. Out of the 17 million premature deaths (under the age of 70) due to noncommunicable diseases in 2015, 82% are in low- and middle-income countries, and 37% are caused by CVDs. Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies. People with cardiovascular disease or who are at high cardiovascular risk such as due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease need early detection and management using counselling and medicines, as appropriate.

METHODOLOGY

Logistic Regression is one of the basic and popular algorithm to solve a classification problem. Logistic regression analysis studies the association between a categorical dependent variable and a set of independent variables. Logistic regression competes with discriminant analysis as a method for analysing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modelling most situations than is discriminant analysis. This is because logistic regression does not assume

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

that the independent variables are normally distributed, as discriminant analysis does. This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values, and provides ROC curves to help determine the best cut off point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

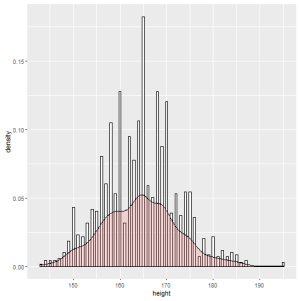


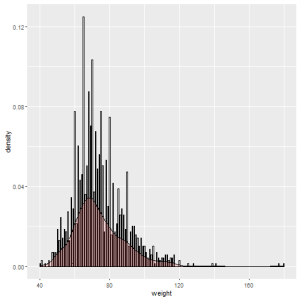
Logistic regression is named for the function used at the core of the method, the logistic function. Where e is the base of the natural algorithms (Euler's number or the EXP(x) function in your spreadsheet) and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made. Definition of the Terms are, Positive (P): Observation is positive (for example: is an apple), Negative (N): Observation is not positive (for example: is not an apple), True Positive (TP): Observation is positive, and is predicted to be positive, False Negative (FN): Observation is positive, but is predicted negative, True Negative (TN): Observation is negative, and is predicted to be negative, False Positive (FP): Observation is negative, but is predicted positive. Kruskal-Wallis H test is a rank-based nonparametric **test** that can be used to determine if there are statistically significant differences between two or more groups of an independent variable on a continuous or ordinal dependent variable. It is used for comparing two or more independent samples of equal or different sample sizes. It extends the Mann–Whitney U test, which is used for comparing only two groups. The parametric equivalent of the Kruskal–Wallis test is the one-way analysis of variance (ANOVA). A significant Kruskal–Wallis test indicates that at least one sample stochastically dominates one other sample. The null hypothesis of the Kruskal–Wallis test is that the mean ranks of the groups are the same. Classification is a two-step process, learning step and prediction step, in machine learning. In the learning step, the model is developed based on given training data. In the prediction step, the model is used to predict the response for given data. Decision Tree is one of the easiest and popular classification algorithms to understand and interpret. Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data. In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. Decision Trees follow Sum of Product (SOP) **r**epresentation. The SOP is also known as Disjunctive Normal Form. For a class, every branch from the root of the tree to a leaf node having the same class is conjunction of values, different branches ending in that class form a disjunction.

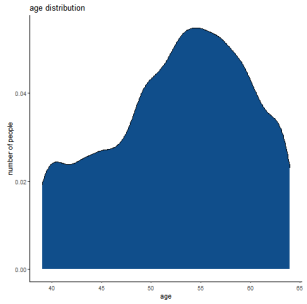
ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

DISCUSSION

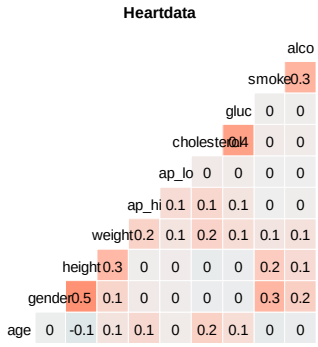
Data Visualisation







From the above three plots, height, weight and age are not normally distributed and they are all somewhat skewed. This should be taken into account in model selection and validation.



From the above graphs, there are correlations between the continuous variables: gender, cholesterol and smoking habits The rest have very low correlation coefficient.

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

```
X = np.asarray(heart[['age', 'gender', 'weight', 'cholesterol', 'gluc', 'smoke', 'alco','active','ap_hi','ap_lo']])
X[0:5]
```

```
array([[ 50.,   2.,  62.,   1.,   1.,   0.,   0.,   1., 110.,  80.],
       [ 55.,   1.,  85.,   3.,   1.,   0.,   0.,   1., 140.,  90.],
       [ 51.,   1.,  64.,   3.,   1.,   0.,   0.,   0., 130.,  70.],
       [ 48.,   2.,  82.,   1.,   1.,   0.,   0.,   1., 150., 100.],
       [ 47.,   1.,  56.,   1.,   1.,   0.,   0.,   0., 100.,  60.]])
```

```
y = np.asarray(heart['cardio'])
y [0:5]
```

```
array([0, 1, 1, 1, 0], dtype=int64)
```

To build the model we considered ‘cardio’ as the target variable and selected 10 independent variables including age, gender, weight, cholesterol, gluc, smoke, alco, active, ap_hi, ap_lo.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=4)
print ('Train set:', X_train.shape,  y_train.shape)
print ('Test set:', X_test.shape,  y_test.shape)
```

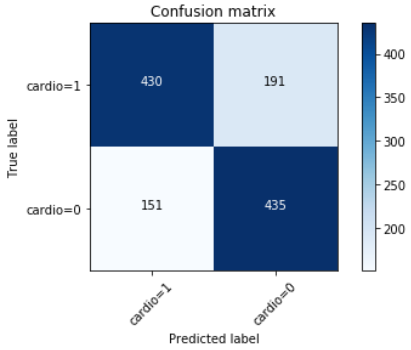
```
Train set: (4828, 10) (4828,)
Test set: (1207, 10) (1207,)
```

Splitting the data into train and test for logistic regression we are taking 80% (4828) of the total data for training the model and 20% (1207) for testing the model.

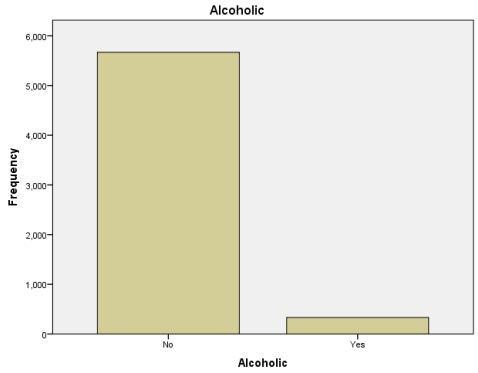
ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

Model Evaluation in Jaccard index and Confusion Matrix

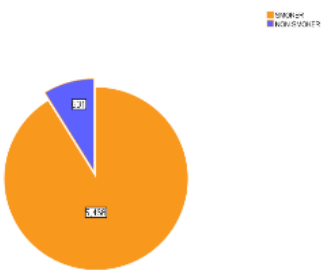
Lets try jaccard index for accuracy evaluation. we can define jaccard as the size of the intersection divided by the size of the union of two label sets. If the entire set of predicted labels for a sample strictly match with the true set of labels, then the subset accuracy is 1.0; otherwise it is 0.0.Another way of looking at accuracy of classifier is to look at confusion matrix



Look at the confusion matrix. The first row refers to individuals with heart disease in the test set. As you can calculate, out of 1207 respondents 621 of them have cardiovascular disease and out of these 621, the classifier correctly predicted 430 have the heart disease and 191 of them have not the heart disease. It means, for 430 respondents, they actually have heart disease in test set, and classifier also correctly predicted those people is suffering from cardiovascular disease. However, while the actual label of 191 respondents have heart disease but the classifier predicted those people have not heart disease, which is not very good. We can consider it as error of the model for first row. In the second row 586 respondents true label is not have heart disease. Out of these 586 the classifier correctly predicted 151 of them have not suffering from cardiovascular disease but predicted 435 of them have heart disease and we consider this as an error.

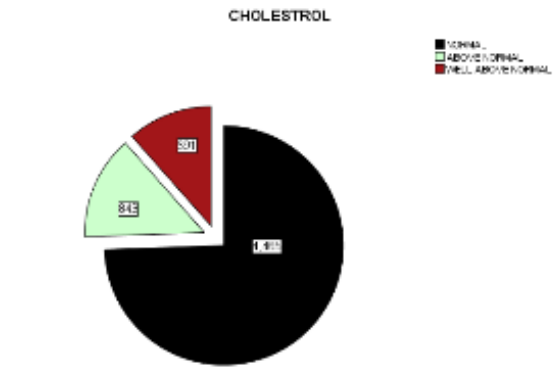


The Kruskal-Wallis test gives us a p-value of 0.05245 which is greater than 0.05. Hence, we have accept our null hypothesis and we conclude that the alcohol use does not affects the cardiovascular disease.

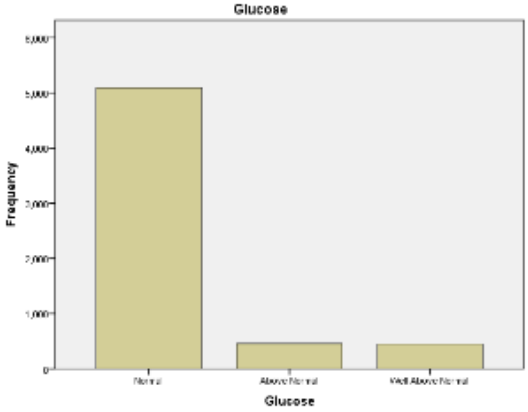


The Kruskal-Wallis test gives us a p-value of 4.182e-05 which is less than 0.05. Hence, we have reject our null hypothesis and we conclude that the smoking habits affects cardiovascular disease.

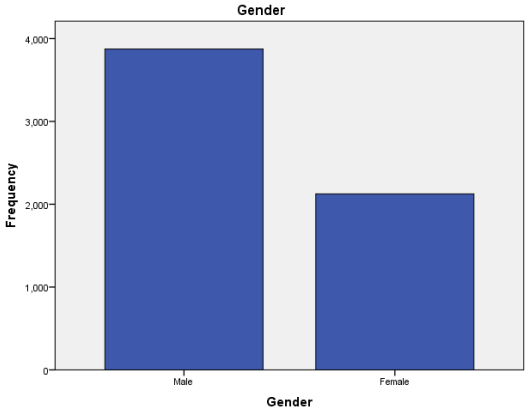
ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026



The Kruskal-Wallis test gives us a p-value is less than 0.05. Hence, we have no evidence to accept our null hypothesis and we conclude that cholesterol level affect cardiovascular disease.

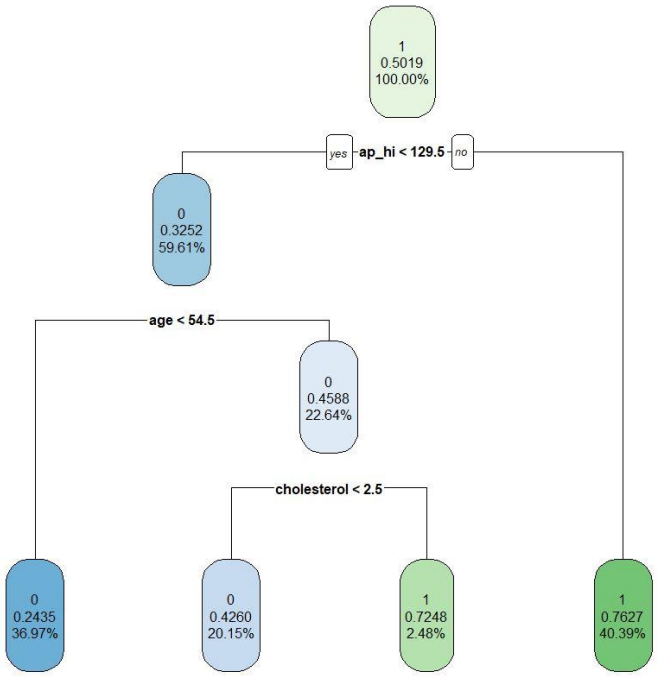


The Kruskal-Wallis test gives us a p-value is less than 0.05. Hence, we have no evidence to accept our null hypothesis and we conclude that glucose level affect cardiovascular disease.



The Kruskal-Wallis test gives us a p-value is 0.03191 which is less than 0.05. Hence, we have no evidence to accept our null hypothesis and we conclude that the gender of the people affects the cardiovascular disease. A decision tree is a flowchart-like structure in which each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label for decision taken after computing all attributes. Decision tree provides a tree-like graph which suggests the classification rules from the training set. These rules are used to predict the class labels for new unclassified observations in test data. Each node contains a splitting condition to separate the instances, and each leaf is given a class label.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026



From the above decision tree chart, the training data is classified into 4 distinct groups based on classifiers: ap_hi , age and cholesterol. The way of interpreting the decision tree is to classify the new observations based on the values of the variables stated in the classification rules. For example, an individual who has systolic blood pressure below 129.5 (ap_hi < 129.5) with age below 54.5 years old (age < 54.5) is less likely to suffer from cardiovascular diseases. Another individual with systolic blood pressure above 129.5 (ap_hi > 129.5) is more likely to have cardiovascular diseases. Cholesterol is another important variable that affect cardiovascular disease. The confusion matrix showing the performance of decision tree on test data is above.The computed performance indicators are shown below. The accuracy when handling unseen test data is 72.14%, which is lower than the accuracy in the logistic model on the test data i.e. 74%. Besides, recall is 64.96%, precision is 76.04%, and F-measure is 0.70064.

CONCLUSION

Cardiovascular diseases (CVDs) have now become the leading cause of mortality in India. A quarter of all mortality is attributable to CVD. Ischemic heart disease and stroke are the predominant causes and are responsible for >80% of Cardio Vascular Disease deaths. The Global Burden of Disease study estimate of age-standardized CVD death rate of 272 per 100 000 population in India is higher than the global average of 235 per 100 000 population. The analysis is mainly performed with the objective to understand the cardiovascular disease and the risk factors such as age, cholesterol, glucose level, alcohol intake, smoking habits, systolic blood pressure and diastolic blood pressure. The cardiovascular disease database was analysed and explored in detail. The patterns identified using Data exploration methods were validated using the modelling techniques employed. Classification models such as Logistic regression, Decision tree and Kruskal Wallis test were built to evaluate the presence of cardiovascular disease. The logistic regression model gives the accuracy of 74% and the decision tree gives the accuracy of72.14%. From the Kruskal Wallis test we can conclude that the cholesterol level, glucose level, smoking habits and systolic blood pressure are the risk factors for heart disease. The following information are observed in this paper: The elderly are more likely to suffer from various types of cardiovascular diseases. Cholesterol level is a very important determinant in leading to cardiovascular diseases. The risks of getting cardiovascular diseases climb significantly when the cholesterol content in human body rises above the normal. People with high alcohol consumption have a lower likelihood to have cardiovascular diseases. Physical activities help people become less susceptible to cardiovascular diseases. As long as people have systolic blood pressure ≥ 129, they are classified as those susceptible to cardiovascular diseases. Therefore, high systolic blood pressure plays an important role in predicting that one experiences greater risks of suffering from cardiovascular

diseases. Among the three splitting factors, two of them are age. This suggests that apart from systolic blood pressure, age is the second most significant factor in deciding whether a person will have a chance of getting cardiovascular disease.

REFERENCE

1. Davis Hosmer and Stanley Lemeshow -Applied Logistic Regression, Wiley series in probability and statistics,2nd edition,2025
2. Gupta S.C and V. K Kapoor- Fundamentals of Mathematical Statistics, Sultan Chand and Sons publisher
3. Yan, H., Chang, Z. & Liu, Z. The risk factors for calcification vary among the different sections of the lower extremity artery in patients with symptomatic peripheral arterial disease. BMC Cardiovasc Disord **20,** 333 (2020).
4. Chen, X., Lin, Y., Tian, L. et al. Correlation between ischemia-modified albumin level and coronary collateral circulation. BMC Cardiovasc Disord **20,** 326 (2020).
5. www.kaggle.com
6. MN Krishnan, Coronary heart disease and risk factors in India – On the brink of an epidemic, 64(4):364–367, Indian Heart Journal, 2022.
7. David Calvin Goff Jr , Sadiya Sana Khan , Donald Lloyd-Jones, Donna K Arnett, Mercedes R Carnethon , Darwin R Labarthe , Matthew Shane Loop , Russell V Luepker, Michael V McConnell , George A Mensah 8, Mahasin S Mujahid9, Martin Enrique O'Flaherty, Dorairaj Prabhakaran, Véronique Roger, Wayne D Rosamond, Stephen Sidney, Gina S Wei 1, Janet S Wright, Bending the Curve in Cardiovascular Disease Mortality: Bethesda + 40 and Beyond, 22;143(8):837–851, Circulation, 2021.
8. Sundeep Salvi, G Anil Kumar, R S Dhaliwal, Katherine Paulson,Anurag Agrawal, Parvaiz A Koul, P A Mahesh, Sanjeev Nair,Virendra Singh, Ashutosh N Aggarwal, D J Christopher,Randeep Guleria, B V Murali Mohan, Surya K Tripathi, Aloke G Ghoshal,R Vijai Kumar, Ravi Mehrotra, D K Shukla, Eliza Dutta, Melissa Furtado, Deeksha Bhardwaj, Mari Smith, Rizwan S Abdulkader, Monika Arora,Kalpana Balakrishnan, Joy K Chakma, Pankaj Chaturvedi,Sagnik Dey, Deesha Ghorpade, Scott Glenn, Prakash C Gupta,Tarun Gupta, Sarah C Johnson, Tushar K Joshi, Michael Kutz,
9. Manu R Mathur, Prashant Mathur, Pallavi Muraleedharan,Christopher M Odell, Sanghamitra Pati, Yogesh Sabde,Dhirendra N Sinha, K R Thankappan, Chris M Varghese,Geetika Yadav, Stephen S Lim, Mohsen Naghavi, Rakhi Dandona,K Srinath Reddy, Theo Vos, Christopher J L Murray,Soumya Swaminathan, †Lalit Dandona.The burden of chronic respiratory diseases and their heterogeneity across the states of India: the Global Burden of Disease Study 1990–2016, 6: e1363–74,Lancet Glob Health 2018.
10. Salim S Virani, Alvaro Alonso, Emelia J Benjamin, Marcio S Bittencourt, Clifton W Callaway, April P Carson, Alanna M Chamberlain, Alexander R Chang, Susan Cheng, Francesca N Delling, Luc Djousse, Mitchell S V Elkind, Jane F Ferguson, Myriam Fornage, Sadiya S Khan, Brett M Kissela, Kristen L Knutson, Tak W Kwan, Daniel T Lackland, Tené T Lewis, Judith H Lichtman, Chris T Longenecker, Matthew Shane Loop, Pamela L Lutsey, Seth S Martin, Kunihiro Matsushita, Andrew E Moran, Michael E Mussolino, Amanda Marma Perak, Wayne D Rosamond, Gregory A Roth, Uchechukwu K A Sampson, Gary M Satou, Emily B Schroeder, Svati H Shah, Christina M Shay, Nicole L Spartano, Andrew Stokes, David L Tirschwell, Lisa B VanWagner, Connie W Tsao, Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association, Circulation, Volume 141, Issue 9, 3 , e139-e596, 2020.
11. Rajeev Gupta, Soneil Guptha, Rajnish Joshi & Denis Xavier , Translating evidence into policy for cardiovascular disease control in India, BMC journals Health Research Policy and Systems, Vol 9, 8-18,2021.
12. Andrew C. Leon, Comprehensive Clinical Psychology,Reference Module in Neuroscience and Biobehavioral Psychology, 243-285, Volume 3,1998.