# A Statistical Approach for early detection and modeling of Heart Diseases.

Jyotismita Talukdar, Dr. Sanjib Kr. Kalita

**Abstract—**Cardiovascular diseases are considered as one of the main causes of morbidity and mortality rate. It causes the narrowing of the blood vessels that prevents supply of oxygen to the heart. This study analyses to design a linear statistical model to detect possibility of heart disease. We have analyzed the correlation and partial correlation coefficients using Pearson method for four primary attributes namely Cholesterol, Thalach, Blood Pressure and Fasting Blood Sugar apart from family history, Hypertension and Diabetes.

**Index Terms:** Pearson correlation, Cholesterol, Blood pressure, Thalach, Fasting blood sugar, Diabetes

## I. INTRODUCTION

Coronary Heart Disease (CHD) is one of the leading global causes of morbidity and mortality. Though in the developed countries the morbidity and mortality rates are declining but this decline is not universal. Some of the compelling underlying reasons are: increase of aging population, socio-economic inequalities, impaired quality of life etc. In coronary heart disease, the small blood vessels supplying blood and oxygen to the heart are narrowing down. This causes slowdown of blood flow to heart, and causing chest pain, shortness of breath, heart attack and other symptoms. Though the medical informatics provides measurable improvements in both the quality of care and effectiveness, still such facilities are far cry for common mass in most developing and underdeveloped countries.

Among the various health issues, the most predominant one is the heart failure. It occurs especially in old patients because of diet, non-steroidal anti-inflammatory drugs etc. One of the most commonly occurred heart disease is the Cardio-Vascular Disease (CVD). Cardio Vascular Disease incorporates coronary heart disease, cerebrovascular (stroke), hypertensive heart disease, congenital heart disease, peripheral artery disease, rheumatic heart disease, and inflammatory heart disease [4]. Globally, the heart attack disease remains as the main cause of death. To prevent or remain safe from such attack it is imperative or essential to detect such possible attack at an earlier stage.

## II. RELATED WORK

their self-generated data, but they are not used effectively for prediction [1]. Even people die of heart attack with experienced symptoms which were not taken into considerations [2]. Among the various factors causing heart disease , the relevant factors

Though the medical practitioners make prediction and do treatment based on hidden information present in

that increase the possibility of heart attack are : high blood pressure(BP), high Cholesterol(CHOL), unhealthy diet, harmful use of alcohol, smoking, lack of physical exercise, and high sugar level etc.[3][4]. There are various types of algorithms and techniques available in literature. Some of the commonly used Data mining algorithms are: J48, Native Bayes, REPTREE, CART, and Bayes Net etc. In addition, Neural Networks (NN), Association Classification and Genetic algorithm are also used by many researchers for predicting and analyzing heart diseases.Chitra, R and Seenivasagam, V [5] used neural network and hybrid intelligent algorithm and showed that hybrid algorithm technique improved accuracy of prediction. Sundar M.A et al [6] predict the probability of a patient receiving heart attack using Native Bayes and Weighted Associative Classifier (WAC). In 2012, Pattekari et al [7] developed via web based intelligent system using Native Bayes to solve complex queries for diagnosing heart disease. There are many predicting models in the domain of CVD [9][1].

The major risk factors for CVD are: 1. Diabetes, 2. High LDL (bad) cholesterol, 3. Low LDL (good) cholesterol, 4. High blood pressure, 5. Less physical activity, 6. Obesity 7. Depression and 7. Smoking. As per

WHO's report, in India, out of 10 deaths, eight are caused by non-communicable diseases, such as, CVD and diabetes. In rural India, 6(six) out of 10(ten) deaths is caused by non-communicable diseases (NCD). It is thus evident that there is an urgent need for development and implementation of suitable prevention approaches to control this epidemic. Of course, sincere bureaucratic, political and social must come in parallel.

In this paper an attempt has been made following statistical approaches to analyze and modeling some primary attributes related to primary issues and causes of heart attack.

## III. RESEARCH METHODOLOGY

In this paper we developed a linear statistical model for predicting the possible heart diseases. The

different attributes we considered for prediction using a statistical model are:

1. Age
2. Sex
3. Family history
4. Blood pressure(BP)
5. Fasting blood sugar(FBS)
6. Cholesterol(CHOL)
7. Thalacy(THAL)
8. Comorbidity ,[ hypertension(HTN) and Type-2 diabetes (T2DM)]

In the present analysis, the correlation coefficients (Pearson) and partial correlation coefficients of four basic attributes, namely, Blood pressure (BP), Cholesterol (CHOL), Thalacy(THAL) and Fasting blood sugar(FBS) are computed. The relation used in computing the coefficients are:

For simple correlation between variable x and y (Pearson),

$$r_{xy} = \frac{\sum_{i=1}^{N}((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \quad (1)$$

$$\bar{x} = \frac{\sum x_i}{N}, \ \bar{y} = \frac{\sum y_i}{N} \quad (2)$$

Where, xi and yi are BO, FBS, CHOL, & THAL. N is the number of patients admitted with symptoms of heart attack. Similarly, the partial correlation coefficients are computed using the relation as given below:

$$r_{x_i x_j (y} = \frac{r_{x_i x_j} - r_{x_i y} r_{x_j y}}{\sqrt{(1 - r_{x_i y}^2)(1 - r_{x_j y}^2)}} \quad (3)$$

Where, $x_i$ and $y_j$ are patient's attributes related to heart attack. The expression (3) represents the partial correlation coefficients between **xi** and **xj** if they obtained the same score on variable **y**. For example,

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \quad (4)$$

The partial correlation coefficients $\mathbf{r_{12.3}}$ indicates the relationship between **variable 1** and **variable 2** when everyone obtained the same score on **variable 3.** The partial correlation coefficients between four variables can be obtained following the expression given as:

$$r_{12.34} = \frac{r_{12.3} - r_{14.3} \cdot r_{24.3}}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}} \quad (5)$$

The general form of equations (4) and (5) can be expressed as:

$$r_{12.34\cdots N} = \frac{r_{12.34\cdots(N-1)} - r_{1N.34\cdots(N-1)} \, r_{2N.34\cdots(N-1)}}{\sqrt{(1 - r_{1N.34\cdots(N-1)}^2)(1 - r_{2N.34\cdots(N-1)}^2)}} \quad (6)$$

The present study is made for heart patients of age between 25 years to 100 years. The total number of patients considered is 3236(including both males and females). The data collected from five leading hospitals, both government and private nursing home, at Guwahati city, Assam, India. The Pearson correlation coefficients (simple correlation coefficients) and partial correlation coefficients between the four attributed, i.e., BP, CHOL, FBS, and THAL are computed using the relations (4), (5) and (6).The **table-1** shows the simple correlation coefficients between BP, FBS THAL and CHOL. **Table-II** show the simple and partial correlation coefficients r12.3, r12.34 at different BP ranges. Here, '1' represents BP, '2' represents FBS, '3' represents THAL and '4' represents CHOL. Similarly, **table-III** represent the simple correlations and partial correlation coefficients, r14, r14.2, and r14.23 taking CHOL and FBS as control variables. **Table-IV** represent the simple and partial correlation coefficients between BP and CHOL at different ranges of Cholesterol, taking FBS and THAL as control variables. **Table-V** represents correlation and partial correlation coefficients between CHOL and FBS taking BP and THAL as control variables. Similarly, **Table-VI** represents the correlation and partial correlation coefficients between FBS and THAL, at different ranges of FBS, taking BP and CHOL as control variables. **Table-VII** represents the relation between CHOL and THAL, taking BP and FBS as control variables, at different ranges of at different ranges of CHOL.

TABLE I BIVARIATE CORRELATIONS AMONG BP, FBS, CHOL, THALACH

| | | BP | FBS | HAL | CHOL |
|---|---|---|---|---|---|
| BP | Pearson Correlation | 1 | -.226(**) | -.111(**) | .034 |
| | Sig. (2-tailed) | | .000 | .000 | .278 |
| | N | 3236 | 3236 | 3236 | 3236 |
| FBS | Pearson Correlation | -.226(**) | 1 | .251(**) | -.083(**) |
| | Sig. (2-tailed) | .000 | | .000 | .009 |
| | N | 3236 | 3236 | 3236 | 3236 |
| THAL | Pearson Correlation | -.111(**) | .251(**) | 1 | -.026 |
| | Sig. (2-tailed) | .000 | .000 | | .412 |
| | N | 3236 | 3236 | 3236 | 3236 |
| CHOL | Pearson Correlation | .034 | -.083(**) | -.026 | 1 |
| | Sig. (2-tailed) | .278 | .009 | .412 | |
| | N | 3236 | 3236 | 3236 | 3236 |

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

TABLE II CORRELATION OF BP, FBS TAKING CHOL, THALACH AS CONTROL VARIABLES

| FBS | $r_{24}$ | $r_{24.1}$ | $r_{24.13}$ |
|---|---|---|---|
| >100,<=120 | -0.173 | -0.173 | -0.174 |
| >120,<=140 | 0.045 | 0.039 | 0.042 |
| >140,<=160 | 0.040 | 0.013 | 0.009 |
| >160,<=180 | 0.154 | 0.145 | 0.139 |
| >180,<=200 | 0.039 | 0.025 | 0.019 |
| >200,<=220 | 0.072 | 0.074 | 0.064 |
| >220,<=240 | 0.081 | 0.073 | 0.073 |
| >240, <= 260 | 0.074 | 0.002 | 0.002 |

TABLE III CORRELATION OF BP, THAL TAKING CHOL, FBS AS CONTROL VARIABLES

| **BP_Range** | $r_{12}$ | $r_{12.3}$ | $r_{12.34}$ |
|---|---|---|---|
| >120,<=140 | -0.293 | -0.298 | -0.285 |
| >140,<=160 | -0.474 | -0.474 | -0.470 |
| >160,<=180 | 0.513 | 0.513 | 0.535 |
| >180 | 0.520 | 0.521 | 0.560 |

TABLE IV CORRELATION OF BP, CHOL TAKING THAL, FBS AS CONTROL VARIABLES

| **BP_Range** | $r_{14}$ | $r_{14.2}$ | $r_{14.23}$ |
|---|---|---|---|
| >120,<=140 | -0.160 | -0.132 | -0.132 |
| >140,<=160 | -0.169 | -0.153 | -0.153 |
| >160,<=180 | -0.182 | -0.183 | -0.183 |
| >180 | 0.053 | 0.056 | 0.5.60 |

TABLE V CORRELATION OF FBS, CHOL TAKING THAL, BP AS CONTROL VARIABLES

| **Chol.** | $r_{13}$ | $r_{13.2}$ | $r_{13.24}$ |
|---|---|---|---|
| >150,<=170 | 0.065 | 0.065 | 0.064 |
| >170,<=190 | -0.019 | -0.019 | -0.020 |
| >190,<=210 | 0.040 | 0.041 | 0.045 |
| >210,<=230 | 0.001 | 0.002 | 0.070 |
| >230,<=250 | 0.133 | 0.135 | 0.136 |
| >250,<=270 | 0.074 | 0.074 | 0.078 |
| >270,<=290 | 0.001 | 0.001 | 0.003 |

TABLE VI CORRELATION OF FBS, THAL TAKING CHOL, BP AS CONTROL VARIABLES

| **Chol.** | $r_{23}$ | $r_{23.1}$ | $R_{23.14}$ |
|---|---|---|---|
| >150,<=170 | -0.004 | 0.003 | 0.012 |
| >170,<=190 | -0.001 | 0.002 | 0.020 |
| >190,<=210 | 0.016 | 0.016 | 0.028 |
| >210,<=230 | 0.035 | 0.002 | 0.055 |
| >230,<=250 | 0.020 | 0.032 | 0.028 |
| >250,<=270 | 0.023 | 0.025 | 0.078 |
| >270,<=290 | 0.008 | 0.015 | 0.006 |

TABLE VII CORRELATION OF THAL, CHOL TAKING FBS, BP AS CONTROL VARIABLES

| Chol. | $r_{34}$ | $r_{34.1}$ | $r_{34.12}$ |
|---|---|---|---|
| >150,<=170 | -0.015 | -0.050 | -0.051 |
| >170,<=190 | -0.011 | 0.013 | 0.013 |
| >190,<=210 | 0.069 | 0.071 | 0.074 |
| >210,<=230 | 0.066 | 0.067 | 0.079 |
| >230,<=250 | 0.017 | 0.025 | 0.019 |
| >250,<=270 | 0.016 | 0.027 | 0.023 |
| >270,<=290 | 0.056 | 0.056 | 0.054 |

It is observed that out of 3236 heart patient's database collected from five hospitals (including both private nursing home and Govt. hospitals), the number of male heart patients admitted with symptoms is 78% and female heart patients admitted is only 22%. This is shown in **figure 1(a,b).** In the present study we are considering the heart patients in two sampled age categories: age ranging from 40 years to 6o years as **sample -1** and age ranging from 61 years to 85years as **sample -2.** It is observed that in **sample-1**, out of 78% male patients, only 3% male patients have positive family history and rest 97% does not have any family history. Also, it is observed that the male patients with positive family history admitted in the hospital with a primary complain of HTN **(Hypertension)**. On the other hand, male patients without any family history, only 17% admitted with a primary complain of **T2DM (Type-2 diabetes)** and rest 83% admitted with a complain of HTN. Similarly, in case of female patients (22%) only 6% found to have positive family history and rest 94% does not follow any family history. In case of female patients with positive family history the common symptom of heart attack is found as HTN. On the other hand, females without any history i.e. 94%, **32%** are suffering from **T2DM** and rest 68% suffering from **HTN.**

Similarly, in **sample-2, figure (1b),** the number of heart patients within age limit 61 years to 85 years, 75% heart patients are male and rest 25% are female. Among the male heart patients, only 10% found to have positive family history and rest 90% does not follow any history. Within this age limit, male with positive history, 13% suffering from T2DM and rest 87% suffering from HTN. On the other hand, male without any family history, 9% suffering from T2DM and rest 91% suffering from HTN. In case of female patients within this age limit (25%), only 8% possess family history and rest 92% does not follow any history. Most of the female patients with positive family history, suffering from HTN (98%) and rest (2%) suffering from T2DM. Similarly, for female patients without any family history (92%), most of the patients suffering from HTN (78%) and only 28% suffering from T2DM prior to heart attack.

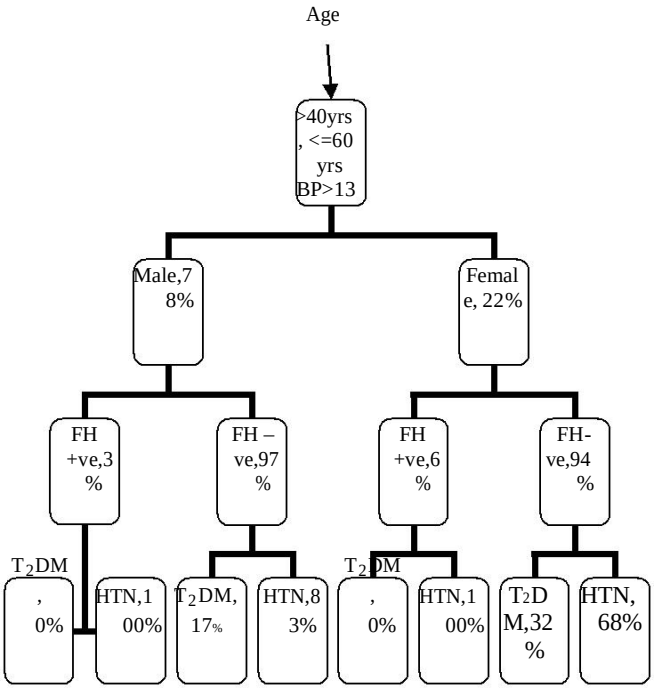


Figure 1a. Heart Patients within age limit >40 years, <=60 years

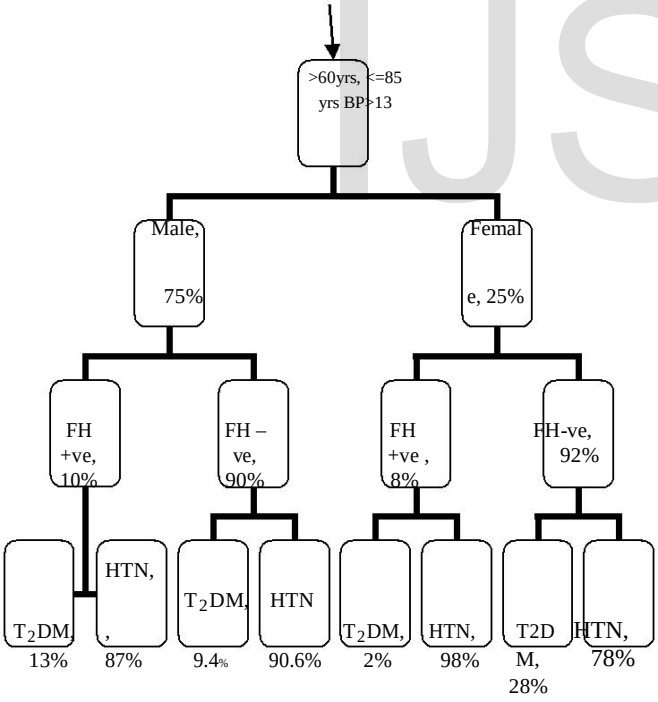


Figure 1b. Heart Patients within age limit >60 years, <=85 years

Taking the four primary attributes responsible for heart attack (BP, CHOL, FBS and THAL) we are developing linear modeling following the equations 7(a), 8(a) 9(a) and 10(a). These models have been developed considering all the heart patients (3236) without distinguishing male and female patients. In table VIII, we have shown the comparison between the dependent variables and independent variables of the attributes. For BP considered as dependent variable we have FBS, CHOL, and THAL as independent variables. Similarly, for FBS as dependent variable and BP, CHOL, THAL as independent variables, we have value of R as 0.143 and $R^2$ as 0.108. Similarly, THAL is considered as dependent variable and BP, CHOL, FBS as independent variables. For CHOL is considered as dependent variable and BP, FBS THAL are independent variables.

Table VIII: Comparison of the dependent and Independent variables

| Dependent variable | Independent variable | R | $R^2$ |
|---|---|---|---|
| BP | FBS, THAL, CHOL | 0.233 | 0.054 |
| FBS | BP,THAL,CHOL | 0.143 | 0.108 |
| THAL | BP,CHOL,FBS | 0.150 | 0.066 |
| CHOL | BP, THAL,FBS | 0.024 | 0.007 |

$$y(BP)=159.04–0.179(FBS)– .048(THAL)+0.008(CHOL) \quad 7(a)$$

$$y(FBS)=132.47–0.234(BP)–0.224(THAL)-0.040(CHOL) \quad 8(a)$$

$$y(THAL)=62.015 – 0.068(BP) - 0.003(CHOL) +0.242(FBS) \quad 9(a)$$

$$y(CHOL)=274.116 + 0.034(BP) – 0.011(THAL) - 0.136(FBS \quad 10(a)$$

## RESULTS AND CONCLUSION

From the present study and analysis of 3236 heart patient's data base, the following results have been obtained:

1. Referring to **Table-I**, irrespective of sex, the simple or Pearson's correlation coefficients is found positive only in case of BP and CHOL(0.034) and FBS& THAL(0.251) , which are significant at the 0.01 level(2-tailed)
2. Referring to tables I to table VII, which show the partial correlation coefficients of four primary attributes(BP,CHOL,FBS,THAL) responsible for heart attack, it is observed that :
   - **(i)** As shown in **Table-II**, in the **BP** range above **160mm/Hg**, attribute **THAL** has a positive impact over the correlation between **BP** and **FBS.**
   - (ii) Above the BP range of **180mm/Hg, the CHOL** shows positive impact on the BP, as shown in **Table-III**.
   - (iii) It is evident from **Table-IV** that within the CHOL range >230mm/dl and <= 250mm/dl, the BP & CHOL possess high correlation with minimum impact of FBS and THAL.
   - (iv) In Table-5 and Table-6, though some positive relationships observed between CHOL & FBS, but not significant. Further, the tables show negligible impact of BP & THAL over the relationship between CHOL & FBS.

3. From **Figure 1(a)** and **1(b),** it is clearly observed that heart attack or heart diseases does not strictly depend on one's family history. Further, males are found more sensitive to

   heart attack (78%) than female (22%). In addition, irrespective of sex and family history, the **HTN (Hypertension**) is found to have been primary and lading symptom of heart attack.
4. From the linear modeling, given by equations: 7a, 8a, 9a and 10a, it is clearly observed that (BP & CHOL) act as a pair and (FBS& THAL) also act as a pair. Constituents of each pair has mutual dependency.

   From the $R^2$ value it is observed that out of BP, FBS, CHOL and THAL, taking as dependent variables, the best-fit linear model among the four primary attributes responsible for heart attack, can be considered as the

   Linear model represented by equation 8(b).

It is thus concluded that BP and CHOL are directly correlated. Likewise, FBS and THAL are dependent on each other. There seems no strong relation or dependency of either BP or CHOL on FBS or THAL and vice-versa. But when the same attributes are studied using partial correlation in different BP and CHOL ranges, it is observed that above 160mm/Hg BP, the FBS, CHOL. & THAL are dependent on each other. The present study reveals that heart attack does not strictly follow any family history. It also reveals that males are more sensitive to heart related health problems than females. Further, HTN is found as more effective symptom than T2DM or other types of co morbidity for earlier prediction of hear related problems of an individual.

**Acknowledgement:** We do hereby acknowledge Prof P.H.Talukdar, professor, Dean, Kaziranga University,Asam for his valuable suggestions and creative criticism while writing this manuscript.

## REFERENCES

[1] A.K.Sen;S.B.Patel and D.P.Sukla : " A data mining technique for prediction of coronary heart disease using Neuro-Fuzzy integrated approach Two level. "International Journal of Computer Engineering & Technology (IJCET), Vol.-3, PP. 30-40, 2012.

[2] S. Ishtake & S.Sanap, "Intelligent heart disease prediction system using data mining technique", "International Journal of Healthcare and Biomedical research",Vol-1, No.3, PP-94 – 101, 2013

[3] D.S Chaitrali & A.S ulabha, "A data mining approach for prediction of heart disease using neural networks", "International Journal of Computer engineering & Technology (IJCET) vol-3, No.3, PP – 30-40, 2012.

[4] V.Chaurasia: "Early prediction of heart diseases using Data mining. " Caribbean Journal of Science and Technology. "Vol.-1, No. 1, PP. 208-217, 2013

[5] R. Chitra and V. Seenivasagam: "Review of Heart Disease Prediction System using Data mining and hybrid intelligent Techniques. "International Journal of Soft computing (ICTACT), Vol.-3, No. 4, PP. 605-609, 2013.

[6] N.A.Sundar; P.P.Latha;and M.R.Chandra:" Performance Analysis of Classification Data mining Techniques over heart disease Data base. ", International Journal of Engineering Science and Advanced Technology. ", Vol.-2, No. 3, PP. 470-478, 2012.

[7] S.A.Pattekari and A. Praveen: "Prediction system for heart disease using Naïve Bayes. ", International Journal of Advanced Computer and Mathematical Science. ", Vol.-3, No. 3, PP. 290-294, 2012.

[8] The Times of India, 14th August, 2011.

[9] S.B.Patil and Y.S.Kumaraswamy: "Extraction of Significant Pattern from Heart Disease Warehouses for Heart attack prediction. ", International Journal of Computer Science and Network Security. ", (IJCSNS), Vol.-9, No. 2, PP. 228-235, 2009.