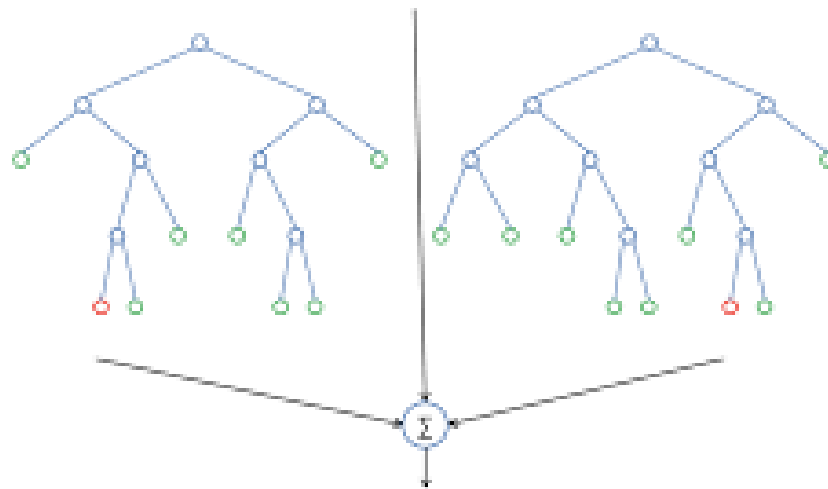


IDENTIFYING MAJOR DEATH CAUSING DISEASE USING DECISION TREE BASED DEATH ANALYSIS



IDENTIFYING MAJOR DEATH CAUSING DISEASE USING DECISION TREE BASED DEATH ANALYSIS

Harshal Jaju & Richa Sharma

Abstract

The strength of any structure lies in its foundation likewise the strength of any nation lies in the health of its citizens. Focusing on the health of the citizens of a country leads to the formation of a prosperous nation as they add in substantial growth and development of the nation. In the present era, the major threat to humankind is the exponential growth of deadly growing diseases. This study deals with decision tree analysis of major death-causing diseases. In this paper, we have considered four major death-causing diseases namely cancer, diabetes, cardiovascular disease and chronic respiratory disease (CRD). The major focus is on the disease cancer as globally it is the second prime cause for death specifically, considering certain tumor types that affect men and women respectively. Further, the decision tree analysis model shows a step-by-step analysis for identifying the death-causing disease. More generally, the decision tree analysis (DTA) assists in achieving the most optimized alternative for a situation.

Keywords

Decision tree, Analysis, Diseases, Cancer, Diabetes, Cardiovascular disease and Chronic respiratory disease

1. Introduction

A disease is an illness condition that affects the functioning of the body. A disease can be caused by many external factors (viruses, bacteria) and internal factors (misfunctioning of organs). The person suffering from diseases senses pain, stress, social problems or sometimes in worse cases it becomes the reason for the death of the people. The disease has always been the biggest reason for death globally. There are numerous numbers of death-causing diseases. Does the question arise how to find out the world's biggest killer disease?

Harshal Jaju
harshaljaju@jklu.edu.in

Once the major death-causing disease is figured, the health institutions can take measures for controlling the death of people due to these diseases.

According to the WHO report on global cancer, it is estimated that there will be 18.1 million new cases and 9.6 million deaths in 2018 due to the cancer burden. It was also found that worldwide one in 5 men and one in 6 women will develop cancer in their lifetime, and one in 8 men and one in 11 women die due to cancer [1].

DTA is used in almost all fields for making a choice that is the most optimal to adopt and at the same time reliable for further study. Bae has proposed in his study that the clinical decision analysis (CDA) using the decision tree which has helped to control the problem of complexity and ambiguity in medical problems [2]. Podgorelec et al. discussed that in medical decision making (classification, diagnosing, etc.) there are many situations where the decision must be made reliably. According to the author, a decision tree is effective in making techniques that provide high classification and simple representation of gathered knowledge and they have been used in different areas of medical decision making [3]. Thus, the decision tree helps us in picking the alternative that has the maximum effectiveness and minimum harm.



In this paper, we have mathematically analysed the official WHO data for various death causing disease, which helps us in identifying the major killer disease due to which many men and women die globally every year. Section 2 depicts the step by step formation of a decision tree with some standard notations. In section 3, the computation of all the diseases is provided. Section 4 provides us a simpler way for doing an analysis of complex DTA with the help of a program made in Java and executed on the NetBeans platform. Section 5 discloses the comparison between various diseases through pie charts and bar graphs. The conclusion is provided in section 6.

2. Model Description

The decision tree graphically demonstrates all the alternatives and evaluation of each alternative helps us in selecting the most suitable outcome. It allows recognizing of all the feasible results that trace each possibility and provides an optimal solution. It is a reliable way of presenting, interpreting and drawing a conclusion from all the possible outcomes.

In our study, we have considered the data of death-causing disease for the last 6 years from 2012- 2018. Fig 1 shows the basic decision tree with all the possible major death-causing

diseases(alternatives). It also provides us the idea of the subtypes of disease and shows probable nodes of making decisions and nodes at which outcomes occur. In Fig 2, the probability at which cancer [4], chronic respiratory disease [5], diabetes [6] and cardiovascular disease [7] affect mankind globally is represented. The last fig 3 depicts the total number of people dead globally due to the death-causing disease. In table 1, we provide various notations for modelling the DTA. These notations will help us in formulating the step -by- step decision tree.

SYMBOL	MEANING
	Represents a time when a decision is made.
	Represents a time when the outcomes occur.
	Represents a time when no nodes come out of the branch.
A	The initial node.
B	The final value of people affected by cancer.
C	The total number of expected people dead globally due to diabetes.
D	The total number of expected people dead globally due to cardiovascular disease.
E	The total number of expected people dead globally due to chronic respiratory disease.
F	The total number of expected men died due to cancer.
G	The total number of expected women died due to cancer.
EMLC	The expected number of men died due to lung cancer.
EMPC	The expected number of men died due to prostate cancer.
EMCC	The expected number of men died due to colorectum cancer
EMSC	The expected number of men died due to stomach cancer.
EMLVC	The expected number of men died due to liver cancer.
EMBC	The expected number of men died due to bladder cancer.
EWBC	The expected number of women died due to breast cancer.
EWLC	The expected number of women died due to lung cancer.

EWCC	The expected number of women died due to colorectum cancer.
EWCUC	The expected number of women died due to cervix uteri cancer.
EWSC	The expected number of women died due to stomach cancer
EWOC	The expected number of women died due to ovary cancer.
EWLVC	The expected number of women died due to liver cancer.
P1	Probability of EMLC
P2	Probability of EMPC
P3	Probability of EMCC
P4	Probability of EMSC
P5	Probability of EMLVC
P6	Probability of EMBC
p1	Probability of EWBC
p2	Probability of EWLC
p3	Probability of EWCC
p4	Probability of EWCUC
p5	Probability of EWSC
p6	Probability of EWOC
p7	Probability of EWLVC

Table 1 Notations

The following steps are used for step- by- step formation of the decision tree for the identification of major death-causing disease.

Step-1: The basic decision tree model with major death-causing diseases(alternatives).

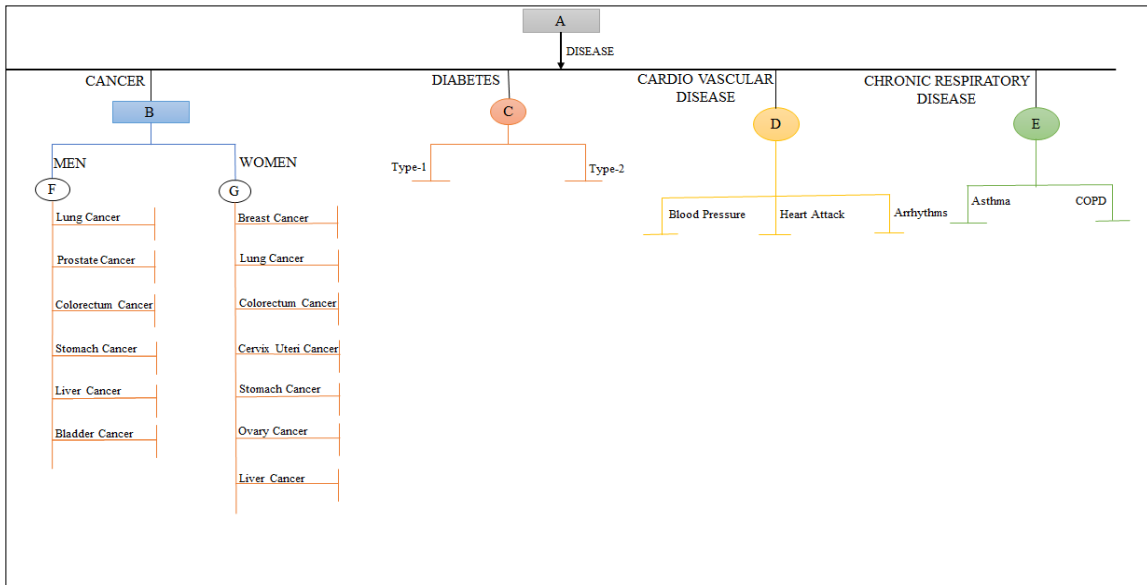


Fig 1: The basic decision tree model for identifying major death-causing diseases

Step-2: In this step, the below-mentioned probabilities in fig 2 are used for modeling of the decision tree.

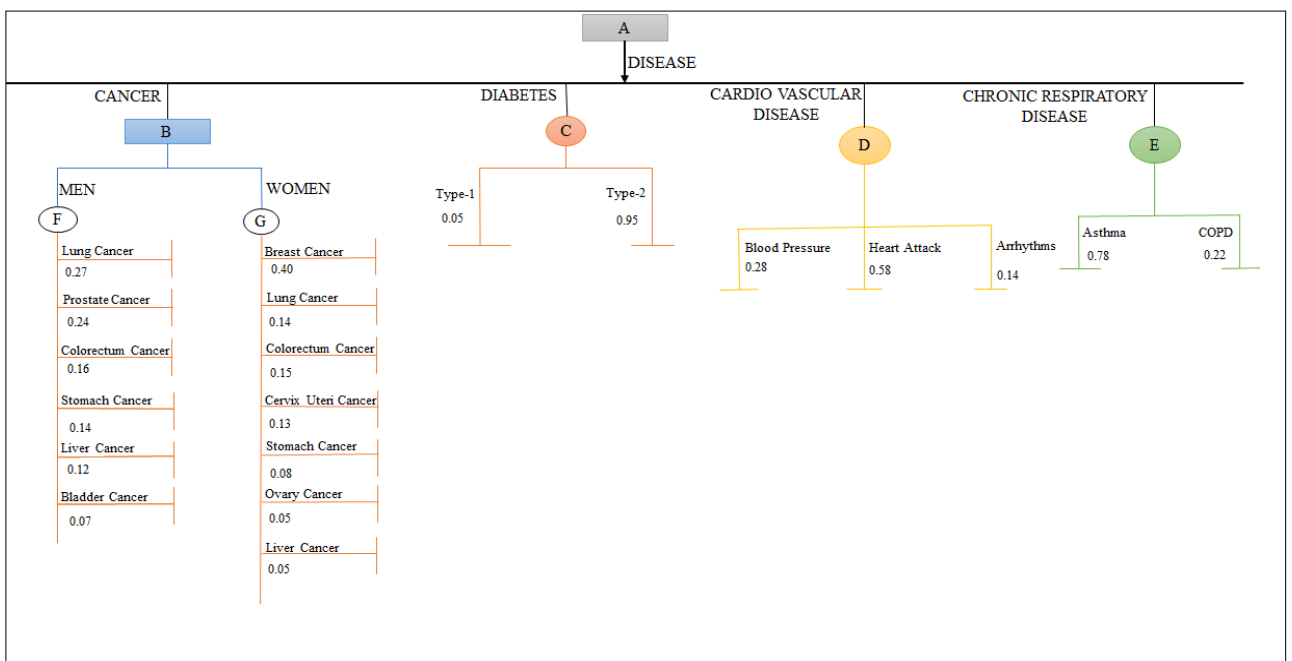


Fig 2: The probability with which these diseases affect people globally

Step-3: The final step is to consider the number of death of people globally (loss) against each alternative for further analysis of the decision tree.

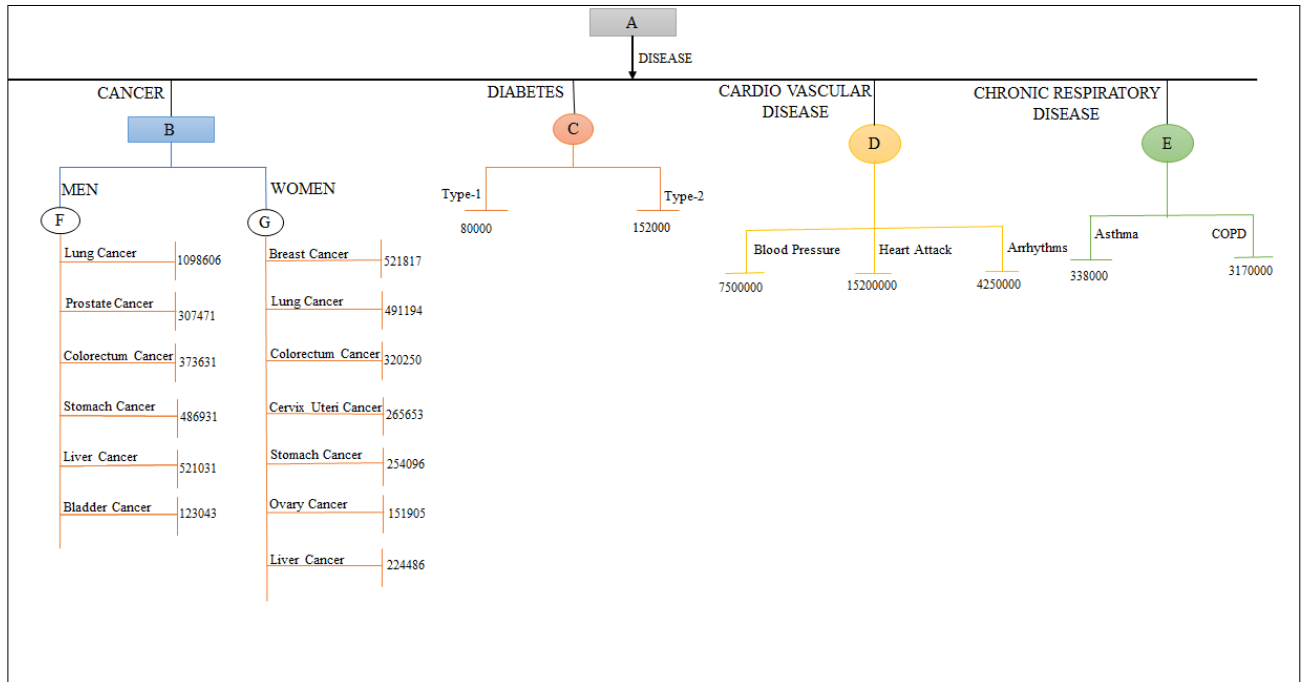


Fig 3: The total number of people died globally due to various diseases

3. Analysis

To find out which disease has affected most people with the provided data following computations are done.

Step-1 Calculation of number of people died due to Cancer

- a) First, we have calculated the total number of men died due to cancer.
 - Each of the tumors types which mostly affects men worldwide is calculated and then the summation of all is done to find out the total number of men died due to cancer. The calculations are done as follows-

$$\text{EMLC} = (0.27 * 1098606) \sim 296,624$$

$$\text{EMPC} = (0.24 * 307471) \sim 73,793$$

$$\text{EMCC} = (0.16 * 373631) \sim 59,781$$

$$\text{EMSC} = (0.14 * 486931) \sim 68,170$$

$$\text{EMLVC} = (0.12 * 521031) \sim 62,524$$

$$\text{EMBC} = (0.07 * 123043) \sim 8,613$$

- The total number of men dead due to cancer are 569,505. **Thus, the value of node F is 569505.**

b) Similarly, we have calculated the total number of women died due to cancer.

$$EWBC = (0.40 * 521817) \sim 208,727$$

$$EWLC = (0.14 * 491194) \sim 68,767$$

$$EWCC = (0.15 * 320250) \sim 48,038$$

$$EWCUC = (0.13 * 265653) \sim 34,535$$

$$EWSC = (0.08 * 254096) \sim 20,328$$

$$EWOC = (0.05 * 151905) \sim 7,595$$

$$EWLVC = (0.05 * 224486) \sim 11,224$$

- The total number of women dead due to cancer are 399,214. **Thus, the value of G node is 399,214.**

c) To finally calculate the value of node B the value of node F and node G is compared. Out of the two nodes, the node having greater values becomes the value of node B. The number of men died due to cancer (node F) is greater than the number of women died due to cancer (node G)

Thus, the value of node B is 569,505.

The same procedure as illustrated in step-1 is followed in the calculation of the number of people died due to all disease.

Step-2 Calculation of the number of people died due to Diabetes.

- The total number of people dead are = $[(0.05 * 80,000) + (0.95 * 152000)]$
= 4000 + 144,400
= 148,400

Thus, the value of node C is 148,400.

Step-3 Calculation of the number of people died due to Cardiovascular Disease.

- The total number of people dead are = $[(0.28 * 7500000) + (0.58 * 15200000) + (0.14 * 4250000)]$
= 2,100,000 + 881,600 + 595,000
= 3,576,600

The value of node D is 3,576,600.

Step-4 Calculation of the number of people died due to Chronic Respiratory Disease.

- The total number of people dead are = $[(0.78 * 338000) + (0.22 * 3170000)]$
= 263,640+697,400
= 961,040

The value of node E is 961,040.

Thus, the value of node A is 3576600 (value of node D). The node with the maximum value becomes the most optimal alternative among all the alternatives.

4. Methodology

There are some complex decision trees whose analysis can be made easy through the inbuilt programs and software. We have also done the analysis of our decision tree with the help of the program. We have made the program in java language and executed it on the NetBeans platform.

```
//@author name
```

```
package solutions;
```

```
// These are libraries
```

```
import java.util.Arrays;
```

```
import java.util.Scanner;
```

```
public class Solutions
```

```
{
```

```
    public static void main(String[] args)
```

```
    {
```

```
        Scanner input = new Scanner(System.in);
```

```
        // Input from the user is taken regarding the starting node.
```

```
        System.out.print("Enter the starting node ");
```

```
        String a = input.nextLine();
```

```
        System.out.println("the starting point is:" + a);
```

```
// The number of sub nodes will help us in executing the loop.

// The loop will run as many times as the value of sub nodes.

// In our decision tree the value of h is 4

System.out.print("Enter the number of sub nodes ");

int h = input.nextInt();

System.out.println("the number of sub nodes are:" + h);

double array2[]= new double[h];

int x=0;

// In this while loop we will take further input from user regarding the tree.

while(x<h)

{

    System.out.print("Further details of tree. 0 means square node is present.

                    1 means circular node is present ");

    int decision;

    decision= input.nextInt();

    switch(decision)

    {

        // This is the case which will execute when the is square node.

        // In our decision tree this case will execute for the disease cancer.

        // In short this case will execute every time when decision is to be made.

        case 0:

            System.out.print("Enter the number of circular nodes: ");

            int c = input.nextInt();

            System.out.print("The number of circular nodes: " +c);
```

```
Double[] arr1=new Double[c];

for(int i=0;i<c;i++)

{

    double sum =Badamultiplication();

    arr1[i] = sum;

    System.out.println("Sum of numbers is "+arr1[i]);

}

double max=0;

int m=0;

if(arr1[m]<arr1[m+1])

    max=arr1[m+1];

System.out.println("maxium value is "+max);

array2[x]=max;

x++;

break;
```

// This is the case which will execute when there is circular.

//In our decision tree this case will execute for the remaning 3 diseases.

// In short this case will execute when there is outcomes and summation

// of all outcomes will five us the value of circular node

case 1:

```
int k=0;

System.out.print("Enter the number of types");

int t = input.nextInt();

System.out.print("The number of types: " +t);
```

```
double array3 []= new double[t];

double array4[]= new double[t];

for(int i=0;i<t;i++)

{

    System.out.print("Enter the values of" +(i+1)+ "probability: ");

    array3[i]= input.nextDouble();

    System.out.print("Enter the values of" +(i+1)+ " rate: ");

    array4[i]= input.nextDouble();

}

double sum=0.0;

double array5 []= new double[t];

for(int z=0; z<t; z++)

{

    array5[z] =array3[z]*array4[z] ;

    sum+= array5[z];

}

System.out.println("The final death of people is:"

+Arrays.toString(array5));

System.out.println("Sum of numbers is =" +sum);

array2[x]=sum;

x++;

break;

default:

System.out.println("Please enter again");
```

```
    }  
  }  
  
  // The array2 stores the final values of all the nodes. In our case node B,C,D and E  
  
  for(int f=0;f<h;f++)  
  
  {  
  
    System.out.println("Array elements are "+array2[f]);  
  
  }  
  
  // Here the comparison is done in all the array values to find out the maximum value.  
  
  // In our case the final maximum values will become the value of node A  
  
  double max1=0.0;  
  
  for( int l=0;l<h;)   
  
  {  
  
    if(array2[l]>max1)  
  
    {  
  
      max1= array2[l];  
  
      l++;  
  
    }  
  
    else  
  
      l++;  
  
  }  
  
  System.out.println("the final maximum element is"+max1);  
  
}
```

// In this function all probability and estimated number of death of people due to each disease
//is asked from user and then multiplied correspondingly.

```
public static double Badamultiplication()
{
    int i;
    Scanner input = new Scanner(System.in);

    System.out.print(" Enter the number of probability ");

    int a = input.nextInt();

    double array6[]=new double[a];

    double array7[]=new double [a];

    for( i=0;i<a;i++)
    {
        System.out.print("Enter the values of" +(i+1)+ "probability ");

        array6[i]= input.nextDouble();

        System.out.print("Enter the values of" +(i+1)+ " rate ");

        array7[i]= input.nextDouble();

        double array8[]= new double[a];

        double sum=0.0;

    for (int k=0; k<a; k++)
        {
            array8[k] =array6[k]*array7[k] ;

            sum+= array8[k];

        }

        System.out.println("The final death of people is:" +Arrays.toString(array8));

        return sum;

    }
}
```

5. Results and Discussion

In this section, we have provided various comparisons between diseases through pie charts and bar graphs. After tracing the path of each alternative following results were obtained from fig 4 - fig10. Fig 4 shows the estimated number of men dead globally, all ages due to different types of cancer as discussed in the decision tree model. From fig 4, we can also observe that lung cancer is the main reason for death, and it accounts for 52% of death among men. Fig 5 shows the estimated number of women dead globally, all ages due to different types of cancer as discussed in the decision tree model. From fig 5, we can also observe that breast cancer is the main reason for death, and it accounts for 52% of death among women. Fig 6 shows the comparison between the estimated number of men and women who died due to some type of cancer. From the bar graph, we can observe that for each type of cancer number of men dead is greater than the number of women dead. Fig 7 shows the estimated number of people who died due to diabetes. We can also observe that a greater number of people are dead due to type-2 diabetes. Fig 8 shows the estimated number of people died due to chronic respiratory disease. We can also see from the fig that the number of people is dead due to COPD. Fig 9 shows the estimated number of people died due to cardiovascular disease. From the fig, we can observe the most common reason for many people to be dead is because of blood pressure which accounts for 59% of total death. Fig 10 shows the total estimated number of people died due to four major disease as discussed in the decision tree. From table 2 and table 3 we can observe the effect of variation in probabilities keeping the rate same on the number of men and women died respectively. We can also see that even change of ± 0.2 in probability brings drastic change by thousands (increase or decrease), on the number of men and women who died globally.

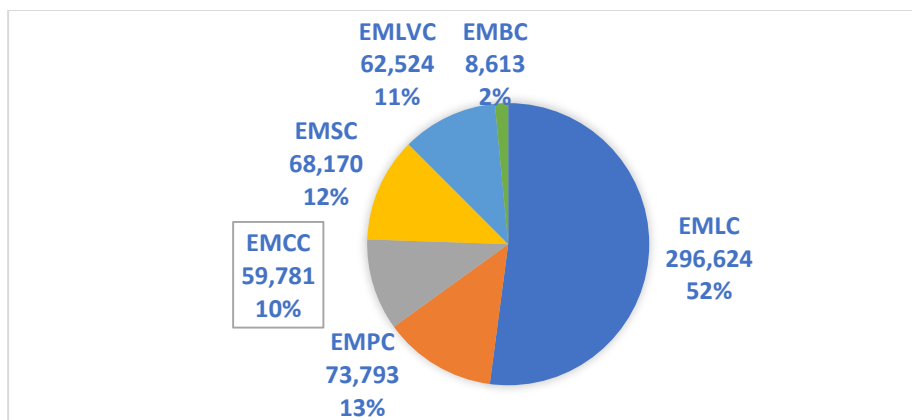


Fig 4: Estimated number of men died due to cancer, all ages

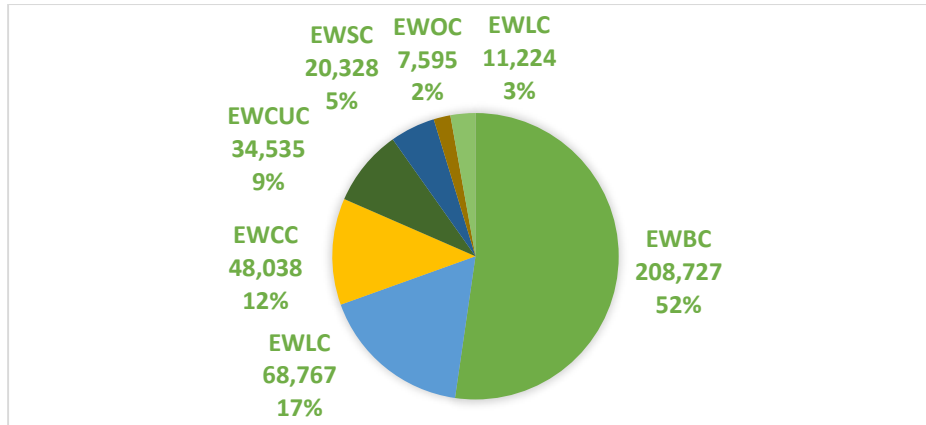


Fig 5: Estimated number of women died due to cancer, all ages

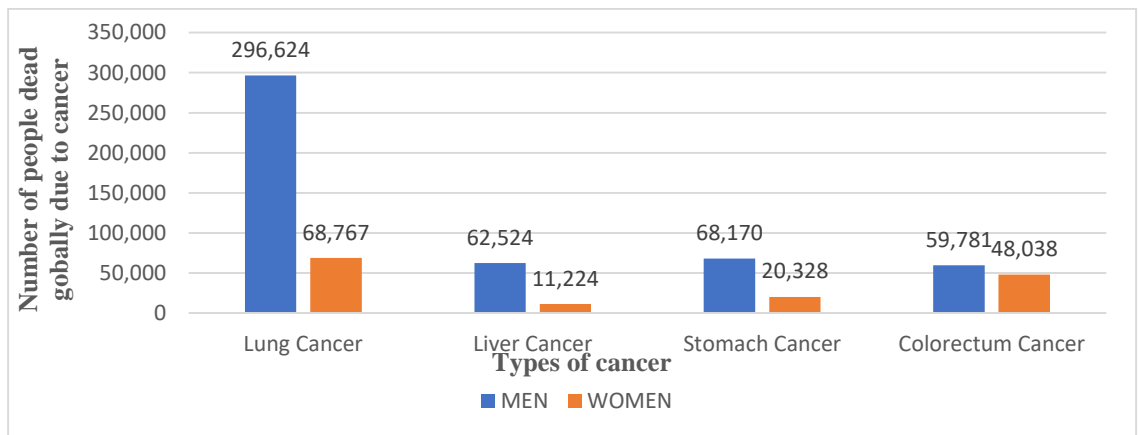


Fig 6: Comparison between the estimated number of men and women died due to cancer.

		Types of cancer in men					
	Probabilities (P1,P2,P3,P4,P5,P6)	EMLC	EMPC	EMCC	EMSC	EMLVC	EMBC
Initial probabilities	(0.27, 0.24, 0.16, 0.14, 0.12, 0.07)	296,624	73,793	59,781	68,170	62,524	8,613
Changed probabilities	(0.27, 0.26 , 0.16, 0.12 , 0.15 , 0.07)	296,624	79,942	59,781	58,432	78,155	8,613
	(0.23 , 0.24, 0.16, 0.14, 0.12, 0.09)	252,679	73,793	59,781	68,170	62,524	11,074
	(0.27, 0.24, 0.13 , 0.14, 0.12, 0.07)	296,624	73,793	48,572	68,170	62,524	8,613

Table 2 Effect of variation in probabilities on number of men died due to types of tumor

		Types of cancer in women					
--	--	--------------------------	--	--	--	--	--

	Probabilities (P1,P2,P3, P4,P5,P6,P7)	EWBC	EWLC	EWCC	EWCUC	EWSC	EWOC	EWLVC
Initial probabilities	(0.40,0.14,0.15, 0.13,0.08,0.05, 0.05)	208,727	68,767	48,038	34,535	20,328	5,595	11,224
Changed probabilities	(0.35 ,0.14,0.15, 0.13,0.08,0.05, 0.07)	182,636	68,767	48,038	34,535	20,328	5,595	15,714
	(0.40, 0.12 ,0.15, 0.10 ,0.08, 0.09 , 0.05)	208,727	58,943	48,038	26,565	20,328	13,671	11,224
	(0.40,0.14, 0.17 , 0.13, 0.09 ,0.05, 0.05)	208,727	68,767	54,443	34,535	22,868	5,595	11,224

Table 3 Effect of variation in probabilities on number of women died due to types of tumor

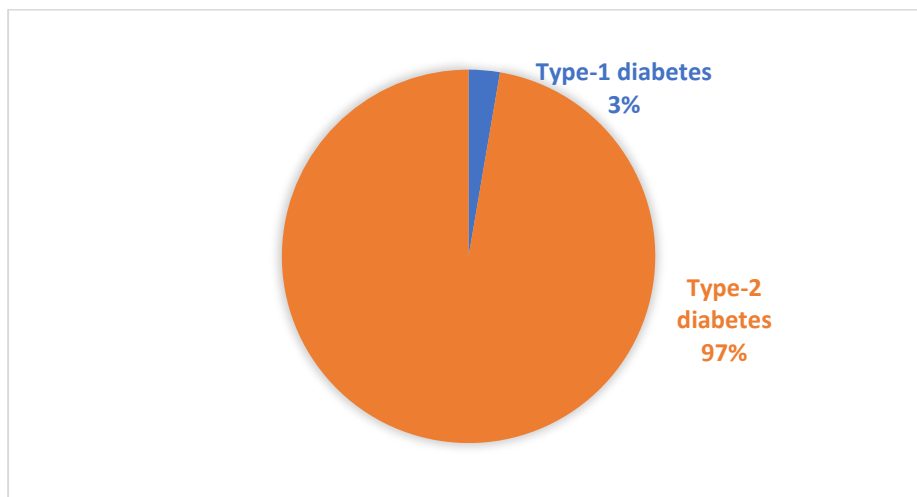


Fig 7: Estimated number of people died due to diabetes

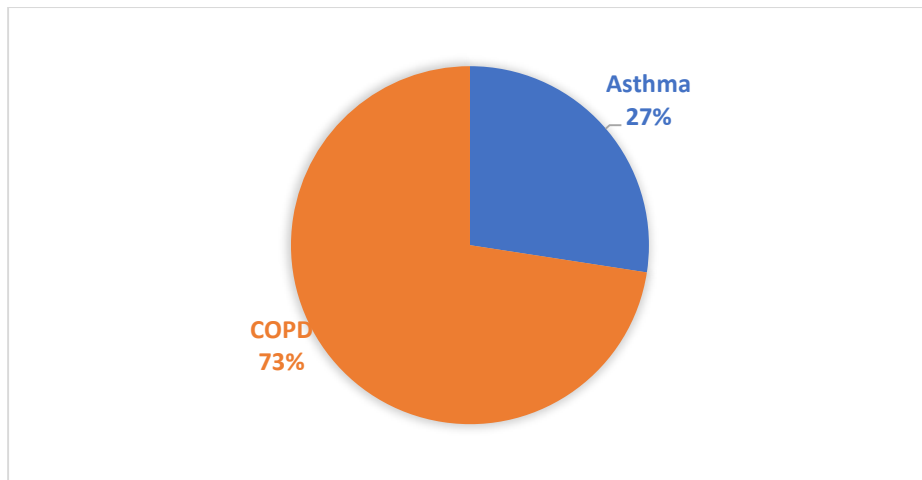


Fig 8: Estimated number of people died due to Chronic respiratory disease.

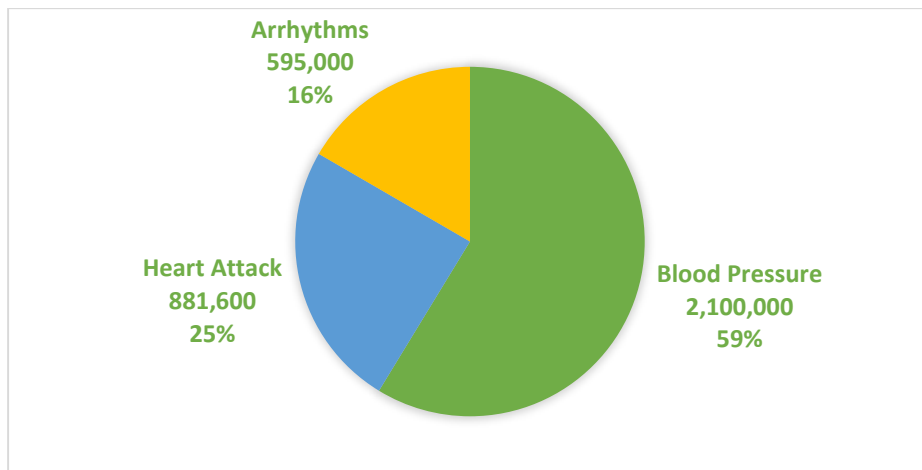


Fig 9: Estimated number of people died due to Cardiovascular Disease

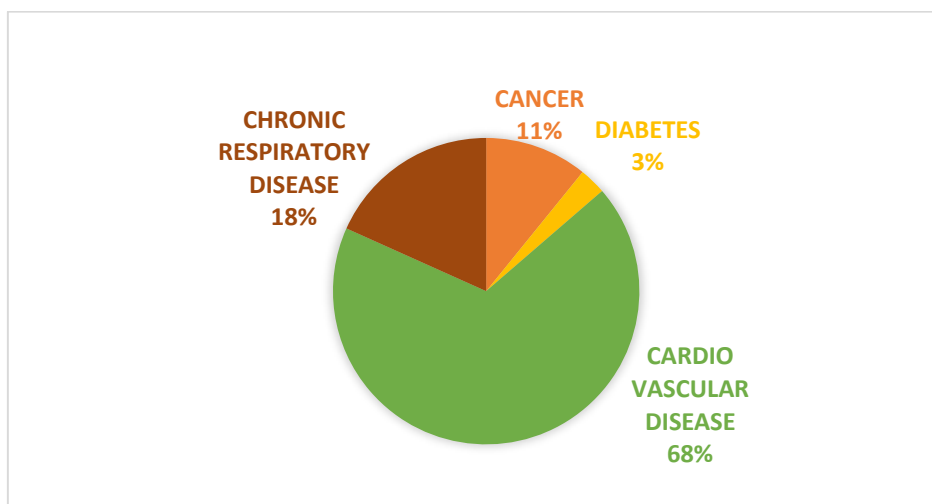


Fig 10: Major death-causing diseases affecting people globally

6. Conclusion

The quality of life depends on many factors but, one of the factors of major concern is health. The growing number of people getting affected due to deadly disease is itself the proof of the ineffectiveness of the country's health system. Our study of death-causing disease, using a decision tree model has helped us to identify the killer disease. Our study concludes that the major death-causing disease is cardiovascular disease followed by chronic respiratory disease followed by cancer and the least number of people were dead due to diabetes globally. We were also able to compare how different types of tumors affect men and women respectively. From the comparison, it was also found that men are more likely to suffer from different types of tumor than women. Globally, where thousands of deadly disease are standing to be fought to save people, our study helps in prioritizing various deadly disease in order to reduce the number of preventable deaths by taking major steps for providing effective treatment for some of the major death-causing diseases so as to improve the health of people.

References

1. <https://www.who.int/cancer/PRGlobocanFinal.pdf>
2. Bae-M. (2014) 'The clinical decision analysis using decision tree' *Epidemiol Health*, vol 36(0): e2014025–2014020. [PMC free article] [PubMed] [Google Scholar]
3. Vili Podgorelec, Peter Kokol, Bruno Stiglic, Ivan Rozman (2002) 'Decision Trees: An Overview and Their Use in Medicine' *Journal of Medical System*, vol 25(5):445-63.
4. https://www.researchgate.net/publication/11205595_Decision_Trees_An_Overview_and_Their_Use_in_Medicine
5. <https://www.drugsandalcohol.ie/28525/1/World%20Cancer%20Report.pdf>
6. <https://www.who.int/respiratory/copd/en/>
7. <https://www.healthline.com/health/diabetes/facts-statistics-infographic#3>
8. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6020177/>