Analysis of Thalassemia Patients Data Using Statistical Methods

Arfa Maqsood[1], Sehrish Iftikhar[2]
[1, 2] Department of Statistics, University of Karachi, Karachi, Pakistan.

Abstract— The purpose of this study is to concern with statistical methods to detect information from thalassemia diseases data. Data collected from Kashif Iqbal Thalassemia care Center of 220 individuals on seven variables. The seven variables are Age, Sex, Weight, Height, Hemoglobin (G/Dl), Packed Cell Volume (PCV, %) and Red Blood Cell (RBC, 10'6/cmm3). Data is in cross sectional form. Firstly, we explain the data with the help of some descriptive analysis. Afterwards, we employ the simple statistical techniques such as correlation, body mass index and regression analysis to explore the dependence relations between the variables.

Index Terms— Statistical Methods, Thalassemia Disease, Correlation, Body Mass Index, Regression Analysis.

1 INTRODUCTION

BIOSTATISTICS or biometry is the application of statistics to a wide range of topics in biology. It has particular applications to medicine and to agriculture. Public health, including epidemiology, health services research, nutrition, and environmental earth. Design and analysis of clinical trials in medicine Genomics, population genetics, and statistical genetics in populations in order to link variation in genotype with a variation in phenotype. This has used in agriculture to improve crops and farm animals. In biomedical research, this work can assist in finding candidates for gene alleles that can cause or influence predisposition to disease in human genetics Ecology Biological sequence analysis Statistical methods are beginning to be integrated into medical informatics, public health informatics, and bioinformatics
Biostatistics is, in effect, two words and two fields of study combined. The *bio* part involves biology, the study of living things. The *statistics* part involves the accumulation, tracking, analysis, and application of data.
Another related term with the diseases study is 'Epidemiology'. In the main, people attempting to define epidemiology have normally done so in the context of their own particular interests or needs. A useful general definition defines the epidemiology as the study of disease in populations. It thus differs from the more conventional medical approaches to the study of disease that is normally concerned with the study of disease processes in affected individuals. While the objective of the latter is to find cures for diseases in individuals already affected, epidemiology concerned with the reasons why those individuals became diseased in the first place. Inherent in the epidemiological approach is the belief that the frequency of occurrence of a disease in a population is governed by the interaction of a large number of different factors or determinants. The epidemiologist believes that by studying these interactions it may become possible to manipulate some of the determinants involved, and so reduce the frequency with which the disease in question occurs in a population.
The paper is organized as follows. Section 2 provides with a brief description about thalassemia disease and its types. The analysis on thalassemia data using statistical methods is carried out in section 3. Lastly, section 4 presents the conclusion.

2 THALASSEMIA AND ITS TYPES: A BRIEF REVIEW

Thalassemia name is derived from the Greek word "thalassa" meaning "the sea" because the condition was first described in populations living near the Mediterranean Sea. (Source: Genes and Disease by the National Center for Biotechnology). Thalassemia is an inherited disease of the blood. In thalassemia quantity of hemoglobin is reduced or absence, it means patient has anemic.
Thalassemia is a genetic disorder (see Marengo-Rowe [1]). It cannot be developed from contact with other people or from the environment. All types of thalassemia, the quantity of hemoglobin reduced or absent. This situation affects the ability of the blood to carry oxygen to all parts of the body. Hemoglobin made with the help of three components (i) alpha globin (ii) beta globin and (iii) heme.
The types of thalassemia described with the help of hemoglobin for example Galanello and Origa [2] for more detail on thalassemia. It distributes it into two parts Alpha thalassemia and Beta Thalassemia, these are briefly discussed in the following sub-sections.

2.1 Alpha Thalassemia

Alpha thalassemia developed when the body has problem to produce alpha globin or person has anemic. It means body has few blood cells or there is too little hemoglobin in them. There are two main types of alpha thalassemia disease: hemoglobin H disease and alpha thalassemia major. Alpha thalassemia major has inherited that results in severe anemia that beings before birth. Most children who are affected; they do not survive to be born or die shortly after birth. Four genes are involved in making the alpha hemoglobin chain (see Arif, F. et al. [3]). Everyone gets two from each of his parents. If anyone inherits:
One mutated gene, it has no signs or symptoms of thalassemia. However, they are carrier of the disease and can pass it on to their children.
Two mutated genes, in this thalassemia signs and symptoms will be mild. This condition called alpha-thalassemia minor, or it may tell an alpha- thalassemia trait.
Three mutated genes, its signs and symptoms will be moderate to severe. This condition called hemoglobin H disease.
Four mutated genes, the condition is called alpha-thalassemia

major or hydrops fetal. It usually causes a fetus to die before delivery or a newborn to die shortly after birth.
In alpha thalassemia, some child has no problem and no symptoms; they do not require treatment. Alpha thalassemia is inheriting from his or her parents. The most severe form of alpha thalassemia major causes stillbirth (death of the unborn baby during birth or the late stages of pregnancy). In thalassemia, major children are normal at time of birth, but developed severe anemia during the first year of life. The symptoms include bone deformities in the face, fatigue, growth failure and yellow skin.

2.2 Beta Thalassemia

Beta thalassemia is known as Cooley's anemia, a well-known type of thalassemia. In beta thalassemia, when body has problem to produce beta globin component of hemoglobin. In people with beta thalassemia, low levels of hemoglobin lead to a lack of oxygen in many parts of the body. Beta thalassemia intermediate is a clinical term that describes the disease in individuals who have moderate anemia that only requires blood transfusions intermittently (see Perrine [4] and Olivieri [5]).
There are two type of beta thalassemia which is depending on the severity of symptoms thalassemia major (also known as Cooley's anemia) and thalassemia intermedia. It is caused by a change in the gene for the beta globin component of hemoglobin. Beta thalassemia causes variable anemia that can range from moderate to severe, depending in part on the exact genetic change underlying the disease. Beta thalassemia usually occurs within three to six months after birth. If left untreated, severe anemia can result in stunted growth and development, as well as other characteristic physical complications that can lead to a dramatically decreased life expectancy. Two genes are involved in making the beta hemoglobin chain. Everyone gets one from each of his parents. If you inherit:
One mutated gene, its mild signs and symptoms. This condition called beta-thalassemia minor or referred to as a beta-thalassemia trait.
Two mutated genes, its signs and symptoms will be moderate to severe. This condition called beta-thalassemia major or Cooley's anemia. Babies born with two defective beta hemoglobin genes usually are healthy at birth, but develop signs and symptoms within the first two years of life.
Humans normally make several types of hemoglobin. An individual's stage in development determines whether he or she makes primarily embryonic, fetal, or adult hemoglobin. All types of hemoglobins are made of three components: heme, alpha globin, and beta globin. All types of thalassemia caused by changes in either the alpha- or beta-globin gene. These changes cause little or no globin to produce. All types of thalassemia recessively inherited, meaning that a genetic change must inherit from both the mother and the father to produce the disease in the child. The severity of the disease influenced by the exact thalassemia mutations inherited, as well as other genetic and environmental factors. There are rare exceptions, notably with beta thalassemia, where globin gene mutations exhibit a dominant pattern of inheritance in which only one gene need altered in order to see disease expression. Signs and symptoms of thalassemia include fatigue, weakness, shortness of breath, pale appearance, irritability, yellow discoloration of skin (jaundice), facial bone deformities, slow growth, abdominal swelling and dark urine. The signs and symptoms you experience depend on the type and severity of thalassemia you have. Some babies show signs and symptoms of thalassemia at birth, while others may develop signs or symptoms during the first two years of life. Some people who have only one affected hemoglobin gene do not experience any thalassemia symptoms.

3 ANALYSIS OF THALASSEMIA

This section presents the analysis of thalassemia using the appropriate statistical methods. Section 3.1 describes the data description and descriptive analysis of thalassemia data is presented in section 3.2. This section also presents the regression analysis.

3.1 Data Description

Thalassemia is a major genetic and hereditary disorder of public health in Pakistan. Around 6,000 new patients added every year to the existing population of the sufferers. 10 million carriers of this disease transmit it to the next generation. Economic cost of treatment of Thalassemia is enormous. The data is concern with thalassemia patients in the form of those variables those depend on thalassemia. Statistical techniques are used in much area of investigations, in the field of medical research. In this study, the data is collected from Kashif Iqbal thalassemia care center. The data is based on cross sectional form. The main aim of collecting data to study the idealistic range of a given factors. The data is taken from 220 thalassemia patients on seven variables. These seven variables are Age (years), Sex, Weight (kg), Height (cm), Hemoglobin (G/Dl), Packed Cell Volume PCV (%) and Red Blood Cell (10'6/cmm3).

Key terms:

i) Hemoglobin: Hemoglobin in the blood carries oxygen from the respiratory organs (lungs or gills) to the rest of the body (i.e. the tissues) where it releases the oxygen to burn nutrients to provide energy to power the functions of the organism and collects the resultant carbon dioxide to bring it back to the respiratory organs to be dispensed from the organism.
ii) Packed cell volume (PCV in %): a measure of the proportion of blood volume that is occupied by red blood cells.
iii) Red Blood Cell (RBC): Red blood cells or erythrocytes are the most common type of blood cell and the vertebrate organism's principal means.

3.2 Descriptive Analysis

Descriptive statistics provides simple summaries about the sample and about the observations that have made. When a sample consists of more than one variable, descriptive statistics may be used to describe the relationship between pairs of variables. Table 1 presents the distribution of Thalassemia patients with respect to age and sex. There are higher probability to get the disease in children whose ages are in between

1 to 5 years. This probability get decreases with increasing age and only 4.3 percent chance to have the disease for the age group 15-20 years. These findings can be better visualized in figure 1.

TABLE 1
Thalassemia Patients With Respect to Age And Sex

Sex/Age	1-5	5-10	10-15	15-20	Total	% Sex
Male	77	36	14	8	135	61.36
Female	43	26	15	1	85	38.63
% Age	54	28.7	13	4.3	100	100

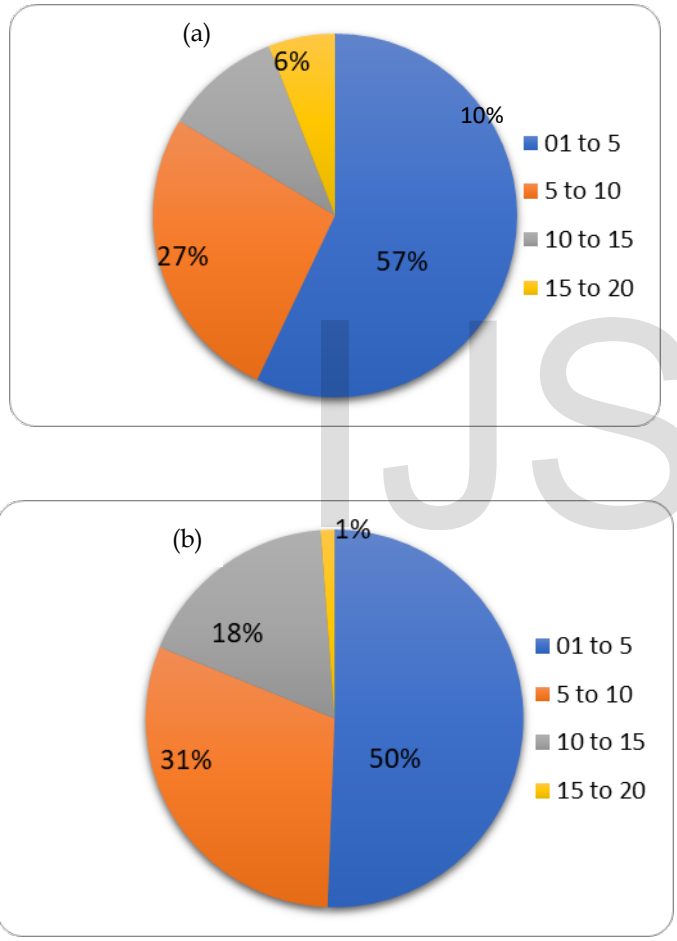

Fig. 1. Thalassemia patients with respect to age (a) for males (b) for females

From the male pie chart, we find that the persons whose ages 1-5 are 57% affected, 5-10 age groups are 27%, 10-15 age groups are 10% and 15-20age groups are only 6% affected. Similarly, by the help of female pie chart, we find that the persons whose ages 1-5 are 50% affected, 5-10 age groups are 31%, 10-15 age groups are 18% and 15-20age groups are only 1% affected. By this result, it can reveal that the persons aged 1-5 years group has greater chance to affect thalassemia disease regardless of gender.

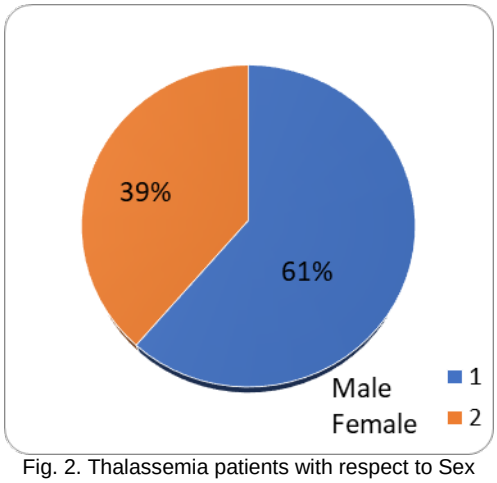

Fig. 2. Thalassemia patients with respect to Sex

Figure 2 presents the pie chart showing the distribution of patients with respect to gender. The male ratio is 61% and female ratio is 39%. We can say that male is highly affected for thalassemia disease as compare to female.

The body mass index (BMI), or Quetelet index, is estimating human body fat based on an individual's weight and height. Frequent use of the BMI is to assess how much an individual's body weight departs from what is normal or desirable for a person of his or her height. BMI is one method used to estimate the total amount of body fat. It is calculated by dividing the weight in kilograms by the height in meters squared (m2) that is;

$$BMI = \frac{Mass\ (kg)}{Height\ (m^2)}$$

The formula was previously called the Quetelet Index for BMI. The BMI generally used as a means of correlation between groups related by general mass and can serve as a vague means of estimating adiposity. Generally, the index is suitable for recognizing trends within sedentary or overweight individuals because there is a smaller margin for errors. Table 2 provides the BMI range for various categories based on the weight of a person. These ranges are provided by World Health Organization (WHO).

TABLE 2
BMI Ranges and BMI Primes given by WHO

Category	BMI range – kg/m²	BMI Prime
Very severely underweight	less than 15	less than 0.60
Severely underweight	from 15.0 to 16.0	from 0.60 to 0.64
Underweight	from 16.0 to 18.5	from 0.64 to 0.74
Normal (healthy weight)	from 18.5 to 25	from 0.74 to 1.0
Overweight	from 25 to 30	from 1.0 to 1.2
Obese Class I (Moderately obese)	from 30 to 35	from 1.2 to 1.4
Obese Class II (Severely obese)	from 35 to 40	from 1.4 to 1.6
Obese Class III (Very severely obese)	over 40	over 1.6

Overweight is defined as BMI 25–30 kg/m^2 and obesity as BMI 30 kg/m^2 and above according to the WHO criteria. Another term BMI prime is calculated by the following formula.

$$BMI\ Prime = \frac{BMI}{23}$$

BMI Prime is the BMI relative to the upper limit for normal weight. i.e. when the BMI is at the upper limit for normal weight, BMI Prime is 1.00. A BMI Prime value larger than 1.0 indicates overweight. For normal weight (BMI 18.5 - 25), the value of BMI Prime is between 0.74 and 1.0.

TABLE 3
BMI Ranges for Thalassemia Patients

BMI Range	C.F	F	Category
1 - 15	92	92	Very severely underweight
15-16	132	40	Severely underweight
16-18.5	185	53	Underweight
18.5-25	213	28	Normal (healthy weight)
25-30	219	6	Overweight
30-35	220	1	Obese Class I (Moderately obese)
Total	-	220	-

Table 3 presents the BMI ranges for thalassemia patients. According to this information 185 patients out of 220 patients are underweight that makes 84 percent of total patients. While only 12.7 percent of the patients are healthy (normal weight). Among these 6 persons are overweight and only 1 is very severely obsess.

In the next phase, the coefficient of correlation is calculated between different pairs of variables to see the linear dependence between the variables. Table 4 presents the coefficients of correlations. The correlation between age and weight, age and hight, and height and weight are highly linear correlated as expected (0.86, 0.83 and 0.78 respectively). They all have a positive relation means one variable is increased (decreased) by increasing (decreasing) the value of other variable.

TABLE 4
Correlations Between Various Pairs of Variables

S.No	Variables	Correlation
1	Age vs. weight	0.86
2	Age vs. height	0.83
3	Height vs. weight	0.78
4	Hemoglobin vs. Age	-0.09
5	Hemoglobin vs. Sex	-0.03
6	Hemoglobin vs. PCV	0.82
7	Hemoglobin vs. RBC	0.94
8	RBC vs. PCV	0.86

The correlation between hemoglobin (HB) and other variables are computed and the results indicate that HB has a positive high correlation when compared with the variables PCV and RBC. However, HB is not directly linearly related with age and sex as the values of coefficients are close to zero that is ignorable. Lastly, the correlation between RBC and PCV is computed and found a highly and positive linear relationship.

When a sample offers evidence that paired variations are linearly related, the next step is to specify the exact nature of that linear relationship. Regression analysis is the process of obtaining displaying and applying linear relationship using a suitable model based on a random sample with independent observations. A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable (see Montgomery, D. C. et al. [6]). The slope of the line is b, and a is the intercept. We have already found the correlations between the variables now employ the regression technique to find the dependency of PCV and RBC on hemoglobin. The regression equations and the associated results are well displayed in table 5.

TABLE 5
Regression Equations and Associated Results

PCV = 4.42 + 6.3 HB	RBC = 0.163 + 0.95HB
SE: (1.09) (0.29)	SE: (0.07) (0.021)
P-value (0.00) (0.00)	P-value (0.03) (0.00)

The regression equation shows if HB increases by one unit then PCV increases by 6.3 while other variables held constants. If HB is zero then PCV is 4.42%. the zero p-value indicate that there is a significant contribution of HB in order to predict the value of PCV as also obtained a high correlation between them. The slandered error of HB is less that shows its reliability. The other simple linear regression equation is obtained between HB and RBC to verify the dependency between them shown by correlation coefficient. The value of RBC is increased by 0.95 with one-unit increment of HB. A small p-value shows the significance of HB in determining the value of RBC, however, HB has a greater impact on estimating PVC of a patient rather than his or her RBC.

4 CONCLUSION

In this article, we used some statistical methods to obtain information from thalassemia diseases data. Data on seven variables were used. The details of data are given in section 3. The descriptive analysis of data showed that the males are at higher risk of getting thalassemia disease as compared to females. Similarly, this risk is also higher at early age of one to five years. The body mass index (BMI) of thalassemia patients revealed that these patients are underweight.

Afterwards, the statistical techniques were employed to explore the dependence relations between the variables under consideration. A high positive correlation was obtained between the variables age versus weight, age versus height, height versus weight, HB versus PCV, HB versus RBC, and RBC versus PVC. Based on the results of correlation coefficients, the regression equations were fitted and found hemoglobin significant variable to predict the PVC and RBC of a thalassemia patient.

REFERENCES

[1] J. Marengo-Rowe, "The Thalassemia and Related Disorders," Baylor University Medical Center Proceedings, vol. 20, no. 1, Jan. 2007.

[2] R. Galanello and R. Origa, "Beta-Thalassemia," Orphanet Journal of Rare Diseases, pp. 5-11, 21 May, 2010.

[3] F. Arif, J. Fayyaz and A. Hamid, "Awareness Among Parents of Children with Thalassemia Major," Journal of the Pakistan Medical Association, vol. 58, no. 11, pp. 621-624, November 2008.

[4] SP Perrine, "Fetal Globin Induction—Can it Cure β Thalassemia?" Hematology, pp. 38-44, 2005.
[5] NF Olivieri, "Reactivation of Fetal Hemoglobin in Patients With ß-Thalassemia," Seminars in Hematology, vol. 33, no. 1, pp. 24-42, 1996.
[6] D.C. Montgomery, E. A. Peck, and G. G. Vining, "Introduction to Linear Regression Analysis," Fifth Edition, Wiley, 2013.