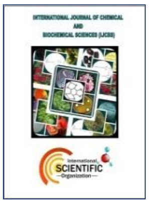
International Journal of Chemical and Biochemical Sciences
(ISSN 2226-9614)

Journal Home page: www.iscientific.org/Journal.html

© International Scientific Organization



Analyzing Blood Abnormalities Using Complete Blood Count Data: A Statistical Approach

Meenakshi Suresh[1*], Kshipra Khadkikar[2], Shravani Daware[3], Siddhi Salvi[4]

[1] Department of Chemistry, [2, 3, 4] Department of Statistics, Fergusson College (Autonomous), affiliated to Savitribai Phule Pune University, 411004*

Abstract

The current research study involved scrutinizing and comparing complete blood count (CBC) test parameters across people with three types of blood disorders: anaemia, leukaemia and thrombosis which aimed to unravel distinctive patterns and variations that could enhance the understanding of disease pathology. The project's main objective included studying the various blood parameters affecting an individual's overall health and checking the relation and association of various blood parameters from CBC test reports. During the study, it was observed that certain CBC parameters were correlated with patterns of clinical symptoms shown by patients and with an increased risk of developing certain blood-related disorders. The study analyzed differences in male and female CBC reports across various age groups.

Keywords: Complete Blood Count (CBC) test, Hematology, Regression and Correlation, Cluster Analysis, Biostatistics

Full length article **Corresponding Author*, e-mail: -meenakshi.suresh@fergusson.edu

1. Introduction

Hematology is a branch of internal medicine that deals with physiology, pathology, aetiology, diagnosis, treatment, prognosis and prevention of blood-related disorders. Blood disorders are pathological conditions which primarily affect the blood or blood-producing organs, causing blood cells to function abnormally, potentially leading to various diseases or impacting overall health. A Complete Blood Count (CBC) test is a widely used hematology test that helps diagnose blood-related disorders and monitor conditions like blood loss or infection [1]. CBC is often conducted during routine health check-ups. It is a fundamental diagnostic tool, providing insights into red and white blood cell counts, platelet levels, and other important metrics. Anaemia, leukaemia, thrombosis, lymphoma, and sickle cell diseases are common blood-related disorders that can be detected and monitored by a CBC test [2]. The current study mainly focuses on three major types of blood disorders: anaemia, leukaemia, and thrombosis. A comprehensive statistical analysis of CBC parameters of the selected three blood disorders was performed, and the results were interpreted. CBC parameters may indicate significant correlations between the three disorders, but more medical tests are essential for a confirmed diagnosis.

2. Methodology

The research was carried out using primary data, which included 610 CBC reports of people collected from three pathology labs in Pune. The data was anonymized to prevent any disclosure of the personal information of the people whose data was used. The hematological parameters were extracted and used only for analysis and research purposes. These reports included 13 hematological parameters: Hemoglobin (Hb), Hematocrit (HCT), White Blood Cell count (WBC), Red Blood Cell count (RBC), Platelet count, Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC), Neutrophils, Lymphocytes, Monocytes, Eosinophils and Red Cell Distribution-Width (RDW-CV). Data cleaning was performed to ensure the accuracy and correctness of all entries. As most parameters did not exhibit characteristics of normal distribution, Box-Cox transformation [3] employed to normalize data. This transformation facilitated use of parametric tests such as correlation and regression, enabling more precise analysis of blood components. By employing these methods, study aimed to gain insights into relationship between age, gender, and hematological disorders.

2.1. Statistical analysis

R software (version 4.3.2) and MS Excel (2007) were used for statistical analysis. Cluster analysis using complete linkage and k-means clustering methods adopted to observe similarities and differences between various blood parameters across different age groups [4]. All blood parameters analyzed using Pearson correlation analysis [5].

Bio statistical analysis of the three blood disorders (Anaemia, Leukaemia, and Thrombosis) was conducted to obtain the percentage of the population at risk of developing disorder by examining male and female CBC test reports across various age groups. Multiple Logistic Regression analysis was executed to determine the significance of age and gender in predicting diagnosis of anaemia, leukaemia and thrombosis.

The model $Y = \prod(X) + \varepsilon$

Where, $\prod(X) = e\text{xp } (\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k) \ / \ 1+ (e\text{xp } (\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)$

And, ε is the error component; β_0, β_1, β_2, ..., β_k are the regression coefficients employed for multiple logistic regression having the hypothesis: H_0: $\beta_1= \beta_2=0$ v/s H_1: $\beta_1= \beta_2 \neq 0$ was utilized to assess the significances of blood disorders around age and gender.

3. Results and discussions

As seen in Figure 1, among all the three blood disorders, the number of people diagnosed with anaemia is higher than that with leukaemia and thrombosis. More females diagnosed with anaemia and leukaemia than males; however, opposite pattern is observed in case of thrombosis. Figure 2 exhibits a right-skewed distribution. Proportion of people having blood disorders is higher in age group of 31 to 40, followed by age groups 21 to 30 and 41 to 50. Likelihood of the blood disorders decreases after the age group of 41-50 [6].

3.1. Cluster analysis

Figures 3, 4 and 5 represent the cluster dendrograms of CBC parameters for age groups 1-10, 41-50 and 81-90, respectively. These age groups were chosen to represent young, middle-aged and older individuals. In all three figures, two primary clusters can be observed. Each figure shows the first subgroup ranging from MCV to MCHC in Figure 3, from MCV to MCH in Figure 4, and from the MCV to the Monocytes in Figure 5, each followed by a second subgroup consisting only of WBC levels. The final cluster consists solely of platelet count, indicating that this parameter is significantly distinct from all the other measured blood parameters or may have unique characteristics. These parameters are grouped together, reflecting their relative similarity in particular age groups [7]. Figure 6 shows a cluster of three groups: leukaemia, anaemia, and the thrombosis. The red portion indicates the Leukaemia, the blue portion indicates Thrombosis, and the green portion indicates the Anaemia. The central area overlaps all three blood disorders, suggesting the similarities between the bloods parameters shared in these conditions [8].

3.2. Correlation

Figure 7 shows the correlation between all the CBC parameters. The Pearson correlation coefficient between Hb and HCT is 0.8, which is close to 1, indicating a strong relation between these two parameters, suggesting the presence of anaemia in the data. There is absolutely no correlation between many blood parameters valued at 0. A negative correlation between neutrophils and lymphocytes of -0.8 indicates no prominent diagnostic findings [9].

3.3. Bio statistical findings

- ***Prevalence rates***

The data analysis results showed variations in the prevalence of different diseases. 28% of the population was diagnosed with anaemia, indicating a substantial incidence of this condition. Moreover, 21% of the population was diagnosed with leukaemia, as against only 1.96% of the population being diagnosed with thrombosis.

Risk Ratio: From the collected data, a comparison between males and females was made to predict which gender was at a greater risk of the disease.

Anaemia: 0.8018949
Leukaemia: 1.005233
Thrombosis: 1.0330

The relative risk for anaemia was less than 1, while for leukaemia and thrombosis, it was found to be greater than 1. Males had a decreased risk of being diagnosed with anaemia compared to females. Further, males had an increased risk of being diagnosed with both leukaemia and thrombosis compared to females [10].

- ***Odds ratio***

The likelihood of being diagnosed with specific diseases differs between males and females. A person diagnosed with anaemia is 0.43% less likely to be male than female. Conversely, a person diagnosed with leukaemia is 1.02% more likely to be male than female. Furthermore, a person diagnosed with thrombosis is 6.28% more likely to be male than female. These differences highlight the varying gender-based risks associated with these disorders.

3.4. Multiple logistic regression

Based on the hypothesis testing, Table 1 shows that females are more likely to develop anaemia than males. In the development of thrombosis, gender is found to be a more significant factor than age. In contrast, age is a more significant factor than gender in the case of leukaemia, indicating that the likelihood of diagnosis increases with age [11]. ** And *** intend the variables to be significant and highly significant, respectively.

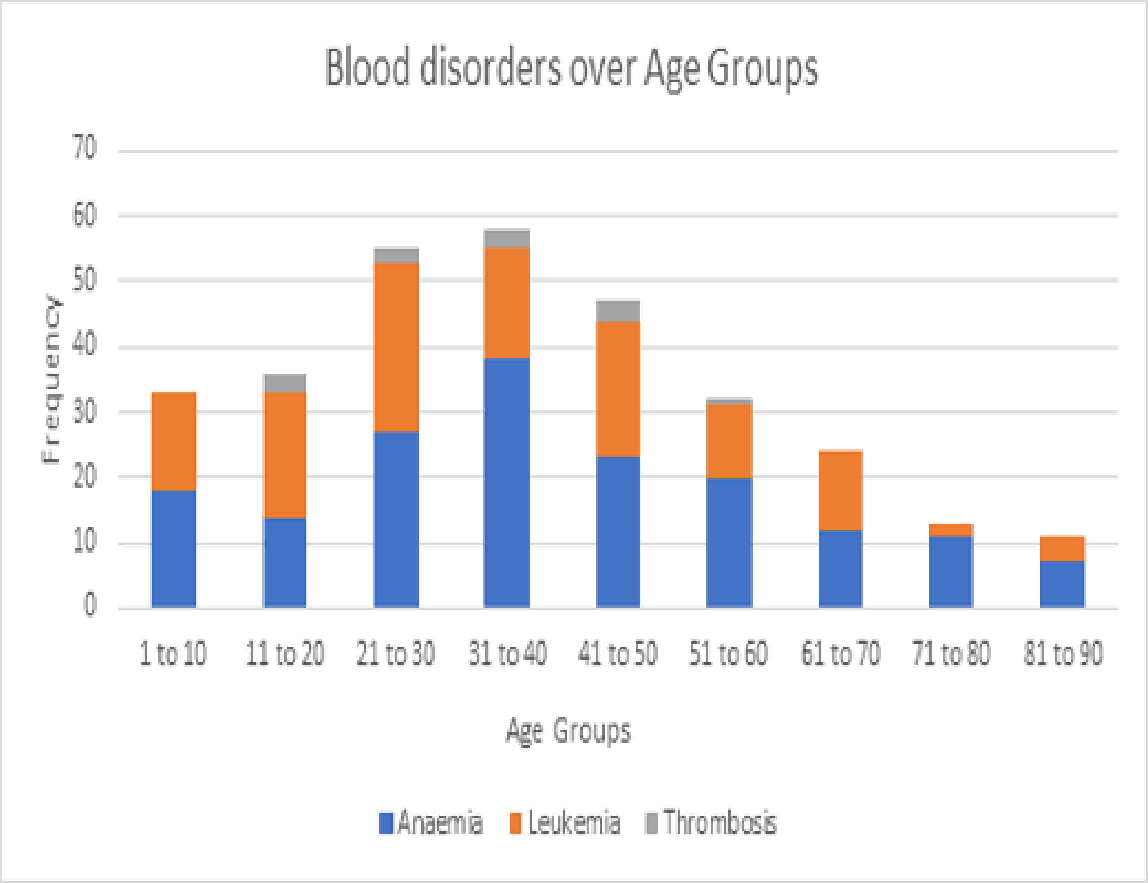


Figure 1: Occurrence of blood disorders in males and females.

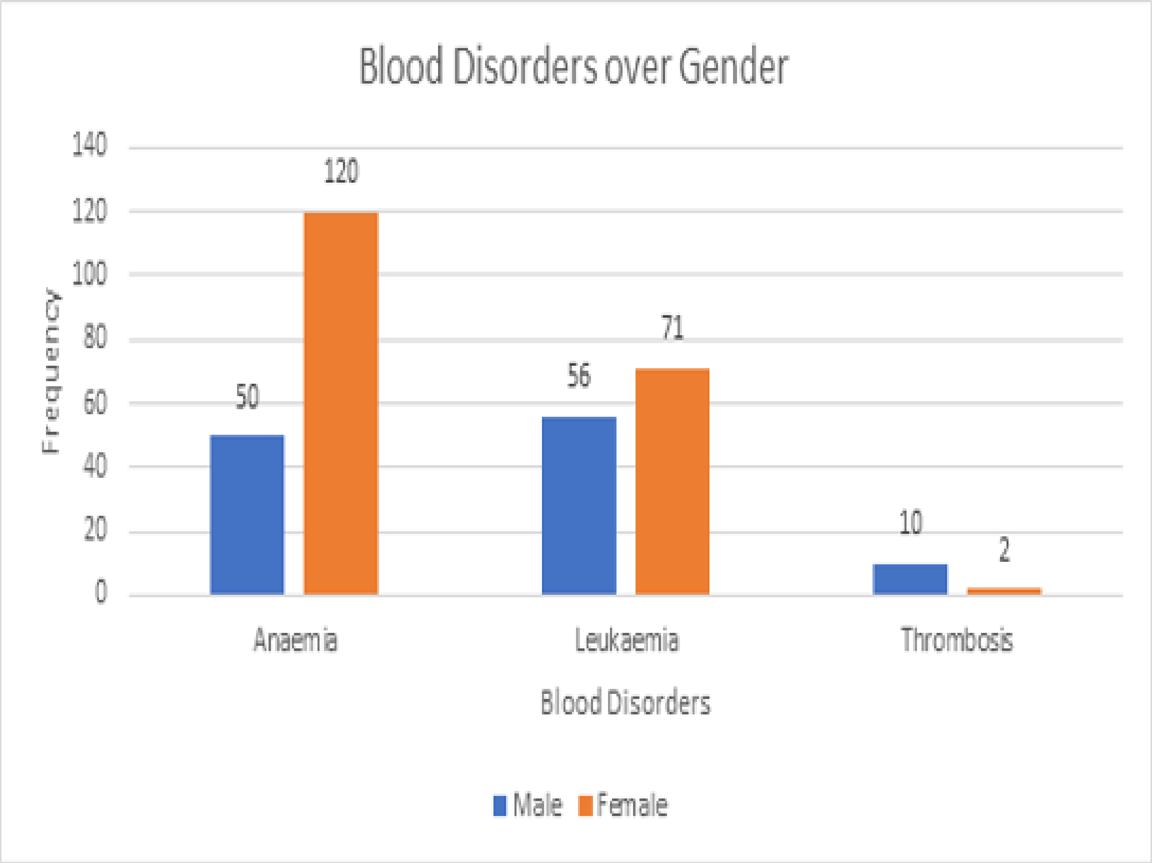


Figure 2: Prevalence of blood disorders in various age groups.

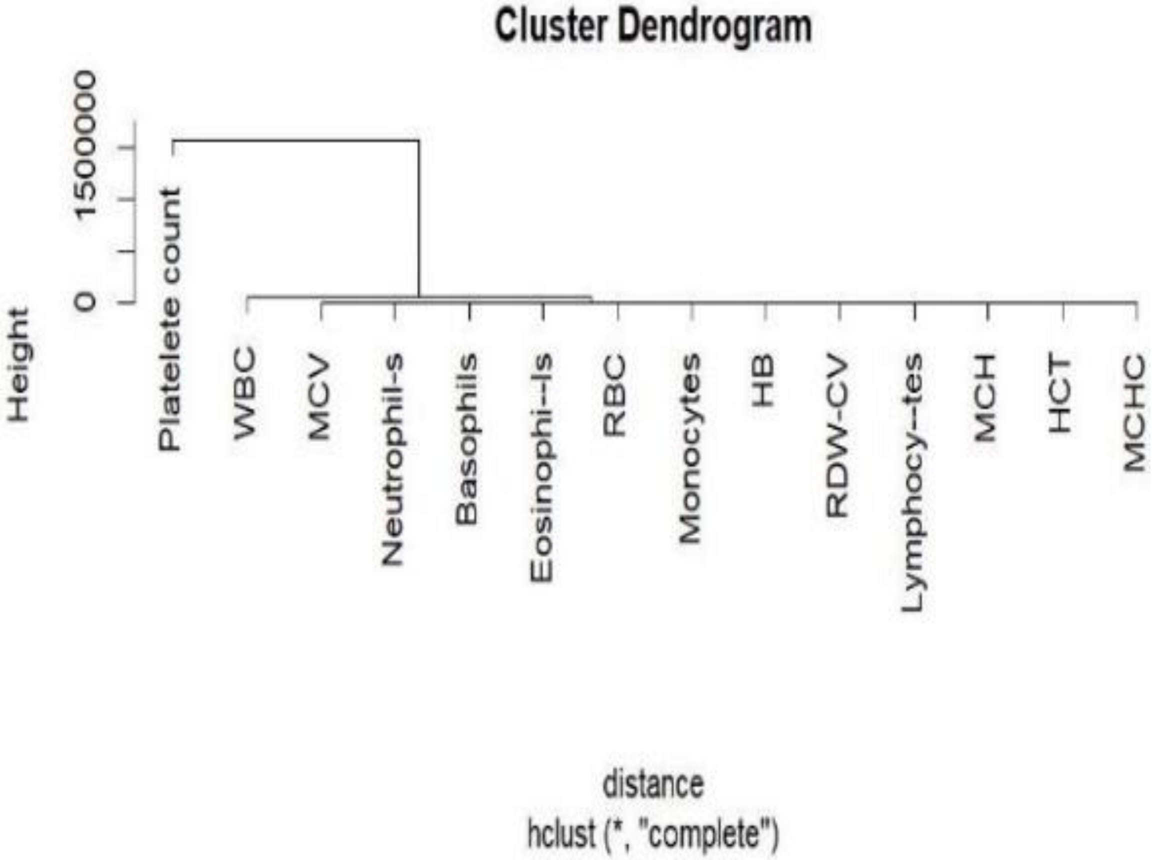


Figure 3: Dendrogram of Age 1-10

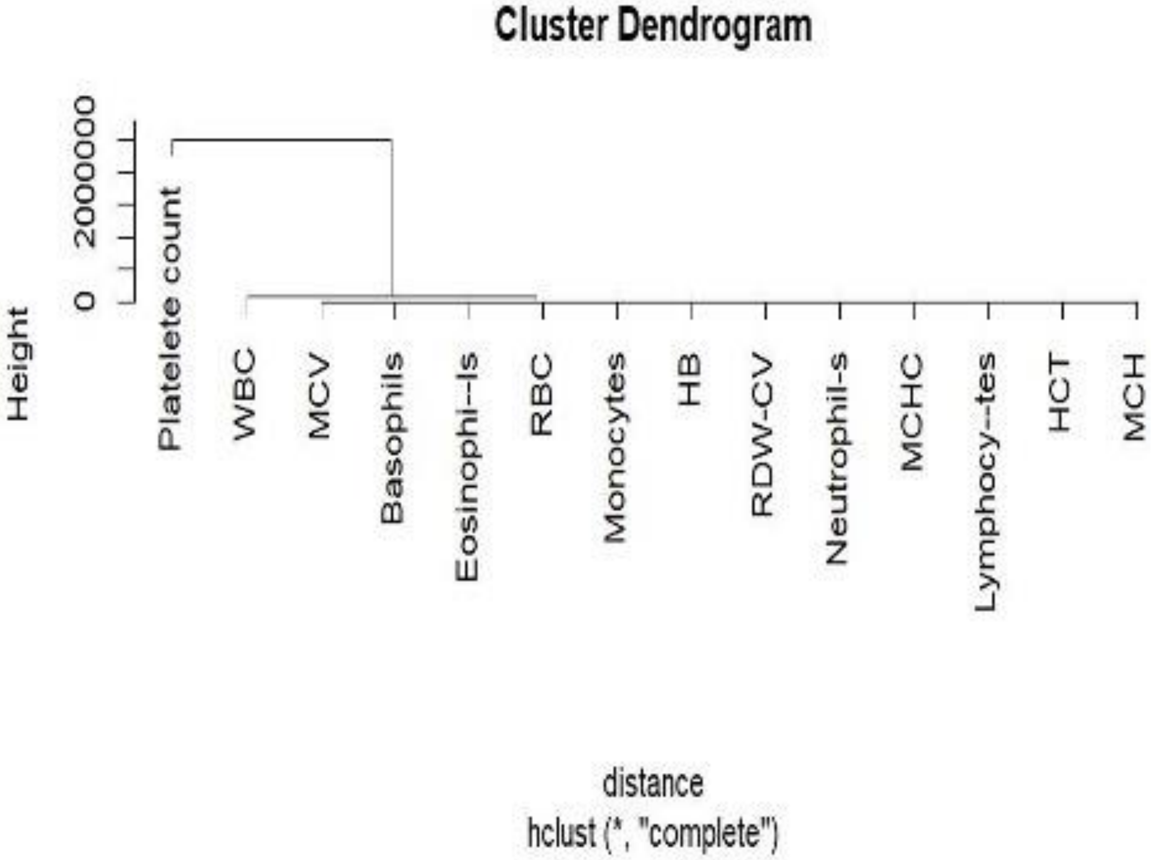


Figure 4: Dendrogram of Age 41-50

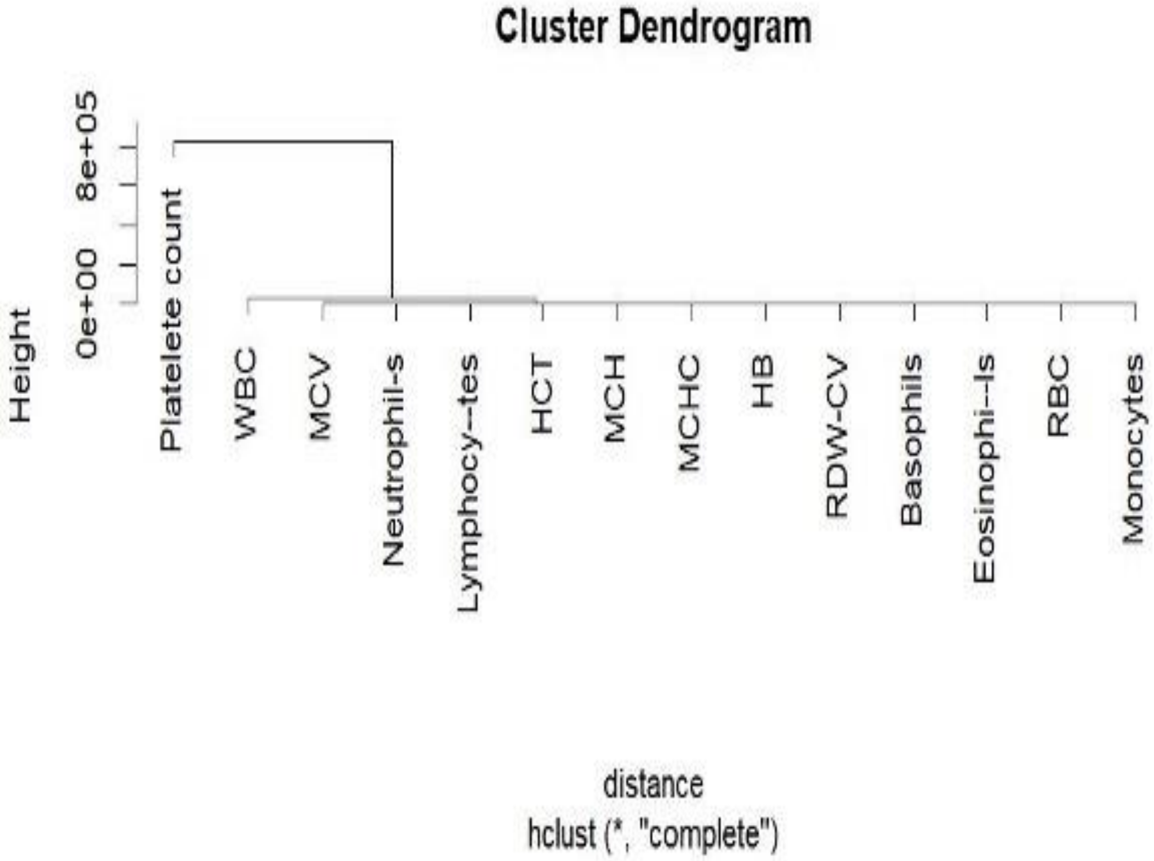


Figure 5: Dendrogram of Age 81-90

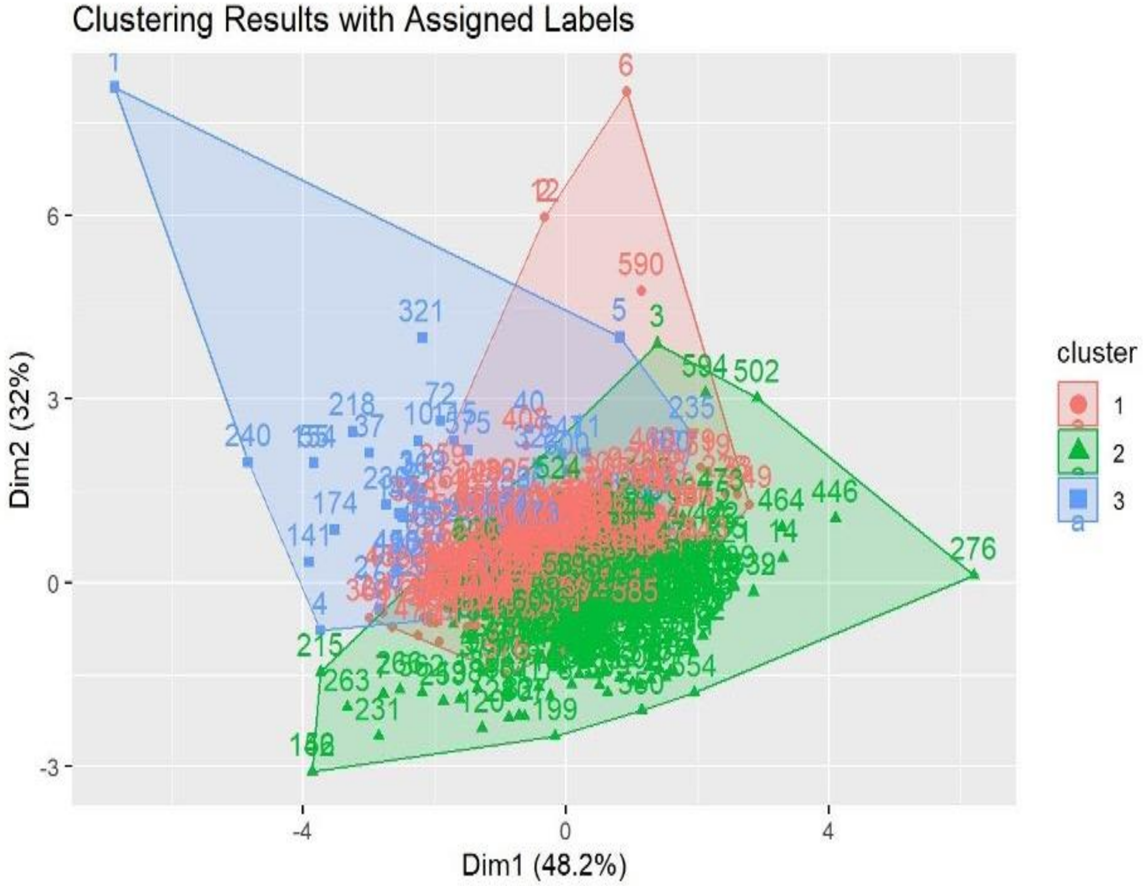


Figure 6: k- means clustering of anaemia, leukaemia, and thrombosis.

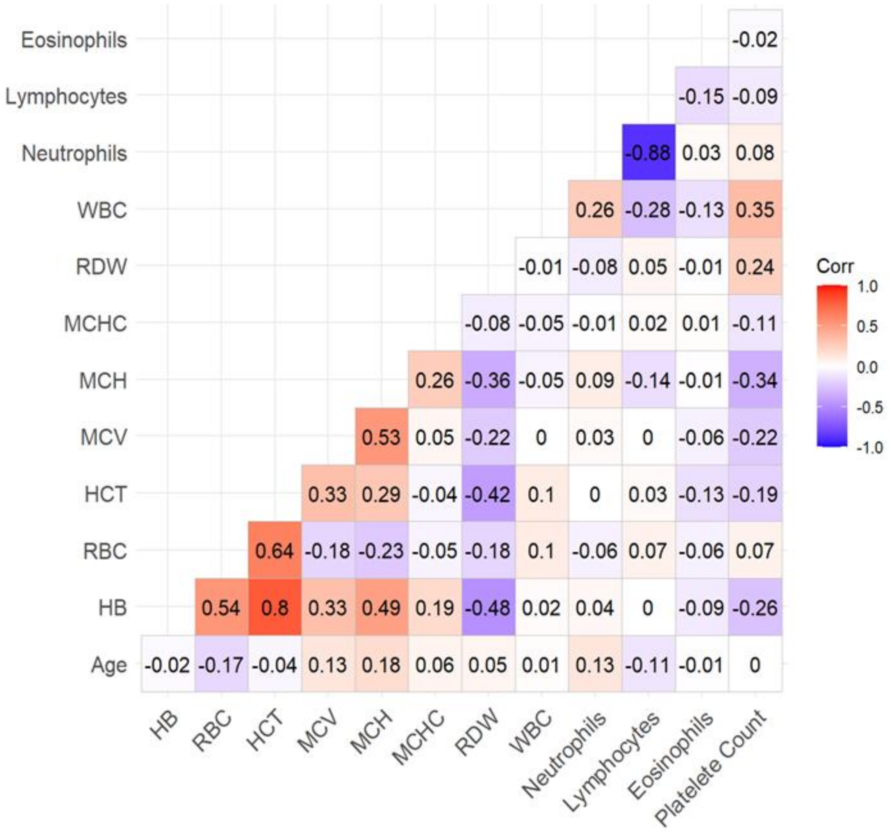


Figure 7: Heat map of all blood components

Table 1: Statistical analysis of blood disorders by logistic regression

Sr no.	Disorders	Predicted variable	Estimate	Standard error	z - value	p-value
1	Anaemia	Intercept	-0.0856627	0.2414342	-0.355	0.72273
		Age	0.0001441	0.0052706	0.027	0.97819
		Gender F	-0.6134432	0.2148962	-2.855	0.00431 **
2	Leukaemia	Intercept	-0.760201	0.225894	-3.365	0.000765 ***
		Age	-0.015674	0.005089	-3.080	0.002072 **
		Gender	0.056236	0.202775	0.277	0.781523
3	Thrombosis	Intercept	-2.93151	0.60932	-4.811	1.5e-06 ***
		Age	-0.01586	0.01538	-1.031	0.3023
		Gender	-1.34906	0.77947	-1.731	0.0835

4. Conclusions

This study particularly highlights gender differences in hematological parameters across different age groups. A strong positive correlation was observed between Hemoglobin (Hb) and Hematocrit (HCT), suggesting those blood metrics as potential indicators for anaemia. The analysis revealed that gender might significantly impact the occurrence of anaemia and thrombosis, whereas age could be a significant factor for leukaemia. Furthermore, by using multiple logistic regression, it confirmed that gender appears to be associated with the risk of developing anaemia and thrombosis, yet independent in case of leukaemia. The risk of developing anaemia was lower in males than females, while risk for leukaemia and thrombosis was higher in males. These insights emphasize the importance of considering demographic factors in diagnosing and managing blood disorders. These analyses may provide a deeper understanding of the correlation of diagnostic results for medical professionals to plan further investigations, treatment and management of disorders.

Acknowledgement

The authors are thankful for the support and guidance provided by the staff of the Department of Statistics and Department of Chemistry, Fergusson College (Autonomous), Pune.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence work reported in this paper.

References

[1] D. Lokwani. (2013). The ABC of CBC: Interpretation of complete blood count and histograms. JP Medical Ltd. 1: 1-15.

[2] A. Engert, C. Balduini, A. Brand, B. Coiffier, C. Cordonnier, H. Döhner, T.D. De Wit, S. Eichinger, W. Fibbe, T. Green. (2016). The European hematology association roadmap for European hematology research: a consensus document. haematologica. 101(2): 115-208.

[3] J. Osborne. (2010). Improving your data transformations: Applying the Box-Cox transformation. Practical Assessment, Research, and Evaluation. 15(1).

[4] P. Arabie, L. Hubert, G. De Soete. (1996). Clustering and classification. World Scientific: pp.

[5] M.M. Mukaka. (2012). A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal. 24(3): 69-71.

[6] J.G. Hollands, I. Spence. (1992). Judgments of change and proportion in graphical perception. Human factors. 34(3): 313-334.

[7] F. Rahim, A. Kazemnejad, M. Jahangiri, A.S. Malehi, K. Gohari. (2021). Diagnostic performance of classification trees and hematological functions in hematologic disorders: an application of multidimensional scaling and cluster analysis. BMC medical informatics and decision making. 21: 1-13.

[8] C. Hennig. (2015). What are the true clusters? Pattern Recognition Letters. 64: 53-62.

[9] S. Puntanen, G.P. Styan, J. Isotalo. (2011). Matrix tricks for linear statistical models: our personal top twenty. Springer. Berlin Heidelberg. 307-310.

[10] S. Wassertheil-Smoller, J. Smoller. (2004). Biostatistics and epidemiology. Springer New York.

[11] D.W. Hosmer Jr, S. Lemeshow, R.X. Sturdivant. (2013). Applied logistic regression. John Wiley & Sons.