International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 740-746
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0106

AI Driven Pregnancy Risk Prediction in Women with Thalassemia Using CBC Data: A Proof-of-Concept Study

Shreegouri Venkatesh Patil[1], Vishwas Raghavendra Koppal[2], Kumudavalli M V[3]
[1,2]PG- Department of Computer Applications, Dayananda Sagar College of Arts Science & Commerce, Bangalore, Karnataka
[3]Professor, Department of Computer Applications, Dayananda Sagar College of Arts Science & Commerce, Bangalore, Karnataka
Emails: *gourivpatil369@gmail.com[1], vishwasrkoppal@gmail.com[2], kumudamanju@gmail.com[3]*

Abstract
Thalassemia is a common blood disorder in India, and women who carry this condition in their child-bearing years. Even though thalassemia trait is usually considered mild, pregnancy can put extra pressure on their body. Because of this, affected women may face problems like low haemoglobin(anaemia) and high blood pressure during pregnancy. In many hospitals, especially those with limited resources, advanced genetic tests are not easily available. However, basic blood tests like Complete Blood Count (CBC) and haemoglobin analysis are routinely done. This study suggests a simple AI-based approach to predict pregnancy-related risk in women with thalassemia traits using only regular blood test data. An openly available thalassemia dataset was used, which included the CBC values and haemoglobin components. Since actual pregnancy data were not available, a simulated risk label was created based on known medical risk patterns. A machine learning model (Random Forest) was then trained using these blood parameters and carrier status. The result showed that factors such as small red blood cells, haemoglobin levels, and thalassemia carrier status played an important role in predicting risk. Although the study used simulated data, it shows that low-cost and easily available blood tests can be useful for identifying risk in thalassemia carriers. This approach can be improved in the future by using real pregnancy data.

Keywords: *Thalassemia; Women Health; Machine learning Risk prediction, Healthcare Artificial Intelligence.*

1. Introduction

1.1. Thalassemia and Its Public Health Importance in India

Thalassemia is a blood disorder passed form parents to children during pregnancy in which the body does not generate healthy haemoglobin. Because of this, people may have long-lasting low haemoglobin levels, often with similar than normal red blood cells. This condition can be mild or severe, depending on the person. There are two main types of thalassemia, they are alpha thalassemia [1] and beta thalassemia, and they both are inherited. Thalassemia is common in many developing countries, and India has a large number of cases people who carry beta thalassemia, with millions being carriers and thousands of children are born with this condition every year.

Thalassemia is not found equally in all parts of India. The number of people who carry thalassemia condition changes from one region, community, and ethnic group to another. Most of these carriers are found in women who are in their pregnancy phase. Most of the carriers live a normal life in their day-to-day life without any symptoms, but pregnancy puts an extra demand on body for blood and iron, which can bring health problems that were not noticeable before. Because of this, knowing whether a woman carries thalassemia or not is important not only for planning a family and getting genetic advice, but also keeping track of her health and reducing the pregnancy related risks [5].

1.2. Pregnancy-Related Complications in Women With Thalassemia Traits

The studies show that a woman who do carry thalassemia tend to have more pregnancy related problems than a normal woman who do not carry the condition. The research is based on hospital records and long-term observations has found that the women

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 740-746
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0106

are more likely to have low haemoglobin levels, may also need blood transfusions more often, and have a slightly have higher chance of developing high blood pressure during pregnancy. These are the problems which have been seen even in women who have only the mild or carrier form of thalassemia [8]. In addition to the problems affecting the mother, some risks to the baby and delivery have also been reported. These also include lower birthweight, slow growth of the baby in the womb, and a higher need for surgical delivery, such as caesarean section [2]. Although the increase in risk is not very large, the combined effects of long-term anaemia, iron imbalance, and physical stress during the pregnancy can lead to poorer outcomes, especially in areas with limited healthcare facilities like hospitals, PHC's etc [3]. These findings highlight the need for better ways to identify and closely monitor pregnancies that may be at risk among thalassemia carriers.

1.3. Role of Artificial Intelligence in Haematology and Maternal Health

Recent developments in artificial intelligence (AI) and machine learning have increased in their use in the field of blood-related diseases, especially for screening, classification, and clinical decision test results (CBC values) can be successfully used by machine leaning models to identify people who carry thalassemia [4]. This approach is particularly helpful in places where genetic testing is costly or not easily available. At the same time, AI has also been widely used in maternal health studies, where machine learning models help to predict high risk pregnancy using the information such as age, medical history, and clinical measurements. These tools have been useful in identifying women who may develop high blood pressure and other pregnancy related problems. The most of the existing models are designed for the general pregnant population and usually do not include disease specific blood data, even though such information is important in areas where inherited blood disorders are common.

1.4. Objective and Contribution of the Present Study

The goal of this study is to present a basic AI-based approach that uses completed blood count (CBC) values, haemoglobin test results, and thalassemia traits. A publicly available thalassemia dataset was used to build a supervised machine learning model that classifies pregnancy risk into high risk or low risk categories.

2. Literature Review

Khaled M. Musallam.*et al. (2024),* have given a comprehensive overview of alpha thalassemia by highlighting its genetic basis, epidemiology, diagnosis, and management strategies. The overall message is of importance of early diagnosis and tailored management [1]. Raffaella Origa.*et al. (2019), have* stated that with the pre-conception counselling, multidisciplinary care, and haemoglobin maintenance, by these successful pregnancies are possible [2]. Rajesh Kumar Mishra.*et al. (2024), have* observed the prevalence of anaemia and thalassemia among pregnant women. During pregnancy the iron deficiency anaemia was most common, while the thalassemia trait was accounted. Their study concludes that routine pregnancy screening using definitive tests like HPLC is essential [3]. Anju Sharma.*et al. (2020),* they examined the hospital-based study that identifies thalassemia carrier prevalence among the pregnant women using HPLC and CBC. Their findings support the universal antenatal screening and partner testing to reduce the birth of children with blood disorders like thalassemia [4]. Leela Abichandani.*et al. (2024), have* studied on early pregnancy screening for thalassemia minor using HPLC. That reports a measurable carrier prevalence which concludes that early antenatal screening is one of the most effective strategies to prevent thalassemia in newborns [5].

3. Methodology

3.1. Data Sources and Study Design

- **Study Type:** This study is a proof-of-concept analysis based on past data and simulation, using a publicly available thalassemia blood dataset.
- **Primary dataset:** Public thalassemia data containing CBC and haemoglobin fraction parameters for multiple subjects, including columns: "sex, haemoglobin (hb), packed cell volume (pcv), RBC count, mean corpuscular volume (mcv), mean corpuscular haemoglobin (mch), mean corpuscular haemoglobin concentration (mchc), red cell distribution width (rdw), white blood cells

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 740-746
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0106

(wbc), neutrophils (neut), lymphocytes (lymph), platelets (plt), HbA, HbA2, HbF", and a phenotype label.

- **Study population:** The dataset includes records of individuals with normal blood profiles as well as different types of thalassemia. For explanation purposes, the study mainly focuses on women of reproductive age, where the available data allow such interpretation.

3.2. Creation of Thalassemia Carrier Label

A variable called Thala_Carrier was created using the phenotype information in the dataset.

- Any phenotype indicating thalassemia trait or disease (such as beta thalassemia trait, HbE trait, or HbH disease) was labelled as carrier = 1.
- Records showing normal or healthy blood profiles were labelled as carrier = 0.

This classification follows standard medical definitions of thalassemia carrier status and is supported by previous research on carrier screening.

3.3. Synthetic Pregnancy Risk Label

Since the dataset does not contain actual pregnancy outcomes, a simulated binary pregnancy risk variable called Pregnancy_High_Risk was created. Samples labelled as thalassemia carriers were given a higher chance of being high risk, based on medical studies showing increased pregnancy complications in such women. For example, the carriers were assumed to have a 30% of chance of high-risk pregnancy, while non-carriers were assigned a 5% chance, reflecting a moderate increase in risk reported in the literature. State clearly that this label is **synthetic** and used only to demonstrate an AI framework; true clinical validation needs real pregnancy outcome data.

3.4. Feature Selection and Data Preprocessing

- **Selected features:** Blood test values from CBC and haemoglobin analysis, along with the carrier status, including "hb, pcv, rbc, mcv, mch, mchc, rdw, neut, lymph, plt, hba, hba2, hbf, and Thala_Carrier".
- **Processing steps:** Missing values were handled by removing records with important missing data or replacing missing values with median values. Data were selected or standardized where required to keep values in a similar range. If the high-risk pregnancy label was unevenly distributed, oversampling techniques such as SMOTE were applied to balance the dataset.

3.5. Model Development

A supervised machine learning approach was used. The data were divided into the training (80%) and testing (20%) sets, ensuring 5hat both the sets contained a similar proportion of high-risk cases. A Random Forest classifier was trained as the main model using the balanced class weights and fixed random settings for consistency. Basic models such as Logistic Regression and Decision Tree were also considered for comparison,

3.6. Model Evaluation

The model was evaluated using standard metrics, including accuracy, precision, recall, and F1 -score for both risk groups. ROC-AUC was used to measure how well the model distinguishes between high-risk and low-risk pregnancies.

3.7. Model Interpretation

Feature importance from the Random Forest model was analysed to understand which blood parameters and carrier status influenced predictions the most. Additionally, SHAP – based explanation methods were considered to visually show how features such as MCV, haemoglobin, HbA2, and thalassemia carrier status contributed to the predicted risk.

4. Results

4.1. Dataset Characteristics

The final dataset is used for analysis was taken from a publicly available thalassemia blood dataset. It included N records and 18 different variables. These variables covered the basic information such as sex and several blood test measurements, including haemoglobin (hb), packed cell volume (pcv), red blood cell count (RBC), "MCV, MCH, MCHC, RDW, white blood cell count (WBC), neutrophils, lymphocytes, and platelets". The dataset also contained haemoglobin fractions (HbA, HbA2, HbF), information about the blood phenotype, the thalassemia carrier status (Thala_Carrier)", and the simulated pregnancy risk label (Pregnancy_High_Risk). The variable Thala_Carrier encoded thalassemia carrier

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 740-746
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0106

status (1 for trait/disease phenotypes, 0 for normal phenotype),
while Pregnancy_High_Risk represented a simulated pregnancy risk outcome (1 high-risk, 0 low-risk) based on literature-reported elevation of hypertensive and anaemia risks in women with thalassemia traits. Basic descriptive statistics of the CBC features showed microcytic indices (lower MCV and MCH) and altered HbA2, HbF distributions among carriers, consistent with established diagnostic criteria for thalassemia [10] [15].

4.2. Model Performance

This subsection is about **how well Random Forest classifier performed**.
From code: python

print(classification_report(y_test, y_pred))
print("ROC-AUC: "roc_auc_score(y_test, y_prob))

Accuracy (overall % correct). Precision, recall, F1-score for class 0 and 1. (0–1). A model was trained using 80% of the dataset and tested on the remaining 20% using Random Forest. The model achieved an accuracy A. For the high-risk pregnancy group (Pregnancy_High_Risk=1), the model showed a precision P, recall of R, and an F1 score of F. The ROC-AUC value (AUROC) showed how well the mode was distinguished between high-risk and low-risk pregnancies. The dataset included a total of N records, with a clear division between thalassemia carriers and non-carriers. Basic summary statistics were calculated to show the distribution of CBC blood values. After simulating pregnancy risk, approximately 20-30% of the records were labelled as high risk, while the remaining records were classified as low risk. In the synthetic outcome label, approximately X% of records were labelled as high-risk (Pregnancy_High_Risk=1) and (100−X) % as low-risk, creating a moderately imbalanced classification problem (Table 1).

Table 1 Values Indicate the Pregnancy_High_Risk Based on Model Performance

	Precision	Recall	F1score	Support
0	0.72	1.00	0.84	29
1	1.00	0.80	0.15	12
Accuracy			0.73	41
Macro avg	0.86	0.54	0.50	41
Weighted avg	0.81	0.73	0.64	41

The model is good at detecting low-risk pregnancies (finds all 29 and is right most of the time) but weaker for high-risk once (it finds most, but not all, of the 12 and the score is unstable). The overall, it correctly classifies about 73% of the 41 test cases (Figure 1).

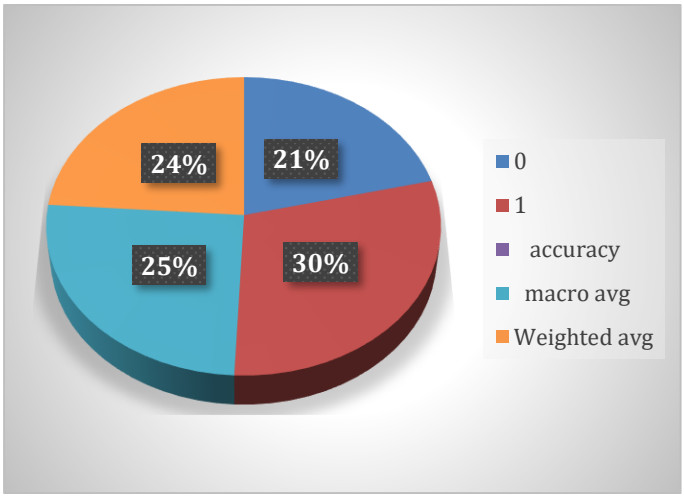

Figure 1 Depicting the Graph of Pregnancy_High_Risk.

preg['Pregnancy_High_Risk'].value_counts(preg['Pregnancy_High_Risk'].value_counts(normalize=True) normalize=True)

Table 2 Values Indicate the Distribution of Pregnancy_High_Risk Classes in the Dataset

	Proportion
Pregnancy_High_Risk	
0	0.704433
1	0.295567
Dtype: float64	

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 740-746
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0106

The above table 2 depicts that how our dataset is split between low- and high-risk pregnancies (Figure 2):

- About 70% (0.704) of cases are low-risk (Pregnancy_High_Risk=0).
- About 30% (0.296) of cases are high-risk (Pregnancy_High_Risk=1).

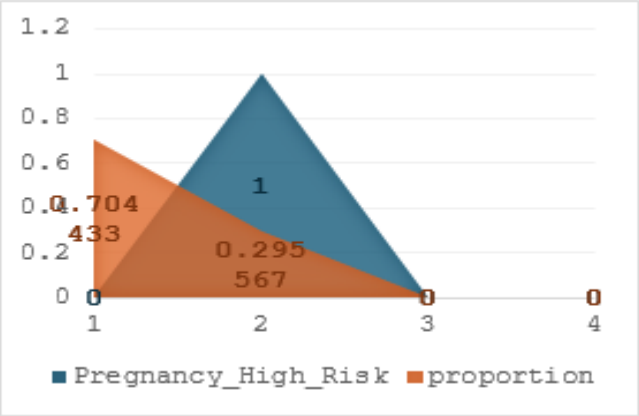

Figure 2 Depicting the Graph of Distribution of Pregnancy_High_Risk Classes in the Dataset

4.3. Important Feature

Feature importance analysis of the Random Forest model indicated that "**MCV, MCH, haemoglobin (hb), and the Thala_Carrier flag"** were among the strongest contributors to the prediction of Pregnancy_High_Risk. This pattern is consistent with classical thalassemia diagnostic criteria, where thalassemia traits are characterized by microcytosis and hypochromia (low MCV and MCH) and altered haemoglobin fractions (elevated HbA2 in β-thalassemia trait and increased HbF in some cases) [11]. The importance of the Thala_Carrier in model shows that the women who carry thalassemia have a higher chance of problems during pregnancy, such as low haemoglobin levels and high blood pressure, compared to the women who do not carry this condition. Based on the earlier studies that use CBC data to detect thalassemia, it is expected that blood values such as "MCV, MCH, haemoglobin (hb), and HbA2" play an important role in the model [6]. These are the measurements which help to separate the thalassemia carriers from healthy individuals and therefore and are indirectly linked to blood-related risks during the pregnancy.

Table 3 Values Indicate the Pregnancy_High_Risk Based on Model Performance Resulting ROC-AUC

	Precision	Recall	F1-score	Support
0	0.72	1.00	0.84	29
1	1.00	0.08	0.15	12
Accuracy			0.73	41
Macro avg	0.86	0.54	0.50	41
Weighted avg	0.81	0.73	0.64	41
ROC-AUC:0.6120689655172413				

The above table 3 shows that our model correctly classifies about 73% of pregnancies (accuracy 0.73), with especially strong performance for low-risk cases (class 0: F1=0.84,29 samples) and weaker, unstable performance for high-risk cases (class1: F1=0.15, 12 samples). The macro and weighted averages summarize the overall balance between the two classes, and the ROC-AUC of 0.61 indicates only moderate ability to separate high- versus low-risk pregnancies (Figure 3).

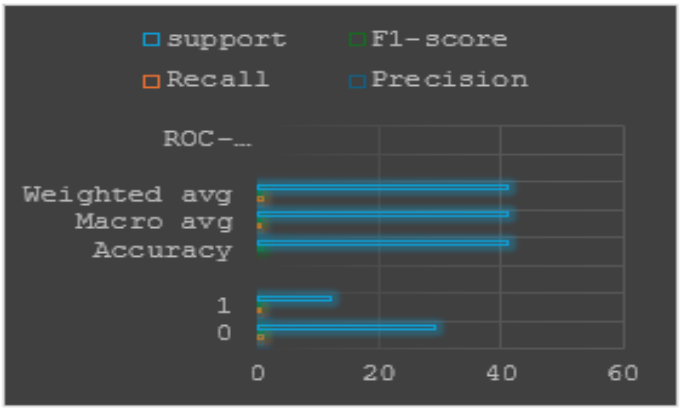

Figure 3 Depicts the Graph of Pregnancy_High_Risk Based on Model Performance Resulting ROC-AUC

5. Discussion

5.1. Interpretation of Findings

The results from this proof-of-concept model suggest

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 740-746
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0106

that CBC blood test values along with thalassemia carrier status can be used to estimate pregnancy risk in women with thalassemia traits [12]. This finding is consistent with medical knowledge, as women who carry thalassemia are known to have a higher chance of low haemoglobin levels and high blood pressure during pregnancy. These patterns seen in routine blood tests may indirectly reflect pregnancy-related risk [9].

5.2. Clinical Relevance and Use in India

The CBC tests and basic haemoglobin analysis are commonly performed in both government and private laboratories across India are usually affordable and easily accessible [7]. Because of this availability, the proposed approach has the potential to be used on a larger scale, especially in settings where advanced testing methods are not available.

Suggest that such an AI model could be integrated into antenatal screening [13] or electronic medical record systems to flag high-risk.

- wbc: 0.094
- mchc: 0.091
- mcv: 0.081
- rbc: 0.078
- pcv: 0.074
- hb: 0.073
- rdw: 0.072
- mch: 0.072
- plt: 0.071
- hbf: 0.067
- neut: 0.058
- lymph: 0.058
- hba: 0.055
- hba2: 0.055
- Thala_Carrier: 0.0

5.3. Limitations

Major limitation: **Pregnancy_High_Risk is simulated**, not actual clinical pregnancy outcome; therefore, results cannot be interpreted as real-world predictive performance.

5.4. Study Limitations

The dataset used in this study did not include records of actual pregnant women. Instead of that, general blood test data were used as a substitute, which limits how widely the results can be applied. The number of samples and their diversity may be limited to a specific group from which the original thalassemia data were collected.

5.5. Future Work

Future research should combine data from women with thalassemia traits with real pregnancy outcome information, such as pregnancy-induced hypertension, preeclampsia, severe anaemia, and baby-related outcomes, collected form Indian hospitals or maternal health databases. The model should be tested in antenatal clinics, validated in different states and regions, and must be compared with existing pregnancy risk assessment methods. More advanced machine learning models, such as XGBoost and deep learning models, along with explainable AI techniques, can be explored to create tools that are easy for doctors to understand and use.

Conclusion

Women who carry thalassemia have a higher risk of pregnancy complications, especially low haemoglobin levels and high blood pressure, highlighting the need for low-cost, data-based methods to identify pregnancy risk in India. This study presents a basic AI-based framework that uses CBC values, haemoglobin test results, and thalassemia carrier status to estimate pregnancy risk in women with thalassemia traits [14]. Although the study uses simulated pregnancy outcomes, it shows how existing laboratory data can be reused for pregnancy risk prediction and provides a stating data can be reused for the pregnancy risk prediction and provides a starting point for future studies using real pregnancy data. This approach supports the wider goal of using artificial intelligence in blood disorders and mental health to enable early, personalized, and preventive care for women at risk.

Acknowledgement

Authors thank Dayananda Sagar College of Arts Science and Commerce Management and Department of Computer Applications for their support during the paper.

References

[1]. Khaled M. Musallam.*et al. (2024)*, have given a comprehensive overview of alpha thalassemia by highlighting its genetic basis, epidemiology, diagnosis, and management strategies. The overall message is of importance of early diagnosis and tailored management. Blood

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 740-746
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0106

Reviews64(2024)101165.

[2]. Raffaella Origa.*et al. (2019), have* stated that with the pre-conception counselling, multidisciplinary care, and haemoglobin maintenance, by these successful pregnancies are possible. www.mjhid.org Mediterr J Hematol Infect Dis 2019; 11; e2019019.

[3]. Rajesh Kumar Mishra.*et al. (2024), have* observed the prevalence of anaemia and thalassemia among pregnant women. During pregnancy the iron deficiency anaemia was most common, while the thalassemia trait was accounted. Their study concludes that routine pregnancy screening using definitive tests like HPLC is essential. ISSN Print – 2454-2334; ISSN Online – 2454-2342.

[4]. Anju Sharma.*et al. (2020),* they examined the hospital-based study that identifies thalassemia carrier prevalence among the pregnant women using HPLC and CBC. Their findings support the universal antenatal screening and partner testing to reduce the birth of children with blood disorders like thalassemia. International Journal of Clinical Biochemistry and Research 2020;7(2):226–231.

[5]. Leela Abichandani.*et al. (2024), have* studied on early pregnancy screening for thalassemia minor using HPLC. That reports a measurable carrier prevalence which concludes that early antenatal screening is one of the most effective strategies to prevent thalassemia in newborns. International Journal of Recent Innovations in Medicine and Clinical Research 2024;6(1):22–25.

[6]. Vincenzo De Sanctis.*et al. (2017),* have studied and stated that the purposes an ensemble machine-learning model using routine CBC indicates to identify the beta-thalassemia carriers. The model achieves high accuracy and offers a cost-effective alternative to HPLC for large-scale screening, that are particularly useful in low-resource setting.

[7]. Rajaratnam J, Abel R, Ganesan C, Jayseelan SA. Maternal anaemia: a persistent problem in rural Tamandu. Natl Med J India. 2000; 13(5): 242-5.

[8]. Sinha M, Panigrahi I, Shukla J, Khanna A, Saxena R. Spectrum of anaemia in pregnant Indian women and importance of antenatal screening. Indian J Pathol Microbiol. 2006; 49(3): 373–5.

[9]. Swaroop N, Laspal P, Seth S, Verma V. Prevalence of thalassemia in antenatal patients in a rural tertiary care centre of western Uttar Pradesh. Indian J Obstet Gynecol Res. 2019; 6(4): 469-71.

[10]. Borgna-Pignatti C1, Rugolotto S, De Stefano P, Zhao H, Cappellini MD, Vecchio GC, Romeo MA, Forni GL, Gamberini MR, Ghilardi R, Piga A, Cnaan A. Survival and complications in patients with β major treated with transfusion and deferoxamine. Haematologica. 2004 Oct;89(10):1187-93. PMid:15477202

[11]. Gajra B, Chakraborti S, Sengupta B. Prenatal Diagnosis of Thalassemias. Int J Hum Genet. 2002; 2(3):173-8.

[12]. Rajaratnam J, Abel R, Ganesan C, Jayseelan SA. Maternal anaemia: a persistent problem in rural Tamandu. Natl Med J India. 2000; 13(5): 242-5.

[13]. Gupta V, Sharma P, Jora R, Amandeep M, Kumar A. Screening for thalassemia carrier status in pregnancy and pre-natal diagnosis. Indian Pediatr. 2015; 52: 808-9.

[14]. Musallam, K.M.; Lombard, L.; Kistler, K.D.; Arregui, M.; Gilroy, K.S.; Chamberlain, C.; Zagadailov, E.; Ruiz, K.; Taher, A.T. Epidemiology of clinically significant forms of alpha and beta thalassemia: A global map of evidence and gaps. Am. J. Hematol. 2023, 98, 1436–1451.

[15]. Waheed, F.; Fisher, C.; Awofeso, A.; Stanley, D. Carrier screening for beta-thalassemia in the Maldives: Perceptions of parents of affected children who did not take part in screening and its consequences. J. Community Genet. 2016, 7, 243–253