Iranian Journal of Blood & Cancer

Journal Home Page: www.ijbc.ir

ORIGINAL ARTICLE

A New Approach Based on Erythrocyte Indices to Exclude Normal Populations from Chromatography Based Thalassemia Screening Programs with Very High Fidelity

Abhishek Samanta[1], Paban Kumar Chaudhuri[2], Ushnish Das[3], Nandan Bhattacharyya[4*]

[1]Department of Zoology, Panskura Banamali College, P.O. – Panskura R.S., West Bengal PIN 721152, India
[2]Thalassaemia Control Unit, School of Tropical Medicine, Kolkata, India
[3]Statistical Programming and Analysis Division, Tata Consultancy Services, Puna, India
[4]Department of Biotechnology, Panskura Banamali College, P.O. – Panskura R.S., West Bengal PIN 721152, India

ARTICLE INFO

Article History:
Received: 20.09.2021
Accepted: 12.11.2021

Keywords:
Erythrocyte indices
Thalassemia
Screening

*Corresponding authors:
Nandan Bhattacharyya,
Department of Biotechnology,
Panskura Banamali College, P.O.
– Panskura R.S., West Bengal PIN
721152, India
Tel: +91 9434453188
Email: bhattacharyya_nandan@rediffmail.com

ABSTRACT

Background: Screening and counselling is the most effective way to prevent the birth of children with thalassemia major. An accurate and relatively less time-consuming protocol is necessary to screen large populations. Separating iron deficiency anaemia from thalassemia trait based on blood cell parameters has been used by hematologists for many years. We aimed to design a new approach to screen the microcytic populations.
Methods: Blood cell parameters and chromatography were used to screen the populations traditionally. Validating the result with a five-point decision tree analysis with two equations based on cut-off values of five blood cell parameters was performed. 2984 participants were screened traditionally, of which 289 were found to be beta-thalassemia trait, 63 were Hemoglobin E carriers, 15 were found to be Hemoglobin D (Hb D) Punjab, 4 hereditary persistent fetal hemoglobin (HPFH), and 14 belonged to beta thalassemia traits with HbA2 levels between 3.3% to 3.8% associated with reduced mean corpuscular volume (MCV) and mean corpuscular haemoglobin (MCH) (borderline cases).
Results: In the decision tree approach, 51.3% with beta thalassemia trait and 11.65% Hb E carriers were detected perfectly. 27% of participants were detected as non-thalassemia carriers which could be excluded from further chromatographic analysis.
Conclusion: During the early stages of the carriers screening program, a large portion of the sample could be excluded, based on segregating the IDA and thalassemia carrier population. Decision tree analysis and equation derived from the regression are essential to from limit of exclusion which implies significant cost reductions.

Please cite this article as: Samanta A, Chaudhuri PK, Das U, Bhattacharyya N. A New Approach Based on Erythrocyte Indices to Exclude Normal Populations from Chromatography Based Thalassemia Screening Programs with Very High Fidelity. IJBC 2021; 13(4): 107-118.

Introduction

Hemoglobinopathies are a group of inherited disorders associated with production of abnormal hemoglobins. According to the World Health Organization[1] and the Thalassemia International Federation,[2] more than 330,000 infants affected with various haemoglobin disorders are born annually. Most of them are originated from Southeast Asia, India, Mediterranean and Middle Eastern ethnic populations. The incidence of thalassemia mutations is different in different parts of the world reported about 13% in Africa, 4% in Asia, and 2% in the United States.[3, 4]

In India, the prevalence of hemoglobinopathies is estimated to be 1.2 per 1000 births.[5] With roughly 27 million births per year,[6] this implies approximately 32,400 children are born annually with a severe hemoglobin abnormality.

According to Sinha *et al.*,[7] the blood requirement for the treatment of thalassemia major will increase to 9.24 million units in India by 2026. The treatment of thalassemia is expensive, so prevention of thalassemia is the best way to control this disease. Birth of children with thalassemia major can be prevented by restricting childbirth in couples with thalassemia minor.[8] Thus, carrier detection and effective counselling could reduce the birth rate of children with thalassemia major.[9] In India, still a high percentage of high-risk couples get married. As half of the Indian population belong to the age group of below 25 years.[1] Premarital screening and counselling of this population would be effective for controlling the major consequences.[10]

Distinctive screening projects have been initiated throughout India with different thalassemia carrier profiles, but there is no adequate data setto be able to validate a localized preventive measure.[11]

Since, the clinical manifestation and laboratory indexes of thalassemia trait and iron deficiency anemia (IDA) may be similar, the discrimination between these two entities is important. The traditional approach to screen a thalassemia carrier is to analyse the hematological parameters and chromatographic pattern of the blood sample.[12] There are numerous reports suggesting a second approach to screen normal subjects and thalassemia carriers based on the hematological parameters which are inexpensive and less time consuming.[13, 14] Many studies have used RBC indices to discriminate thalassemia carriers from normal population. These studies suggested mathematical models which can exclude 3.8-29.32% of populations with 71.43-92.08% of fidelity.[13, 14] Since these models have a minor chance to attribute a thalassemia carrier as normal, there may be major drawbacks. Although there are very few chances, this might adversely results in birth of newborns with thalassemia major. Therefore, a reliable alternative model is necessary to be able to precisely and reliably exclude normal samples by chromatographical studies.

The objective of this study was to screen thalassemia carriers using RBC parameters and chromatographic techniques that create an approach to differentiate normal population from thalassemia carriers. In this model, we consider the fidelity of the approach over its efficiency.

Materials and Methods

This study; approved by the Institutional Ethical Approval Committee, Panskura Banamali College, West Bengal, India, was carried out over three consecutive years (2017-2020). Written consent was obtained from participants for evaluation of thalassemia and other hemoglobinopathies status as per the institutional norms. A unique identification number was assigned to each participant, which was used throughout the study.

Sample Collection and Preparation

Five millilitres of blood sample was collected in a tube containing ethylene diamine tetraacetic acid (EDTA) and stored at 2-8 °C. The complete blood count was done by an automated cell counter (Sysmex XT-2000i).

High-performance Liquid Chromatography (HPLC)

The samples were screened by cationic exchange high-pressure liquid chromatography method using the VARIANT II instrument, Bio-Rad Laboratories (Hercules, CA, USA). Chromatograms were generated based on the ratio between the hemoglobin fragments and retention time (RT).

Score Construction and Equation Formation

The statistical analyses were done in this study using the IBM® *SPSS*® version 26 statistical software package. The total population was subdivided into the non-microcytic (MCV>78 fl) and microcytic groups (MCV<78 fl), respectively.[15, 16] A cut-off value was set using Hb, RBC, HCT, MCV, and MCH by the decision tree analysis method.[14] The beta-thalassemia minor population was separated from the microcytic group using the cut-off values obtained from decision tree analysis. Similarly, another cut-off value was set as the above-described method for separating the normal population from the microcytic population (Tables 1 and 2).

Table 1: Derived limits using decision tree method for beta thalassemia minor population

	5.0435	11.0831	21.8827	40.0997	78.7204	Percent Correct
5.0435	1	0	0	0	0	100.0%
11.0831	1	0	0	0	0	0.0%
21.8827	1	0	0	0	0	0.0%
40.0997	1	0	0	0	0	0.0%
78.7204	1	0	0	0	0	0.0%
Overall Percentage	100.0%	0.0%	0.0%	0.0%	0.0%	20.0%

Growing Method: Exhaustive CHAID; Dependent Variable: beta thalassemia minor

Table 2: Derived limits using decision tree method for normal population

	4.0921	11.8030	26.8886	39.6264	78.5701	Percent Correct
4.0921	1	0	0	0	0	100.0%
11.8030	1	0	0	0	0	0.0%
26.8886	1	0	0	0	0	0.0%
39.6264	1	0	0	0	0	0.0%
78.5701	1	0	0	0	0	0.0%
Overall Percentage	100.0%	0.0%	0.0%	0.0%	0.0%	20.0%

Growing Method: Exhaustive CHAID; Dependent Variable: Normal

To develop a prediction algorithm for target variables among the above-mentioned parameters and establishing classification systems, the decision tree analysis method was used.[17] After selection of the most likelihood parameter, an equation was formed by using multiple linear regression to derive a cut-off value.[18]

For further validation, a commonly used indexing protocol was determined. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), efficiency, and Youden's index (YI) were determined and compared with the different previous models for efficiency. The parameters were calculated as follows.

Sensitivity=[True Positive/(True Positive+False Negative]×100

Specificity=[True Negative/(True Negative+False Positive)]×100

Positive Predictive Value=[True Positive/(True Positive+False Positive)]×100

Negative Predictive Value=[True Negative/(True Negative+False Negative)]×100

Youden's Index=(Sensitivity+Specificity)-100.[14]

Results

During the mentioned period, 2984 participants were screened using chromatographic method. 385 were found to be carriers for thalassemia mutations (12.89% of the sample), which was higherthan the carrier frequency of West Bengal.[19] Among the 385 carriers, 289 were found to be beta-thalassemia trait, 63 were heterozygote for haemoglobin E, 15 heterozygotes for haemoglobin D Punjab, 4 for hereditary persistent fetal hemoglobin (HPFH), and 14 belonged to beta thalassemia traits with HbA2 levels between 3.3% to 3.8%.[20]

Using the above dataset, an alternative screening approach was developed to separate the true negative normal population from the rest without any false-negative result. The dataset was first divided into two sub-populations based on the HPLC results. The descriptive statistics like mean, standard deviation, Skewness, and Kurtosis and range of beta thalassemia minor (Table S1) and normal population (Table S3) were calculated.

Decision tree analysis was drawn with the mean of two above populations by chi-square values under 95% confidence level shown in Table 1 and Table 2. Exhaustive Chi-square automatic interaction detection was used as growing method in the decision tree analysis.

According to the results of the decision tree analysis, the following cut-off points of the variables (Hb, RBC, HCT, MCV, and MCH) were derived for separating the beta thalassemia minor population: Hb≤11.083 g/dL, RBC≥5.0435/mm^3, HCT≤40%, MCV≤78.7 fL, and MCH≤21.8827 pg/cell.

Similarly, for separating the normal population from the microcytic population, the cut-off points were: H ≥11.8 g/dL, RBC≥4.09 mm^3, HCT≥39.6%, MCV≥78.5 fL, and MCH≥26.8 pg/cell.

The microcytic and non-microcytic populations were segregated by a cut-off value of MCV (78 fl).[15, 16] Multiple linear regression methods were used for separating the normal population from beta thalassemia minor and other microcytic groups accompanied with Durbin-Watson auto correction and 95% confidence level and t score (Table 3).

The equation for separating the true normal population from microcytic subpopulation was derived according to the results from Table 3.

Cut-off limit for each red blood cell parameter was derived using decision tree analysis on the training population (Table 1 for thalassemia minor and Table 2 for normal population).

The cut off value equation: 3.667-0.01Hb+0.001HCT-0.004MCV+0.064MCH (Eq.1) was drawn by parameters in multiple linear regression (Table 3).

After using the limits derived by decision tree analysis for normal population in the above equation, a threshold

Table 3: T-score and 95% confidence interval for equation formation using multiple linear regression

Model		Unstandardized Coefficients		Standardized Coefficients	t-score	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	(Beta)			Lower Bound	Upper Bound
1	(Constant)	3.666	1.863		1.968	0.051	-0.017	7.349
	HBE	-0.01	0.033	-0.026	-0.315	0.753	-0.075	0.055
	HCT	0.001	0.013	0.01	0.12	0.905	-0.023	0.026
	MCV	-0.004	0.011	-0.028	-0.338	0.736	-0.026	0.019
	MCH	0.064	0.049	0.108	1.293	0.198	-0.034	0.161

[a]Dependent Variable: RBC

Table 4: Comparative outcomes of proposed scoring mechanisms with existing and derived indices.

Index	Formula	Beta thalassemia minor	Sensitivity	Specificity	PPV	NPV	Youden's Index
Mentzer[23]	MCV/RBC	<13	71	95	95	97	75
Srivastava[24]	MCH/RBC	<3.8	62	90	78	76	52
Shine & Lal[21]	(MCV2×MCH)/100	<1530	98	71	91	87	69
Jayabose *et al.*[25]	(MVC×RDW)/RBC	<220	53	95	65	78	48
Sirdah*et al.*[26]	MCV-RBC-3Hb	<27	67	97	58	73	64
Ehsani *et al.*[22]	$MCV - (10 \times RBC)$	<15	90	97	58	96	87
The current study)	Equation 1	<4.961	100	94	66	100	94

level of ≥4.961 was obtained.

This approach was further validated by a test dataset of 108 samples (Table S6) of HPLC screening reports. Among 108 participants, 11 samples were screened as beta thalassemia minor and 3 samples were heterozygotes for Hb E. The rest of population was normal. After applying the proposed model (Figure 1) and equation 1 to the test dataset, 29 true normal samples (Table S7) were detected, where no further HPLC test was needed, which is around 26.85 % without any false-negative for the dataset.

Discussion

Several studies were conducted to determine and compare the efficiency of the RBC indices for differentiating IDA from beta thalassemia minor as shown in Table S8. Shine and Lal index,[21] and Ehsani index[22, 23] have a very high sensitivity (Table 4). Youden's Index value increases where both the sensitivity and specificity are high. This study showed a higher value than Youden's index without false negative results. Bhargavetet al. also demonstrated different cut-off values proposed by different scientists (Table 4). Roth et al.[12] have proposed a new algorithm to separate the IDA population, but none of these separate the false negative ones with 100 % efficiency.

Das et al. proposed a five-point decision tree method based on the multilayer perceptron model with more than 99% accuracy to separate beta thalassemia minor from IDA population[14] Considering the limits and cut-off values of this study, more than 21% of the separated normal population were "True negative". By applying this model, two Hb E carriers found to be false negative samples. The proposed model of this study could decrease the necessity of HPLC with no false-negative results.

Conclusion

Based on the combined impact of Hb, RBC counts, MCV, MCH, and RDW a decision strategy was developed. The goal was not to miss any, even if it means examining additional HPLC instances which may turn out to have a normal HPLC pattern. Most importantly, the scores effectively predicted the true positive rate. As a result, a major portion of the population can be screened out at the early stages of the carrier screening programm, leading to significant savings in health expenditure. After the implementation of the proposed approach, around 26.85% of the test dataset could be eliminated from further HPLC screening. The results of this study could almost exclude chromatographic screening with 100% fidelity in our dataset. This approach needs to be further assessed with different datasets.

Acknowledgement

The authors are thankful to Calcutta School of Tropical Medicine, 108, Chittaranjan Ave, Calcutta Medical College, College Square, Kolkata, West Bengal 700073; and also, to Thalassaemia Society of Midnapore District, Midnapore Medical College and Hospital Campus, Midnapore, West Bengal, 721101 for providing the necessary facilities for HPLC test both training and testing dataset used in this paper.

Author contributions

All authors contributed to the study conception and design. Abhisek Samanta designed the study, collected data, and prepared manuscript. Paban Kumar Choudhury performed HPLC analysis, Ushnish Das performed statistical analysis, Nandan Bhattacharyya designed and supervised the study, and prepared final manuscript. All authors read and approved the final manuscript.

Ethics Approval

This study was approved by the Institutional Ethical Approval Committee, Panskura Banamali College, West Bengal, India. Consent to participate (include appropriate statements)

Availability of Data and Material

Test data set and other materials are available in Supplementary data

Conflict of Interest: None declared.

References

1. United Nations. World population prospects 2019: Highlights. 2019. 43 pages. **doi:** https://doi.org/10.18356/13bf5476-en.
2. Soteriades ES, Weatherall D. The Thalassemia International Federation: a global public health paradigm. *Thalass. Rep.* 2014; *4*: 7-12.doi: 10.4081/thal.2014.1840.
3. Pant L, Kalita D, Singh S, Kudesia M, Mendiratta S, Mittal M, et al. Detection of abnormal hemoglobin variants by HPLC method: common problems with suggested solutions. Int Sch Res Notices. 2015; 2015: 257805. doi: 10.1155/2014/257805.
4. De Sanctis V, Kattamis C, Canatan D, Soliman AT, Elsedfy H, Karimi M, et al. β-Thalassemia Distribution in the Old World: an Ancient Disease Seen from a Historical Standpoint. Mediterr J Hematol Infect Dis. 2017;9(1): e2017018. doi: 10.4084/MJHID.2017.018. eCollection 2017. PubMed PMID: 28293406. PubMed Central PMCID: PMC5333734.
5. Christianson A, Howson CP, Modell B. March of dimes: global report on birth defects, the hidden toll of dying and disabled children. Research report. March of Dimes Birth Defects Foundation, White Plains, USA, 2006.
6. Carnevale, E. World Population Highlights: Key Findings from PRB's 2007 World Population Data Sheet. Population Reference Bureau. 2008.
7. Population Reference Bureau. World Population Highlights: Key Findings from PRB's 2007 World Population Data Sheet. Population Bulletin. 2008;63(3): 1-16.
8. Available form: https://www.prb.org/wp-content/uploads/2008/08/Population-bulletin-2008-63.3highlights.pdf
9. Sinha S, Black ML, Agarwal S, Colah R, Das R, Ryan K, et al. Profiling β-thalassaemia mutations in India

at state and regional levels: implications for genetic education, screening and counselling programmes. Hugo J. 2009; 3(1-4): 51–62. doi: 10.1007/s11568-010-9132-3. PubMed PMID: 21119755. PubMed Central PMCID: PMC2882644.

10. Karimzaei T, Masoudi Q, Shahrakipour M, Navidiyan A, Jamalzae AA, Bamri AZ. Knowledge, attitude and practice of carrier thalassemia marriage volunteer in prevention of major thalassemia. Glob J Health Sci. 2015;7(5):364-70. doi: 10.5539/gjhs.v7n5p364. PubMed PMID: 26156937.
11. Cousens NE, Gaff CL, Metcalfe SA, Delatycki MB. Carrier screening for Beta-thalassaemia: a review of international practice. Eur J Hum Genet. 2010;18(10):1077-83. doi: 10.1038/ejhg.2010.90. PubMed PMID: 20571509.
12. Kukreti R, Dash D, Vineetha KE, Chakravarty S, Kr Das S, De M, et al. Spectrum of beta-thalassemia mutations and their association with allelic sequence polymorphisms at the beta-globin gene cluster in an eastern Indian population. Am J Hematol. 2002;70(4):269-77. doi: 10.1002/ajh.10117. PubMed PMID: 12210807.
13. Chatterjee T, Chakravarty A, Chakravarty S. Population screening and prevention strategies for thalassemias and other hemoglobinopathies of eastern India: Experience of 18,166 cases. Hemoglobin. 2015;39(6):384-8. doi: 10.3109/03630269.2015.1068799. PubMed PMID: 26428539.
14. Roth IL, Lachover B, Koren G, Levin C, Zalman L, Koren A. Detection of β-thalassemia carriers by red cell parameters obtained from automatic counters using mathematical formulas. Mediterr J Hematol Infect Dis. 2018; 10(1): e2018008. doi: 10.4084/MJHID.2018.008. PubMed PMID: 29326805.
15. Bhargava M, Kumar V, Pandey H, Singh V, Misra V, Gupta P. Role of hematological indices as a screening tool of beta thalassemia trait in eastern uttar pradesh: an institutional study. Indian J Hematol Blood Transfus. 2020; 36(4):719-724. doi: 10.1007/s12288-020-01282-z. PubMed PMID: 33100716.
16. Das R, Datta S, Kaviraj A, Sanyal SN, Nielsen P, Nielsen I, et al. A decision support scheme for beta thalassemia and HbE carrier screening. J Adv Res. 2020; 24:183–190.
17. Soliman AR, Kamal G, Walaa AE, Mohamed THS. Blood indices to differentiate between β-thalassemia trait and iron deficiency anemia in adult healthy Egyptian blood donors. Egypt J Haematol. 2014 ;39(3):91-7.
18. Cao A, Kan YW. The Prevention of Thalassemia. Cold Spring Harb Perspect Med. 2013; 3(2): a011775. doi: 10.1101/cshperspect.a011775.
19. Jahangiri M, Rahim F, Malehi AS. Diagnostic performance of hematological discrimination indices to discriminate between βeta thalassemia trait and iron deficiency anemia and using cluster analysis: Introducing two new indices tested in Iranian population. Scientific Reports. 2019; 9(1):18610. doi:10.1038/s41598-019-54575-3.
20. Machuca C, Vettore MV, Krasuska M, Baker SR, Robinson PG. Using classification and regression tree modelling to investigate response shift patterns in dentine hypersensitivity. BMC Med Res Methodol. 2017;17:120. doi: 10.1186/s12874-017-0396-3.
21. Maji SK, Mandal PK, Bera R, Dolai TK. The prevalence and characterization of β-thalassemia trait by using high-performance liquid chromatography among the rural population in West Bengal, India. Thal Rep. 2014. doi: 10.4081/thal.2014.2188.
22. Perseu L , Satta S, Moi P, Demartis FR, Manunza L, Sollaino MC, et al. KLF1 gene mutations cause borderline HbA2. Blood. 2011;118(16):4454-8. doi: 10.1182/blood-2011-04-345736. PubMed PMID: 21821711.
23. S Lal S. A strategy to detect beta-thalassaemia minor Lancet. 1977;1(8013):692-4. doi: 10.1016/s0140-6736(77)92128-6.
24. Ehsani MA, Shahgholi E, Rahiminejad MS, Seighali F, Rashidi A. A New Index for discrimination between iron deficiency anemia and beta-thalassemia minor: results in 284 patients. Pak J Biol Sci. 2009;12(5):473-5. doi: 10.3923/pjbs.2009.473.475.
25. William M. Differentiation of iron deficiency from thalassæmia trait. Lancet. 1973;1(7808):882. doi: 10.1016/s0140-6736(73)91446-3. PubMed PMID: 4123424.
26. Srivastava PC, Bevington JM. Iron deficiency and/or THALASSÆMIA trait. The Lancet.1973; 1(7807):832. doi: 10.1016/s0140-6736(73)90637-5.
27. Jayabose S, Giamelli J, LevondogluTugal O, Sandoval C, Ozkaynak F, Visintainer P. # 262 Differentiating iron deficiency anemia from thalassemia minor by using an RDW-based index. J Pediatr Hematol Oncol.1999;21(4):314
28. Sirdah M, Tarazi I, Al Najjar E, Al Haddad R. Evaluation of the diagnostic reliability of different RBC indices and formulas in the differentiation of the -thalassaemia minor from iron deficiency in Palestinian population. Int J Lab Hematol. 2008; 30(4):324-330. doi: 10.1111/j.1751-553X.2007.00966.x.
29. Lafferty JD, Crowther MA, Ali MA, Levine M. The Evaluation of Various Mathematical RBC Indices and Their Efficacy in Discriminating Between Thalassemic and Non-ThalassemicMicrocytosis. Am J ClinPathol [Internet]. 1996; 106(2):201-5. doi: 10.1093/ajcp/106.2.201.
30. Jiang F, Chen G-L, Li J, Xie XM, Zhou JY, Liao C, et al. Pre Gestational thalassemia screening in mainland China: the first two years of a preventive program. hemoglobin. 2017; 41(4-6):248-253. doi: 10.1080/03630269.2017.1378672.
31. Old JM, Varawalla NY, Weatherall DJ. Rapid detection and prenatal diagnosis of β-thalassaemia: studies in Indian and Cypriot populations in the UK. Lancet. 1990; 336(8719):834-7. doi: 10.1016/0140-6736(90)92338-i.
32. Rathod DA, Kaur A, Patel V, Patel K, Kabrawala R, Patel V, et al. Usefulness of cell counter-based

parameters and formulas in detection of beta-thalassemia trait in areas of high prevalence. Am J ClinPathol. 2007; 128(4):585-9. doi: 10.1309/R1YL4B4BT2WCQDGV.

33. Sahli CA, Bibi A, Ouali F, Hadj Fredj S, Dakhlaoui B, Othmani R,et al. Red cell indices: differentiation between β-thalassemia trait and iron deficiency anemia and application to sickle-cell disease and sickle-cell thalassemia. Clin Chem Lab Med. 2013;51(11):2115-24. doi: 10.1515/cclm-2013-0354.
34. Pornprasert S, Panya A, Punyamung M, Yanola J, Kongpan C. Red cell indices and formulas used in differentiation of β-thalassemia trait from iron deficiency in Thai adults. Hemoglobin. 2014;38(4):258-61. doi: 10.3109/03630269.2014.930044.

SUPPLEMENTARY DATA

Model formation:

Determination of score from decision tree analysis:

Decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. Based on the desired method of selecting input variables, the user goes to the dialogue box for the corresponding algorithm.

Table S1: Descriptive Statistics in beta thalassemia minor

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Std. Error	Statistic	Std. Error
RBC	289	4.45	3.14	7.59	5.0435	0.04290	0.72929	0.214	0.143	0.660	0.286
Haemoglobin	289	7.30	7.90	15.20	11.0831	0.09188	1.56200	0.329	0.143	-0.087	0.286
HCT	289	25.50	26.80	52.30	40.0997	0.37805	6.42687	0.096	0.143	-0.998	0.286
MCV	289	64.90	39.20	104.10	78.7204	0.84994	14.44904	-0.053	0.143	-1.282	0.286
MCH	289	12.90	16.30	29.20	21.8827	0.19548	3.32324	0.368	0.143	-0.774	0.286
Valid N (listwise)	289										

Table S2: Descriptive Statistics in IDA and beta thalassemia minor

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Std. Error	Statistic	Std. Error
RBC	804	3.53	4.06	7.59	5.1983	0.02339	0.66325	1.022	0.086	1.212	0.172
HAE	804	11.50	7.60	19.10	12.5846	0.07496	2.12556	0.411	0.086	0.967	0.172
HVT	804	29.00	26.80	55.80	38.9512	0.16942	4.80396	0.510	0.086	1.655	0.172
MCV	804	38.60	39.20	77.80	71.7949	0.19529	5.53737	-1.233	0.086	2.043	0.172
MCH	804	14.50	16.30	30.80	25.3050	0.12735	3.61107	-0.736	0.086	-0.507	0.172
Valid N (listwise)	804										

Table S3: Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Std. Error	Statistic	Std. Error
RBC	1992	3.35	3.76	7.11	5.0133	0.01335	0.59564	1.135	0.055	2.278	0.110
HAE	1992	11.50	7.60	19.10	12.8030	0.04276	1.90830	0.152	0.055	1.860	0.110
HCT	1992	42.50	13.30	55.80	39.6264	0.10476	4.67548	0.372	0.055	2.304	0.110
MCV	1992	17.60	78.50	96.10	84.3701	0.09669	4.31530	0.350	0.055	-1.159	0.110
MCH	1992	14.10	16.90	31.00	26.8886	0.05630	2.51281	-1.275	0.055	2.306	0.110
Valid N (listwise)	1992										

Equation formation by multiple linear regression:

Using multilayer perceptron analysis, the Likelihood Ratio was derived which suggested MCH as a primary predictor variable, multiple logistic regression also proved that. The Model Information table describes the data set, the response variable, the number of response levels, the type of model, and the algorithm used to obtain the parameter estimates.

The Optimization Technique was the iterative numerical technique that PROC LOGISTIC uses to estimate the model parameters. The model was assumed to be binary logit when there were exactly two response levels.

For MCH the Model Fit Statistics table reports the results of three goodness-of-fit tests measures:

AIC or Akaike Information Criterion adjusts for the number of predictor variables

SC or Schwarz's Bayesian Criterion or SBC adjusts for the number of predictor variables and the number of observations. SC uses a bigger penalty for extra variables, and therefore, favours more parsimonious models

-2 Log L, which is -2 times the natural log of the likelihood. The score could be reduced by adding more regression parameters to the model. Which was used to compare the fit of models that use different numbers of parameters of nested models using likelihood ratio tests for measuring relative fit for comparing models and smaller values for all these measures indicate better fit.

The univariate measures of parameter MCH were found to be the lowest as compared to all other parameters, leading to its selection as a primary predictor in the discussed model.
The Global Tests table, Testing Global Null Hypothesis: BETA=0, provides three statistics to test the null hypothesis that all regression coefficients of the model were 0. A significant p-value (Pr>ChiSq) provides evidence of regression coefficients for the predictor variable, MCH was significantly different from 0. It was most significant when fared against other contributing parameters in the discussed model. The Likelihood Ratio Chi-Square value was calculated as the difference between the -2 Log L value of the baseline model (intercept only) and the -2 Log L value of the hypothesized model (intercept and covariates). The Score and Wald tests were also used to test whether all the regression coefficients were 0, and all three tests were asymptotically equivalent and often give very similar values. The Wald Chi-Square and associated p-value ensure the parameter, MCH estimate is statistically significant.
The degrees of freedom were equal to the difference in several parameters between the hypothesized model and the baseline model. Here predictor, MCH, has been compared to the intercept-only model. The Parameter Estimates table, Analysis of Maximum Likelihood Estimates, lists the estimated model parameters, their standard errors, Wald Chi-Square values, and p-values.
Here, the logistic regression equation is logit (p-hat) = 8.2864 (-0.4039) * MCH.
The p-value for the variable MCH is significant at the 0.05 alpha level and that the probability of being Thalassemia detection decreases as MCH value increases.
The odds and the odds ratio from the logistic regression model were calculated.
Logit (p)=log(odds)=$\beta0+\beta1$* MCH
For a continuous predictor variable, such as MCH, the odds ratio measures the increase or decrease in odds associated with a one-unit difference of the predictor variable. The odds ratio for MCH indicates that the odds of detecting Thalassemia decrease by 0.332% for each increase in one unit MCH value. The 95% confidence interval, 0.643 to 0.693, does not include 1.000, the odds ratio was significant at the 0.05 alpha level, and therefore, the predictor MCH was significantly different from 0. Percentages of concordant, discordant, and tied pairs as goodness-of-fit measures to compare one model to another. Higher percentages of concordant pairs and lower percentages of discordant and tied pairs indicate a more desirable model. ***The table***also shows the four rank correlation indices that were computed from the numbers of concordant, discordant, and tied pairs of observations: Somers' D, Gamma, Tau-a, and c.

Table S4: Multiple logistic regression model

Association of Predicted Probabilities and Observed Responses				
Percent Concordant		87.6	Somers'D	0.755
Percent Discordant		12.1	Gamma	0.757
Percent Tied		0.3	Tau-a	0.178
Pairs		1068189	c	0.877
Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95%ConfidenceLimits	
MCH	1.0000	0.651	0.622	0.679
HGB	1.0000	0.705	0.657	0.756
MCV	1.0000	1.026	1.011	1.042
HCT	1.0000	1.089	1.061	1.118

In general, a model with higher values for these indices has better predictive ability than a model with lower values. The c, concordance statistic, estimates the probability of observation with the event having a higher predicted probability than an observation without the event. The c value was calculated as the number of concordant outcomes plus one-half times the number of ties divided by the total number of pairs. The range of possible values is 0.5 to 1.0, where 1.0 is perfect prediction. The value of 0. 0.844 shows a very strong ability of MCH to discriminate between Thalassemia detection or otherwise.
The profile likelihood confidence intervals were different from the Wald-based confidence intervals.
This difference was visible because the Wald confidence intervals use a normal error approximation, whereas the profile likelihood confidence intervals were based on the value of the log-likelihood. These likelihood-ratio confidence intervals require a much greater number of computations but were generally preferred to the Wald confidence intervals, especially for sample sizes less than 50.
The Odds Ratio plot displays the results of the Odds Ratio table graphically. This plot was obtained by applying the parameter estimates from the logistic model to values of the predictors and then converting the predictions to the probability scale. A reference line shows the null hypothesis, an odds ratio equal to 1. When the confidence interval crosses the reference line, the effect of the variable is not significant. Calculating and interpreting odds ratios for categorical variables was similar to that of continuous variables. A logistic regression model with the predictor Lot_Shape_2 instead of Basement_Area. Lot_Shape_2 has only two levels, Regular and Irregular were fitted.
The logit of p was also equal to the linear predictor the redundant level represents the regular level. Regular lot shapes were coded as 0 and Irregular lot shapes are coded as 1. To obtain the odds for an Irregular lot shape, the linear predictor for the level was exponentiated. First, we substitute 1 for Lot_Shape_2 to get $\beta0 + \beta1$ as the linear predictor. Then, we add the parameter estimates that we got for $\beta0$ and $\beta1$ and exponentiate the sum. To obtain the odds for a Regular lot shape, the same process was followed.
First, we substitute 0 for Lot_Shape_2 to get $\beta0$ as the linear predictor. Then we take the parameter estimate that we got for $\beta0$ and exponentiate it. The odds ratio was then the odds for the Irregular lot shape divided by the odds for a Regular lot shape.
The profile likelihood confidence intervals were different from the Wald-based confidence intervals.
This difference was because the Wald confidence intervals use a normal error approximation, whereas the profile likelihood confidence intervals were based on the value of the log-likelihood. These likelihood-ratio confidence intervals require a much greater number of computations, generally preferred to the Wald confidence intervals, especially for sample sizes less than 50.
The Odds Ratio plot displays the results of the Odds Ratio table graphically. This plot was obtained by applying the parameter estimates from the logistic model to values of the predictors and then converting the predictions to the probability scale. A reference line shows the null hypothesis, an odds ratio equal to 1. When the confidence interval crosses the reference line, the effect of the variable was not significant.

Calculating and interpreting odds ratios for categorical variables was similar to that of continuous variables. A logistic regression model with the predictor Lot_Shape_2 instead of Basement Area was imagined. Lot_Shape_2 has only two levels, Regular and Irregular. The logit of p was also equal to the linear predictor for our model. In this case, we use the level regular to represent the redundant level. So, Regular lot shapes were coded as 0 and Irregular lot shapes are coded as 1. To obtain the odds for an Irregular lot shape, the linear predictor for the level was exponentiated.

First, 1 for Lot_Shape_2 to get β0 + β1 as the linear predictor was substituted. The parameters were estimates that we got for β0 and β1 and exponentiate the sum.

To obtain the odds for a Regular lot shape, the same process was followed. First, 0 for Lot_Shape_2 to get β0 as the linear predictor was a substitute and divided by the odds for a Regular lot shape.

Table S5: Logistic regression models predicting Thalassemia detection status

	Model 1: Unadjusted			Model: Fully Adjusted		
		95% Confidence Interval			95% Confidence Interval	
Category	**Odds Ratio**	**Lower Limit**	**Upper Limit**	**Odds Ratio**	**Lower Limit**	**Upper Limit**
MCH	0.668	0.643	0.693	0.651	0.622	0.679
Reference	Reference	Reference	Reference	Reference	Reference	Reference
HGB				0.705	0.657	0.756
MCV				1.026	1.011	1.042
HCT				1.089	1.061	1.118

* MCH - Mean Corpuscular Hemoglobin, HGB - Hemoblobin, MCV - Mean Corpuscular Volume, HCT - Hematocrit value.

Validation:

Table S6: Validating data set with 108 participants were used as testing dataset

Variant	RBC	Hemoglobin	HCT	MCV	MCH	RDW	Cut off value
0	5.21	13.3	44.1	84.6	25.5	13.3	4.8647
0	4.72	13	41	86.9	27.5	13.2	4.9834
0	4.45	12.4	40.2	90.3	27.9	13.3	5.0006
0	4.91	15.3	46.7	95.1	31.2	13.2	5.1701
0	5.06	14.2	44.4	87.7	28.1	14.2	5.01
0	3.68	10.1	31.5	85.6	27.4	14.9	5.0017
0	4.62	13.4	41.5	89.8	29	12.7	5.0643
0	4.45	13.8	42	94.4	31	12.8	5.1704
0	4.63	2.4	40	86.4	26.8	13.6	5.0456
0	4.87	13.7	41.6	85.4	28.1	14.2	5.0214
0	4.48	12.1	37.7	84.2	27	14.9	4.9679
0	4.26	10.4	33.9	79.6	24.4	14.2	4.8331
0	4.25	11.6	37.6	88.5	27.3	15.1	4.9748
0	4.84	13.9	43.4	89.7	28.7	14.2	5.0424
0	4.6	13.2	39.5	85.9	28.7	14.5	5.0607
0	4.62	11.9	39.2	84.8	25.8	14.2	4.8922
0	5.55	12.6	41.9	75.5	22.7	16	4.7267
0	4.99	15.3	46.3	92.8	30.7	14.5	5.1469
0	5.44	13.9	45.6	83.8	26.1	13.5	4.9018
0	4.41	12	39.4	89.4	27.2	12.3	4.9626
0	4.13	11.6	37.2	90.1	28.1	14.8	5.0192
0	4.96	10.7	36	72.6	21.6	14.6	4.681
0	4.42	12.5	39.2	88.7	28.3	13.7	5.0306
0	5.93	15.8	48.6	82	26.6	13.8	4.925
0	5.72	14	45.4	79.4	24.5	14.8	4.8158
0	5.34	11.3	37.8	70.8	21.2	15.2	4.6584
0	5.52	14.9	44.2	80.1	27	12	4.9628
0	4.41	12	37.9	85.9	27.2	12.8	4.9751
0	6.07	14.6	46.3	75.3	24.1	14.1	4.8015
0	5.78	14.1	46.6	80.5	24.4	12.3	4.8052
0	5.85	15.3	48.8	83.4	26.2	12.4	4.899
0	6.46	13.5	44.6	69	20.9	15.3	4.6312
0	5.1	14	44	86.3	27.5	12.2	4.9788
0	5.06	13.6	43.2	81.4	26.9	13.3	4.9632
0	4.95	13.9	42.4	85.7	28.1	13.7	5.019
0	5.11	13.8	42.9	84	27	14.4	4.9569

0	5.1	13.5	42.7	82.7	26.5	14.1	4.9329
0	4.44	10.4	35.7	80.4	23.4	16.9	4.7677
0	4.09	11.2	36.3	88.8	27.4	13.5	4.9827
0	5.53	14.3	44.3	80.1	25.9	13.6	4.8985
0	4.55	12.7	39.5	86.8	27.9	13.2	5.0109
0	4.85	14	42.1	86.8	28.9	12.6	5.0645
0	4.46	12.5	39.2	87.9	28	13.5	5.0146
0	4.42	11.4	37.2	84.2	25.8	15.3	4.8976
0	4.25	12.9	40.4	95.1	30.4	14.3	5.1366
0	4.49	11.7	36.1	80.4	26.1	15	4.9279
0	5.09	15.1	45.9	90.2	29.7	14.3	5.0949
0	5.29	14.9	45.6	86.2	28.2	13.4	5.0166
0	5.9	14.4	45.6	77.3	24.4	15	4.814
0	4.74	14.7	44.5	93.9	31	13	5.1659
0	4.92	14	43.3	88	28.5	15.2	5.0353
0	5.61	11.7	38.3	78	23.5	15	4.7733
0	5.22	14.8	46.6	89.3	28.4	13.3	5.019
0	4.71	14	42	89.2	29.7	13.8	5.106
0	5.57	14.8	46.4	83.3	26.6	12.4	4.9276
0	4.79	11.7	37.2	77.7	24.4	12.8	4.831
0	4.07	11.5	35.7	87.7	28.3	12.1	5.0411
0	5.05	13.7	43.6	86.3	27.1	13.2	4.9558
0	4.28	11.9	37.3	87.1	27.8	11.6	5.0091
0	4.66	12.1	37.3	80	26	13.4	4.9203
0	5.17	12.1	39	75.4	23.4	13.3	4.774
0	5.15	9.6	33.4	64.9	18.6	16.1	4.5282
0	5.1	11	36.5	71.6	21.6	14.9	4.6825
0	4.52	13	39.7	87.8	28.8	13.1	5.0617
0	4.67	12.8	39.7	85	27.4	12.9	4.9853
0	4.32	11.6	35.7	82.6	26.9	12.4	4.9709
0	4.73	11.7	38.5	81.4	24.7	12.8	4.8367
0	5.57	14.5	46.3	83.1	26	11.9	4.8929
0	4.97	13.4	41.1	82.7	27	12.5	4.9643
0	5.52	13.9	44.8	81.2	25.2	13.4	4.8538
0	5.62	13.5	42.4	75.4	24	14.3	4.8018
0	5.02	13.4	42.5	84.7	26.7	12.4	4.9385
0	5.65	14.4	47.2	83.5	25.5	13.4	4.8612
0	4.89	12.3	39.9	81.6	25.2	13	4.8633
0	5.09	13.8	44.4	87.2	27.4	12.6	4.9712
0	4.88	14	41.9	85.9	28.7	12.7	5.0551
0	5.41	13.4	40.6	75	24.8	15.4	4.8538
0	5.35	13.9	45.4	84.9	26	14	4.8908
0	5.89	12.8	42.4	72	21.7	14.8	4.6752
0	4.71	12.7	40.8	86.6	27	13.2	4.9554
0	5.11	13.3	42.2	82.6	26	12.4	4.9028
0	4.93	14.1	41.5	84.2	28.6	13.4	5.0541
0	4.28	12	36.8	86	28	12.3	5.0248
0	4.31	10.3	33	86.6	23.9	15.6	4.7732
0	4.55	11.3	37.5	82.4	24.8	14.1	4.8421
0	5.54	14.2	45.5	82.1	25.6	12.5	4.8735
0	4.59	10.7	36.8	80.2	23.3	12.1	4.7602
0	5.25	10.9	36.8	70.1	20.8	14.7	4.6386
0	4.57	11.9	38.8	84.9	26	12.6	4.9042
0	4.29	9.3	33.1	77.2	21.7	17.2	4.6801
0	4.87	11.9	39.7	81.5	24.4	14.3	4.8163
0	5.85	15.1	47.1	80.5	28.5	12.1	5.0581
1	4.23	12.3	49.2	95.2	25.2	12.6	4.8182
1	4.23	11.2	46.2	90.2	24.2	12.3	4.7822
1	5.23	14.2	52.3	98.2	27.2	12.3	4.9183
1	4.23	11.2	46.2	90.2	24.2	12.3	4.7822

1	4.23	12.3	49.2	95.2	25.2	12.6	4.8182
1	4.93	12.9	48.2	97.8	26.2	14.1	4.8648
1	6.76	11.4	46.1	68.2	16.9	15.6	4.4009
1	4.1	9.5	38.6	94.1	23.2	20.2	4.712
1	4.92	12.7	44	89.4	25.8	12.6	4.8706
1	4.15	10.2	42.1	101.4	24.6	16.9	4.7689
1	5.33	14	51	95.7	26.3	11.8	4.8714
1	5.05	8.5	42.8	84.8	16.8	15.2	4.3538
2	5.95	10.8	44.2	74.3	18.2	13.5	4.4638
2	4.77	12.1	43.9	92	25.4	11.2	4.8405
2	4.26	11.1	40.4	94.8	26.1	13.6	4.8806

Variant column denote HPLC results, whereas 0 denote normal, 1 denote BTT and 2 denote HbE

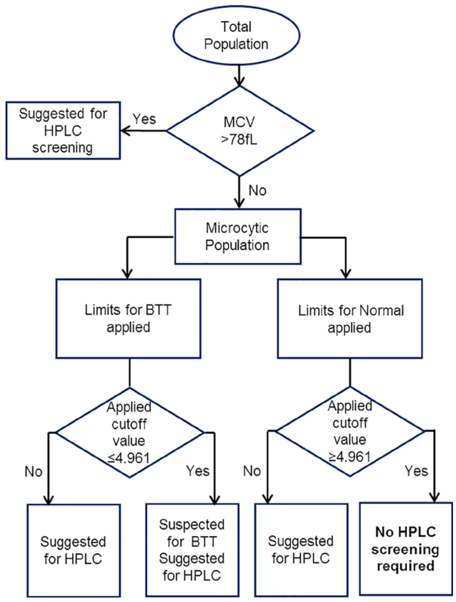

Figure S1: Decision tree analysis with two critical parametric limits for separating beta thalassemia minor and true normal from microcytic population. Using derived parametric limit and equation, 29 samples were separated as a true negative population which was more than 26%.

Table S7: Final derived true negative normal data

Variant	RBC	Hemoglobin	HCT	MCV	MCH	RDW	Cut-off Value
0	4.72	13	41	86.9	27.5	13.2	4.9834
0	4.45	12.4	40.2	90.3	27.9	13.3	5.0006
0	4.91	15.3	46.7	95.1	31.2	13.2	5.1701
0	5.06	14.2	44.4	87.7	28.1	14.2	5.01
0	4.62	13.4	41.5	89.8	29	12.7	5.0643
0	4.45	13.8	42	94.4	31	12.8	5.1704
0	4.87	13.7	41.6	85.4	28.1	14.2	5.0214
0	4.84	13.9	43.4	89.7	28.7	14.2	5.0424
0	4.99	15.3	46.3	92.8	30.7	14.5	5.1469
0	5.52	14.9	44.2	80.1	27	12	4.9628
0	5.1	14	44	86.3	27.5	12.2	4.9788
0	5.06	13.6	43.2	81.4	26.9	13.3	4.9632
0	4.95	13.9	42.4	85.7	28.1	13.7	5.019
0	4.85	14	42.1	86.8	28.9	12.6	5.0645
0	4.25	12.9	40.4	95.1	30.4	14.3	5.1366
0	5.09	15.1	45.9	90.2	29.7	14.3	5.0949
0	5.29	14.9	45.6	86.2	28.2	13.4	5.0166
0	4.74	14.7	44.5	93.9	31	13	5.1659
0	4.92	14	43.3	88	28.5	15.2	5.0353
0	5.22	14.8	46.6	89.3	28.4	13.3	5.019
0	4.71	14	42	89.2	29.7	13.8	5.106
0	4.52	13	39.7	87.8	28.8	13.1	5.0617
0	4.67	12.8	39.7	85	27.4	12.9	4.9853
0	4.97	13.4	41.1	82.7	27	12.5	4.9643
0	5.09	13.8	44.4	87.2	27.4	12.6	4.9712
0	4.88	14	41.9	85.9	28.7	12.7	5.0551
0	4.93	14.1	41.5	84.2	28.6	13.4	5.0541
0	5.85	15.1	47.1	80.5	28.5	12.1	5.0581

Table S8: Cut-off values for haematological parameters in some existing literature.

Index	MGH (pg)	MCV(fL)	RBC($10^6 \mu L^{-1}$)	RDW
Lafferty *et al.*,[27]	-	<72	-	-
Jiang *et al.*,[28]	-	<80	-	-
Old *et al.*,[29]	<27	<79	-	-
Rathod *et al.*,[30]	<27	<76.5	>5	>13.5
Sahli *et al.*,[31]	<23	<75	>5	>14
Cao *et al.*,[16]	<27	<78	-	-
Pornprasert *et al.*,[32]	<27	<76	>5	>14