ORIGINAL ARTICLE

Feasibility of a Model-based HbA2 Prediction for Low-cost Screening of β-thalassemia Trait Using Blood-derived Indices: Opportunities and Challenges

Ramadevi Peraka[1*], Pradeep Kumar Reddy M[2], Vanajakshi S[3], Sumedha Avadhanam[4], Manimekhala Parsa[5], Aditya Koppula[6]

Received: 02 February 2026; *Accepted:* 06 July 2026

Visual Abstract

Supplement Material

Abstract

Background: β-thalassemia trait (β-TT) is a common genetic condition with significant public health implications, particularly in regions with high prevalence. Diagnosis conventionally relies on HbA2 quantification using high-performance liquid chromatography or Hemoglobin electrophoresis, which are costly and inaccessible in resource-limited settings. Developing low-cost screening/triage tools is therefore essential.

Materials and methods: A retrospective single-center cross-sectional study was conducted at the Department of Pathology, Apollo Institute of Medical Sciences & Research, Hyderabad, between February 2023 and December 2025. Data from 129 patients with complete blood counts, iron profiles, and HbA2 measurements were analyzed. LASSO method was applied to identify salient predictors of HbA2 from demographic, hematological, and iron parameters. Two ordinary least squares regression models (OLS) were constructed: one excluding iron parameters (model 1) and another incorporating iron parameters (model 2). LASSO-then-OLS approach was used for both 5-fold cross-validation and final model development. ROC analysis was also done.

Results: Of the 129 patients, 15 (11.6%) had β-TT (HbA2 >3.5%). The 5-fold cross-validation found lower root mean squared error (RMSE) and superior variable selection stability with an iron parameter-inclusive model. Salient predictors in the final model were age, RBC count, MCV, and RDW-CV in model 1, while model 2 additionally included serum ferritin, transferrin saturation, and total iron binding capacity (TIBC). Age and RBC count were positively associated with HbA2, while MCV, RDW-CV, serum ferritin, and TIBC were negatively associated. Model 1 explained 40% of HbA2 variance, while model 2 explained 47% variance. Model imprecision was more pronounced for β-TT cases. ROC analysis showed an area under the curve (AUC) value of 0.868 (0.766–0.969) for model 1 and 0.904 (0.829–0.98) for model 2. Optimal cutoff (3.31% from model 1 and 3.26% from model 2) for β-TT detection was lower than the conventional threshold.

Conclusion: Regression-based machine learning models have the potential to estimate HbA2 using routine blood indices, offering a cost-effective screening/triage before specialized assays. Such models are particularly valuable in populations with prevalent iron deficiency anemia (serum ferritin <12 ng/ml), where fixed HbA2 cutoffs are unreliable. Larger studies incorporating peripheral smear image analysis are needed to improve the predictive accuracy of models before they can be deployed clinically.

Keywords: β-thalassemia trait, Least absolute shrinkage and selection operator, Machine learning, OLS regression, ROC analysis, Screening.

Journal of The Association of Physicians of India (2026): 10.59556/japi.74.1706

Introduction

Globally, the burden of β-thalassemia is considerable. A recent systematic review and meta-analysis has revealed that nearly 3–5% of the population are carriers of thalassemia genes, called β-thalassemia trait (β-TT).[1] The progeny of such carriers are at a high risk of full-blown β-thalassemia, underscoring the importance of accurate screening and diagnosis of the β-TT. In regions with high prevalence,[2] such as India and other South Asian countries, the Middle East, and the Mediterranean, Population-level screening and detection of β-TT are challenging and have profound public health implications. Failure to identify carriers of β-TT not only affects individual patient management but also hampers preventive strategies aimed at reducing the incidence of severe thalassemia syndromes in future generations. The β-TT is conventionally diagnosed by HbA2 >3.5% with high-performance liquid chromatography (HPLC) or hemoglobin electrophoresis.[3] Both techniques are expensive and not widely available in the remote, resource-limited settings in low- and middle-income countries, posing a formidable public health challenge. It is therefore imperative to develop low-cost mass screening tools for β-TT.

β-thalassemia is due to mutations in both of the β-globin genes, leading to reduced or no production of β-globin protein, while the α-globin synthesis continues unabated, leading to excess accumulation of α-globin chains that form toxic precipitates in developing RBC precursor cells in the bone marrow. This leads to excessive cell death (apoptosis) of erythroid precursors (ineffective erythropoiesis) and premature hemolysis of the RBC in the spleen. Serious consequences include chronic anemia, hypoxia, bone marrow expansion, skeletal deformities, and a lifelong requirement of periodic blood transfusions and their attendant adverse consequences like systemic iron overload. The affected patients may overproduce fetal hemoglobin, which reduces the severity of the disease. In contrast, β-TT is associated with a single mutated β-globin gene, which reduces but does not eliminate β-globin production, leading to a milder and very often asymptomatic clinical phenotype.[4,5] The progeny of two such β-TT individuals has a 25% risk of major β-thalassemia and a 50% risk of carrier state. The β-TT individuals overproduce HbA2 as a compensatory strategy to mitigate the imbalance in α-β synthesis. This high HbA2 (>3.5%) is the diagnostic hallmark of β-TT. Iron deficiency anemia (IDA), widely prevalent in India,[6] can mask β-TT by reducing HbA2 levels. IDA is defined as serum ferritin <12 ng/ml and low hemoglobin for the age and sex.[7–9] On the other hand, mild-to-moderate pure IDA causes a similar clinical and hematological profile and renders the task of

[1]Associate Professor; [2]Assistant Professor; [3]Associate Professor, Department of Pathology; [4]Final MBBS Student; [5]Professor and Head, Department of Pathology, Apollo Institute of Medical Sciences and Research; [6]Assistant Professor, Department of Physiology, ESIC Medical College and Hospital, Hyderabad, Telangana, India; *Corresponding Author

How to cite this article: Peraka R, Reddy MPK, Vanajakshi S, *et al.* Feasibility of a Model-based HbA2 Prediction for Low-cost Screening of β-thalassemia Trait Using Blood-derived Indices: Opportunities and Challenges. J Assoc Physicians India 2026;74(9):19–24.

β-TT screening challenging. Co-existing β-TT and IDA will further lower HbA2, complicating the interpretation using a fixed cutoff (like HbA2> 3.5%). Hence, it is necessary to characterize the factors affecting HbA2 levels to support its judicious interpretation in a clinical context. Identification of all the salient sources of variability in HbA2 is also a prerequisite for devising a robust and reliable machine learning model for the estimation of HbA2 levels.[10]

Several discriminant indices based on routine blood count parameters have been proposed as a screening test of β-TT, including Mentzer's Index (MI).[11] The MI (MCV/RBC count), introduced in 1973, remains one of the most widely used indices, and values greater than 13 suggest IDA and less than 13 indicate β-TT. While simple and cost-effective, these discriminant indices alone are not universally reliable, as overlap between IDA and β-TT cases can occur. Consequently, there is a growing interest in refining predictive models by integrating multiple hematological parameters. In this work, we devise and assess the feasibility of combining traditional red blood cell indices[12] with predictive machine learning algorithms,[10] to develop models that estimate HbA2 levels indirectly, thereby offering an accessible adjunct to specialized laboratory techniques. Such models could serve as valuable triage tools in primary care and resource-limited settings, enabling clinicians to make informed decisions regarding further testing, treatment, and genetic counseling. By leveraging readily available hematological data, this model has the potential to improve patient outcomes, optimize resource utilization, and contribute to global efforts in reducing the burden of thalassemia.

Although HbA2 testing is available at referral laboratories, access remains inconsistent across India. Many peripheral facilities lack HPLC/capillary gel electrophoresis capability, and referral requires temperature-controlled transport and multi-day turnaround times; high ambient temperatures frequently compromise sample integrity and HbA2 accuracy. Because most referral pathways lack reliable cold-chain support, test quality is variable, and delays disrupt antenatal screening workflows. Moreover, while the per-test price is modest, the logistical and programmatic costs of scaling population-level screening are substantial. These constraints explain the limited uptake of HbA2 testing in many regions and support the need for low-cost triage tools. In addition, IDA can confound HbA2 interpretation as discussed previously. Our model is intended for preliminary screening in such settings, and not as a replacement for definitive laboratory testing. This study has two-fold objectives: 1. To examine the feasibility of a machine-learning model to predict HbA2 from blood-derived parameters for β-TT screening. 2. To determine the factors affecting the variability of HbA2.

Materials and Methods

This retrospective single-center cross-sectional study was conducted at the Department of Pathology, Apollo Institute of Medical Sciences and Research (AIMSR), Hyderabad. Inpatients and outpatients of all age groups, both males and females, who visited AIMSR between February 2023 and December 2025 for whom the complete blood picture, iron profile, and HbA2 were available, were included in the study. Thalassemia major and other hemoglobinopathies were excluded from the study. Routine screening of medical records identified the participants, and their data were retrieved from the hospital databases. The institutional ethics committee-Apollo Institute of Medical Sciences and Research (AIMSR), Hyderabad (EC/NEW/INST/1527/2025/12/010 dated December 23, 2025) approved the study.

Following enrollment, the following data were collected from hospital databases:

HbA2 estimated by hemoglobin capillary gel electrophoresis; Complete blood picture was done using Mindray 7-part Coulter. The Coulter-derived parameters are:

- RBC parameters, i.e., RBC count, Hemoglobin (Hb), Hematocrit (Hct), red cell distribution width (RDW-CV%), Mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), and mean corpuscular hemoglobin concentration (MCHC). From the MCV and the RBC count, MI was computed.
- WBC parameters: total WBC count, absolute (Absolute neutrophil, lymphocyte, monocyte, eosinophil, and basophil count) and differential cell counts (neutrophil%, lymphocyte%, monocyte%, eosinophil%, basophil%).
- Platelet indices: platelet count, mean platelet volume (MPV), and platelet distribution width (PDW).

Serum iron profile[1,3,13]: Serum iron, serum ferritin, Total iron binding capacity (TIBC), unsaturated iron binding capacity, and transferrin saturation (%).

Raw continuous variables were summarized as mean ± standard deviation (SD), while categorical variables were represented as frequencies. Missing data-related information is summarized in Supplementary information, Table S1. Missing data were statistically recovered using multiple imputation analysis under a missing at random (MAR) assumption in accordance with the STROBE guidelines for missing data. Continuous hematological predictors were imputed using PMM (predictive mean matching) (m = 1, 40 iterations). Categorical variables (sex, β-TT status) were excluded from the imputation model. PMM was chosen because it preserves the observed distribution and avoids implausible values. The completed dataset was used for all subsequent analyses.

As the analysis involved a large number of variables, a two-stage modeling approach was adopted to predict HbA2 from blood and demographic variables. The approach consisted of: 1. Least Absolute Shrinkage and Selection Operator (LASSO) based feature-selection followed by: 2. Post-selection ordinary least squares (OLS) regression. This two-stage approach was used for both *Model validation* (5-fold cross-validation) and *Final model development*. In the first stage, salient variables were identified from the large variable set using the LASSO technique. All continuous variables were first scaled to unit variance and zero mean before the LASSO. The LASSO was implemented with a 5-fold inner cross-validation within the training dataset to determine the optimal penalty parameter ('λ'). As the dataset contains a small number of β-TT cases relative to non-β-TT patients, it creates an inherent class imbalance. This imbalance was addressed by using observation-level weighting within the LASSO framework, assigning higher weights (w = 5) to β-TT cases and lower weights (w = 1) to non-β-TT cases during model training. This weighted penalized regression model increases the influence of minority observations while preserving the original data distribution. All the variables with nonzero regression coefficients were considered salient variables for the prediction of HbA2. Using the salient predictors determined by the LASSO method, two multiple linear regression models (OLS regression) were devised from the training dataset to predict HbA2, the first model without iron parameters and the second one including iron parameters.

Overall model performance and generalization were assessed using 5-fold cross-validation. The dataset was randomly divided into five approximately equal subsets; in each iteration, four subsets were used for training and one for validation, with each subset serving as the validation set once. For each fold, the partitioning was done, ensuring that β-TT patients were divided in a 70:30 ratio between the training and validation sets, respectively. Within each fold,

the full modeling pipeline outlined above, consisting of LASSO followed by OLS-fit on selected features, was implemented. Predictive performance was evaluated using the root mean squared error (RMSE), computed between observed and predicted values in each validation fold. The overall performance was summarized by reporting the mean RMSE across folds along with its standard deviation. The stability of feature selection was assessed by recording the frequency with which each predictor was selected across the 5-fold cross-validation. This provides an indication of the consistency of variable selection under different data partitions.

Two final models, with and without iron profile parameters, were constructed using the full dataset, with both the 5-fold cross-validation and final model development performed separately for each predictor set. The goodness-of-fit for each model was expressed as adjusted R^2 and AIC (Akaike Information Criterion). The results of the multiple regression models are expressed as regression coefficients, their standard errors, and 95% confidence intervals (95% CI), *t*-values, and the *p*-values for each variable. A $p < 0.05$ was considered statistically significant. The ability of the model-fitted values from models 1 and 2 to correctly classify β-TT (defined as measured HbA2 > 3.5%) was assessed using ROC (receiver operating characteristic) analysis. The area under the curve was estimated, and Youden's index-based optimal cutoff for the model fit value was determined. From the optimal cutoff, sensitivity, specificity, positive predictive value, and negative predictive value were determined. Statistical analysis was done using R software version 4.5.2. Multiple imputation for missing data was done using the mice R package. Weighted LASSO with 5-fold cross-validation was implemented using the glmnet R package with weights passed directly to the penalized regression objective function. ROC analysis was implemented using the pROC package.

Results

The data of 129 participants were included in the study. Of them, 15 patients (11.6%) had β-TT, based on the HbA2 cutoff of >3.5%. The demographic and hematological profile of the cohort is summarized in Table 1, and the iron profile results are shown in Table 2. Missing data are summarized in supplementary information, Table S1.

The 5-fold cross-validation was performed separately for predictor sets including and excluding iron profile parameters. The model including iron parameters demonstrated improved predictive performance, with a mean RMSE of 1.03 (SD: 0.14), compared to a higher RMSE of 1.09 (SD: 0.15) for the model without these variables. Fold-wise RMSE values ranged from 0.83 to 1.18 in the iron-inclusive model and from 0.98 to 1.35 in the model excluding iron parameters. The number of predictors selected by the LASSO was more consistent in the model with iron parameters (8–9 variables across folds) than in the model without iron parameters (5–9 variables). Feature selection stability analysis showed that age, RBC count, RDW-CV, and MCHC were consistently selected across all folds in both models, indicating robust associations with the outcome. In the iron-

Table 1: Demographic and hematological characteristics of the Non-β-TT and the β-TT group

S. no	*Variable*	*Non-β-TT (HbA2 ≤ 3.5%) Mean (SD)*	*β-TT (HbA2 >3.5%) Mean (SD)*
1	Age (years)	15.1 (14.8)	27.8 (12.4)
2	WBC count (/μL)	9967.2 (4953.0)	8972.2 (2843.2)
3	Neutrophil%	49.2 (20.3)	62.0 (13.1)
4	Lymphocyte%	41.4 (20.2)	28.0 (14.6)
5	Monocyte%	6.4 (2.7)	8.6 (6.1)
6	Eosinophil%	2.8 (3.6)	2.1 (1.4)
7	Basophil%	0.3 (0.6)	0.2 (0.4)
8	Absolute neutrophil count (/μL)	4646.2 (4140.7)	5570.0 (2020.8)
9	Absolute lymphocyte count (/μL)	3960.7 (3165.4)	2651.8 (2261.5)
10	Absolute monocyte count (/μL)	581.7 (359.2)	778.8 (682.5)
11	Absolute eosinophil count (/μL)	278.2 (521.2)	148.8 (101.3)
12	Absolute basophil count (/μL)	30.5 (59.5)	18.2 (31.6)
13	RBC count (millions/μL)	4.5 (0.9)	5.5 (0.9)
14	Hemoglobin (gm/dL)	9.1 (2.6)	10.5 (2.2)
15	Hematocrit (Hct)%	29.9 (7.5)	33.9 (6.1)
16	Mean corpuscular volume (MCV) (fL)	67.5 (13.0)	62.0 (7.3)
17	Mean corpuscular hemoglobin (MCH) (pg)	20.6 (5.5)	19.1 (2.5)
18	Mean corpuscular hemoglobin concentration (MCHC) (gm/dL)	30.1 (2.8)	30.8 (1.4)
19	Red cell distribution width-coefficient of variation (RDW-CV)%	22.0 (7.8)	17.8 (3.5)
20	Mentzer's index (MI) (MCV/RBC count)	15.8 (5.7)	11.7 (3.1)
21	Platelet count ($\times 10^3$ cells/μL)	309.3 (155.5)	291.5 (63.2)
22	Mean platelet volume (MPV) (fL)	9.8 (8.0)	10.1 (1.3)
23	Platelet distribution width-coefficient of variation (PDW-CV)%	15.6 (0.6)	15.5 (0.5)
24	Hemoglobin A2 (HbA2)%	1.97 (0.49)	4.73 (0.43)
	Sample size (total = 129)	114 (females = 61)	15 (females = 9)

SD: Standard deviation

Table 2: Iron profile of the non-β-TT and β-TT groups

S. no	*Variable*	*Non-β-TT Mean (SD)*	*β-TT Mean (SD)*
1	Serum ferritin (ng/mL)	133.3 (318.0)	103.0 (178.7)
2	Serum iron (µg/dL)	22.7 (39.5)	78.0 (85.9)
3	Unsaturated iron binding capacity (µg/dL)	398.4 (127.9)	291.8 (136.9)
4	Total iron binding capacity (TIBC) (µg/dL)	421.0 (120.7)	369.8 (51.5)
5	Transferrin saturation%	6.0 (8.7)	24.1 (28.1)

Table 3: Regression model 1 to predict HbA2 from hematological (excluding iron profile) parameters selected by LASSO (AIC = 428.93)

Model. no.	*Model*	*Model parameters*					R^2
		Variables	*Regression Coefficients (SE)*	*Regression Coefficients 95% CI*	*t-value*	*p-value*	
1	HbA2 ~age + RBC count+ + MCV + MCHC + RDW-CV	Age	0.492 (0.115)	(0.264,0.720)	4.272	3.83×10^{-5}***	0.40 (model 1)
		RBC count	0.291 (0.104)	(0.085,0.497)	2.801	0.00592**	
		MCV	−0.498 (0.149)	(−0.793, −0.202)	−3.335	0.00113***	
		MCHC	0.241 (0.149)	(−0.054, 0.536)	1.62	0.10782	
		RDW-CV	−0.549 (0.121)	(−0.789, −0.309)	−4.532	1.37×10^{-5}**	

SE: Standard error of the regression coefficient; 95% CI: 95% confidence intervals; R^2: adjusted coefficient of determination; AIC: Akaike information criterion; ***p-value ≤ 0.001; **0.001 < p-value ≤ 0.01.

Table 4: Regression model 2 to predict HbA2 from hematological (including iron profile) parameters selected by LASSO (AIC = 417.03)

Model no.	*Model*	*Model parameters*					R^2
		Variables	*Regression Coefficients (SE)*	*Regression Coefficients 95% CI*	*t-value*	*p-value*	
2	HbA2 ~ age + RBC count + MCV + MCHC + RDW-CV + serum ferritin + transferrin saturation + TIBC	Age	0.413 (0.124)	(0.170, 0.660)	3.349	0.00108**	0.47 (model 2)
		RBC count	0.179 (0.104)	(−0.030, 0.383)	1.730	0.08626	
		MCV	−0.469 (0.145)	(−0.755, −0.182)	−3.234	0.00158**	
		MCHC	0.225 (0.151)	(−0.073, 0.523)	1.493	0.13796	
		RDW-CV	−0.401 (0.123)	(−0.645, −0.157)	−3.255	0.00147**	
		Serum Ferritin	−0.363 (0.132)	(−0.624, −0.102)	−2.757	0.00675**	
		Transferrin saturation	0.093 (0.097)	(−0.100, 0.285)	0.995	0.34158	
		TIBC	−0.403 (0.126)	(−0.653, −0.153)	−3.192	0.00181**	

SE: Standard error of regression coefficient, 95% CI: 95% confidence intervals, R^2: adjusted Coefficient of determination, AIC: Akaike information criterion. **0.001 < p-value ≤ 0.01.

inclusive model, additional predictors viz., serum ferritin, TIBC, and transferrin saturation, were also consistently selected in all folds, suggesting added predictive value from iron profile variables. In contrast, predictors such as Hct, MCV, and MI exhibited moderate selection frequency, while Hb showed low stability when iron parameters were excluded. Overall, inclusion of iron profile parameters improved model performance and enhanced the stability of predictor selection.

The final weighted regression models were constructed on the full dataset for predictor sets with and without iron profile parameters. In the model excluding iron parameters (model 1), LASSO selected five predictors: age, RBC count, MCV, MCHC, and RDW-CV. In the subsequent weighted OLS analysis (Table 3), age and RBC count were positively associated with the outcome (HbA2); MCV and RDW-CV were negatively associated, while MCHC did not have a significant association. This model demonstrated moderate explanatory power (adjusted R^2 = 0.40). Similarly, the model including iron parameters (Model 2) had eight predictors selected by LASSO: age, RBC count, MCV, MCHC, RDW-CV, serum ferritin, TIBC, and transferrin saturation. The subsequent OLS regression analysis results are summarized in Table 4. Age was positively associated with HbA2; MCV, RDW-CV, serum ferritin, and TIBC were negatively associated; no significant association was seen for RBC count, MCHC, and transferrin saturation in this model. Overall, the inclusion of iron profile parameters enhanced both predictive performance (adjusted R^2 = 0.47) and model stability, supporting their added value in the final model.

We show the relationship between measured and model-predicted HbA2 for both models in Figure 1. The scatterplots comparing measured and predicted HbA2 illustrate that the model exhibits modest imprecision in normal/non-β-TT individuals, and it's particularly more pronounced with elevated HbA2. The RMSE between measured and model-predicted HbA2 is 1.014 for Model 1 and 0.928 for Model 2. The ROC analysis curves are shown in Figure 2. The optimal cutoff and the other ROC metrics for the model-derived multivariate regression estimator of HbA2 (model fit) are summarized in Table 5. It can be seen that inclusion of iron and hematology parameters reduced the optimal cutoff for the detection of β-TT.

Table 5: ROC analysis metrics for classification of β-TT (HbA2 > 3.5%) using model fit-based estimator of HbA2

Metric	*Model 1*	*Model 2*
Optimal cutoff (based on Youden's index)	3.32%	3.26%
Sensitivity	76.5%	88.3%
Specificity	82.1%	86.6%
Positive predictive value	40%	50%
Negative predictive value	96%	98%
ROC AUC (95% CI)	0.868 (0.766–0.969)	0.904 (0.829–0.98)

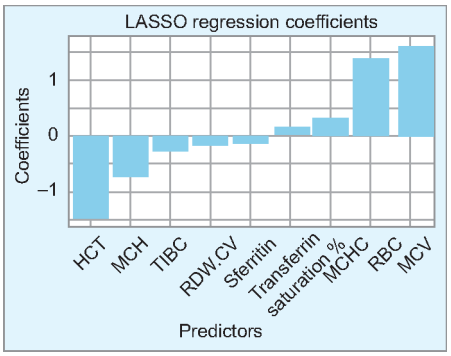

Fig. 1: Scatter plot showing the relationship between measured and model-predicted HbA2 for models 1 (without iron parameters) and 2 (with iron parameters)

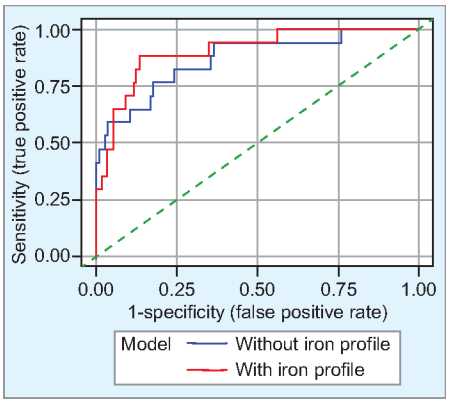

Fig. 2: ROC curves for the final models without and with iron parameters

Discussion

The present study has explored the feasibility of designing a machine learning-based regression model and the relevant variables to predict the HbA2 levels for screening of β-TT using blood-derived indices. In the study cohort, HbA2 was found to be positively associated with age, RBC count, while it was negatively associated with MCV, RDW-CV, serum ferritin, and TIBC.

HbA2, a tetramer of α- and δ-globin chains, is a marker of β-TT. The decrease in β-globin synthesis due to a mutation in one copy of the gene is known to induce compensatory δ-globin overproduction to reduce the toxic effects of α-β imbalance on developing and mature RBC. This accounts for elevated HbA2 in β-TT. Positive association of HbA2 with RBC count reflects compensatory overproduction of RBC, triggered by ineffective erythropoiesis characterized by increased apoptosis of RBC precursors due to the toxic effects of α-globin precipitates.[5,14] Ineffective hemoglobin synthesis delays the cytoplasmic maturation of RBC precursors, while nuclear maturation continues unabated, leading to a greater number of cell divisions before maturation is completed, resulting in smaller-sized RBCs. This explains the low MCV values (Table 1) and negative relationship between HbA2 and MCV seen in the study. Because MCHC is known to be preserved in β-thalassemia trait,[15] the absence of a significant association with HbA2 in our model is consistent with the underlying pathophysiology. The RDW-CV[16] reflects the variability in red blood cell size (anisocytosis). In β-TT, most red cells are uniformly small and hypochromic because the defect is genetic and affects all cells consistently, reducing the size variability. This explains the negative association between HbA2 and RDW-CV. A negative regression coefficient with TIBC and serum ferritin reflects coexisting anemia in the cohort. Several studies found that β-TT is associated with low MI[4,6,11,15,17,18] characterized by high RBC and low MCV. Our study found a positive association between HbA2 and RBC counts and a negative association with MCV, although MI was not selected by LASSO to be included in the model. Patients have higher RBC counts with β-TT. Our RDW-CV Results are concordant with other studies, which also found similarly diminished RDW-CV in β-TT. The negative association between HbA2 and TIBC found in our study is consistent with the results of other studies.[6,19] We did not find any significant association between HbA2 and serum iron and transferrin saturation. Although HbA2 is physiologically stable beyond infancy, the regression model showed a positive association between age and HbA2. This likely reflects a statistical rather than biological relationship. Age may be acting as a proxy for unmeasured factors such as referral patterns, comorbidities, or cohort differences in laboratory testing, rather than exerting a direct effect on HbA2. Such associations are well-recognized in multivariable models when predictors share correlations or when residual confounding is present. Moreover, the average age of β-TT patients is higher than that of the non-β-TT patients, explaining the positive association with HbA2. Therefore, the observed age coefficient should be interpreted as a modeling and sampling-related artifact rather than a true physiological effect.

Several studies have attempted to devise a binary classification type of machine learning[10,20–23] to distinguish the β-TT group from the non-β-TT group based on a fixed cutoff value for HbA2. The present study and other studies discussed previously have shown the dependence of HbA2 on various blood indices, which are also sensitive to other conditions, like concurrent iron deficiency. ROC analysis in our study also has found a lower cutoff compared to the conventional HbA2 threshold for β-TT detection. Therefore, the use of a fixed cutoff value renders the models biased and unreliable in the assessment of β-TT. By using regression-based machine learning models, the present study offers a novel and cost-effective adjunct to the determination of HbA2 that does not suffer from the pitfalls of using a fixed cutoff value.

The study has the following limitations. A key limitation of this study was the presence of missing data, which may have influenced the completeness of the analyses and the generalizability of the findings. A few predictors had substantial missingness (3 had ~40%, one had 53.5%), but most had <15% missing data, because certain laboratory tests were not uniformly ordered. Although multiple imputation was used to retain sample size (by estimating missing data) and reduce bias, imputed values cannot fully replace observed measurements, and some residual uncertainty may remain. The average age of the β-TT patients is higher than that of the control group due to the liberal selection criteria and the observational nature of the study. The number of β-TT cases in the dataset was small, resulting in a limited predictor-to-event ratio. This tends to reduce model stability and increase the risk of overfitting. To mitigate this, we used LASSO methods for variable selection and evaluated model performance using 5-fold cross-validation. However, we acknowledge that the small number of β-TT cases remains a key limitation of the study and contributes to the reduced predictive accuracy that was more pronounced in β-TT patients, as shown in Figure 1. Larger β-TT cohorts will enable more precise and accurate model estimates of HbA2. The two-stage LASSO-then-OLS may introduce bias, as the multiple

regression model in the second stage tends to remove the shrinkage applied with the penalization parameter in the first stage (LASSO) for variable selection. However, unpenalized coefficients (from OLS) are more interpretable and hence clinically useful, while fully penalized coefficients are known to over-shrink clinically meaningful effects, reducing interpretability without improving predictive accuracy. LASSO-then-OLS, therefore, represents a pragmatic balance between regularization and interpretability. The study did not explicitly include or model the rich cytomorphological features in the peripheral blood smear relevant to the β-TT that could be extracted by the digital image analysis of the peripheral blood smear. The scatterplots comparing measured and predicted HbA2 illustrate that the model exhibits modest imprecision, with predictions deviating from the identity line, particularly at higher HbA2 values. Clinically, this level of imprecision is less likely to affect interpretation for individuals with clearly normal range values. The model has more pronounced imprecision in the β-TT range. In this range, even small prediction errors could shift a value across the cutoff, potentially leading to misclassification or unnecessary confirmatory testing. Thus, while the model may support preliminary screening or triage, definitive diagnostic decisions should continue to rely on direct HbA2 measurement. Future work with larger datasets and more comprehensive predictors may help reduce uncertainty around threshold-adjacent values.

Conclusion

The present study has explored the feasibility of devising a regression-based machine learning model to estimate HbA2 from various blood-derived metrics, including RBC indices and iron profile. Such a model could be developed as a mass-screening/triage tool in the context of India, where the use of a fixed cutoff value for HbA2 is untenable in the background of pervasive IDA. However, the best model was able to explain only 47% of the overall variance, and the predictive performance deteriorated at the clinically critical higher HbA2 values. Nevertheless, the model was able to identify some of the relevant variables needed to develop such a predictive model. Further development of the model, particularly through expansion of the β-TT cohort and training on a larger dataset, will be essential to improve predictive accuracy. Incorporating an expanded feature set, including quantitative digital image analysis of peripheral blood smears, may also enhance model performance. These steps are necessary before the model can be considered sufficiently robust for clinical deployment.

Acknowledgments

The authors would like to thank Dr. Asma Niloufer, Senior Consultant, Department of Biochemistry, Apollo Hospitals, Hyderabad, for her support. The authors also thank Dr. N Balakrishna for statistical support in the design of the study.

Ethical Approval

The study was approved by the Institutional Ethics Committee, Apollo Institute of Medical Sciences and Research (AIMSR), Hyderabad (approval no. EC/NEW/INST/1527/2025/12/010, dated 23 December 2025).

Source of Funding

None.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflict of Interest

None.

Supplementary Material

The supplementary Table S1 is available online on the journal website.

Table S1: Missing data in the study variables

Orcid

Ramadevi Peraka https://orcid.org/0000-0002-5567-143X
Pradeep Kumar Reddy M https://orcid.org/0000-0002-4190-1927
Vanajakshi S https://orcid.org/0000-0001-8423-9163
Sumedha Avadhanam https://orcid.org/0009-0007-4523-0343
Manimekhala Parsa https://orcid.org/0000-0003-0160-4394
Aditya Koppula https://orcid.org/0000-0002-0978-4142

References

1. Sumedha D, Anita K. Prevalence of beta thalassemia carriers in India: a systematic review and meta-analysis. J Community Genet 2023;14(6):527–541.
2. Tuo Y, Li Y, Li Y, et al. Global, regional, and national burden of thalassemia, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. EClinicalMedicine 2024;72:102619.
3. Baliyan M, Kumar M, Nangia A, et al. Can RBC indices be used as screening test for beta-thalassemia in Indian antenatal women? J Obstet Gynaecol India 2019;69(6):495–500.
4. Tabassum S, Khakwani M, Fayyaz A, et al. Role of Mentzer index for differentiating iron deficiency anemia and beta thalassemia trait in pregnant women. Pak J Med Sci 2022;38(4 Pt 2):878–882.
5. Budania P, Patidar AK, Loyal K, et al. Reliability of different RBC indices to differentiate between beta thalassemia trait and iron deficiency anemia in children. Int J Life Sci Biotechnol Pharma Res 2023;12(2):837–843.
6. Jasim RH, Aouda RH. Evaluation of RBC indices and significance of Mentzer index for differentiation between iron deficiency anemia and beta thalassemia trait. Al-Mustaqbal J Pharm Med Sci 2025;3(1):1.
7. Ministry of Health and Family Welfare, Government of India. Anemia Mukt Bharat: Operational Guidelines for Programme Managers—Intensified National Iron Plus Initiative (I-NIPI). New Delhi: Ministry of Health and Family Welfare, Government of India; 2018.
8. Bhargava M, Kumar V, Pandey H, et al. 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.
9. Kamath SA, Shah SN, Munjal YP, et al. API Textbook of Medicine, 12th edition. New Delhi: Jaypee Brothers Medical Publishers; 2022.
10. Jahan A, Singh G, Gupta R, et al. Role of red cell indices in screening for beta thalassemia trait: an assessment of the individual indices and application of machine learning algorithm. Indian J Hematol Blood Transfus 2021;37(3):453–457.
11. Bhattacharya S, Sinha S, Chahal KT, et al. Role of Mentzer index in diagnostic dilemma of thalassemia trait viz a viz iron deficiency anaemia in a teaching hospital. Int J Pharm Clin Res 2024;16(1):659–664.
12. Bordbar E, Taghipour M, Zucconi BE, et al. Reliability of different RBC indices and formulas in discriminating between β-thalassemia minor and other microcytic hypochromic cases. Mediterr J Hematol Infect Dis 2015;7(1):e2015022.
13. Thomas C, Thomas L. Biochemical markers and hematologic indices in the diagnosis of functional iron deficiency. Clin Chem 2002;48(7):1066–1076.
14. Thein SL. The molecular basis of β-thalassemia. Cold Spring Harb Perspect Med 2013;3(5):a011700.
15. Shankar G, Mohan A, Jain M, et al. Mentzer index as a reliable and economical tool for screening of beta-thalassemia trait. Indian J Public Health Res Dev 2025;16(3):96–102.
16. Vehapoglu A, Ozgurhan G, Demir AD, et al. Hematological indices for differential diagnosis of beta thalassemia trait and iron deficiency anemia. Anemia 2014;2014:576738.
17. AlQarni AM, Althumairi A, Alkaltham NK, et al. Diagnostic test performance of the Mentzer index in evaluating Saudi children with microcytosis. Front Med (Lausanne) 2024;11:1361805.
18. Bose S, Maimoon S. Is Mentzer index a reliable diagnostic screening tool for beta thalassemia trait? IOSR J Dent Med Sci 2018;17(7):7–11.
19. Singh V, Baranwal AK, Biswas AK, et al. Prevalence and assessment of the impact of iron-deficiency anemia in beta-thalassemia trait subjects: a study from a tertiary care center of Western India. Asian J Transfus Sci 2025;19(1):48–52.
20. Kabootarizadeh L, Jamshidnezhad A, Koohmareh Z, et al. Differential diagnosis of iron-deficiency anemia from β-thalassemia trait using an intelligent model in comparison with discriminant indexes. Acta Inform Med 2019;27(2):78–84.
21. Barnhart-Magen G, Gotlib V, et al. Differential diagnostics of thalassemia minor by artificial neural networks model. J Clin Lab Anal 2013;27(6):481–486.
22. Uçucu S, Azik F. Artificial intelligence-driven diagnosis of β-thalassemia minor and iron deficiency anemia using machine learning models. J Med Biochem 2024;43(1):11–18.
23. Saleem M, Aslam W, Lali MIU, et al. Predicting thalassemia using feature selection techniques: a comparative analysis. Diagnostics (Basel) 2023;13(22):3441.