Machine learning-based prediction of hemoglobinopathies using complete blood count data

Schipper, A.; Rutten, M.; Gammeren, A. van; Harteveld, C.L.; Urrechaga, E.; Weerkamp, F.; ... ; Kurstjens, S.

Clinical Chemistry 70:8
1064–1075 (2024)

Machine Learning-Based Prediction of Hemoglobinopathies Using Complete Blood Count Data

Anoeska Schipper,[a,b] Matthieu Rutten,[b,c] Adriaan van Gammeren,[d] Cornelis L. Harteveld,[e] Eloísa Urrechaga,[f] Floor Weerkamp,[g] Gijs den Besten,[h] Johannes Krabbe,[i] Jennichjen Slomp,[i] Lise Schoonen,[g,j] Maarten Broeren,[k] Merel van Wijnen,[l] Mirelle J.A.J. Huijskens,[m] Tamara Koopmann,[e] Bram van Ginneken,[b] Ron Kusters ,[a,n,†] and Steef Kurstjens[a,*,†]

BACKGROUND: Hemoglobinopathies, the most common inherited blood disorder, are frequently underdiagnosed. Early identification of carriers is important for genetic counseling of couples at risk. The aim of this study was to develop and validate a novel machine learning model on a multicenter data set, covering a wide spectrum of hemoglobinopathies based on routine complete blood count (CBC) testing.

METHODS: Hemoglobinopathy test results from 10 322 adults were extracted retrospectively from 8 Dutch laboratories. eXtreme Gradient Boosting (XGB) and logistic regression models were developed to differentiate negative from positive hemoglobinopathy cases, using 7 routine CBC parameters. External validation was conducted on a data set from an independent Dutch laboratory, with an additional external validation on a Spanish data set (n = 2629) specifically for differentiating thalassemia from iron deficiency anemia (IDA).

RESULTS: The XGB and logistic regression models achieved an area under the receiver operating characteristic (AUROC) of 0.88 and 0.84, respectively, in distinguishing negative from positive hemoglobinopathy cases in the independent external validation set. Subclass analysis showed that the XGB model reached an AUROC of 0.97 for β-thalassemia, 0.98 for α^0-thalassemia, 0.95 for homozygous α^+-thalassemia, 0.78 for heterozygous α^+-thalassemia, and 0.94 for the structural hemoglobin variants Hemoglobin C, Hemoglobin D, Hemoglobin E. Both models attained AUROCs of 0.95 in differentiating IDA from thalassemia.

CONCLUSIONS: Both the XGB and logistic regression model demonstrate high accuracy in predicting a broad range of hemoglobinopathies and are effective in differentiating hemoglobinopathies from IDA. Integration of these models into the laboratory information system facilitates automated hemoglobinopathy detection using routine CBC parameters.

Introduction

Normal adult hemoglobin (Hb A) is a tetramer composed of 2 α and 2 β globin chains, which facilitates oxygen transport via a reversible binding mechanism. Hemoglobinopathies stem from mutations or deletions affecting α-globin (*HBA1*, *HBA2*) genes, β-globin (*HBB*) genes, γ-globin (*HBG1*, *HBG2*) genes, and δ-globin *(HBD)* genes. Structural variants of hemoglobin typically result from qualitative changes such as amino acid substitutions, whereas quantitative alterations, including gene deletions or mutations, manifest as thalassemias that reduce globin chain production (1, 2).

[a]Laboratory of Clinical Chemistry and Hematology, Jeroen Bosch Hospital's, Hertogenbosch, the Netherlands; [b]Diagnostic Image Analysis Group, Radboudumc, Nijmegen, the Netherlands; [c]Department of Radiology, Jeroen Bosch Hospital's, Hertogenbosch, the Netherlands; [d]Laboratory of Clinical Chemistry and Laboratory Medicine, Amphia Hospital, Breda, the Netherlands; [e]Department of Clinical Genetics, Laboratory for Genome Diagnostics, Leiden University Medical Center, Leiden, the Netherlands; [f]Laboratory of Hematology, Hospital Universitario Galdakao Usansolo, Galdakao, Spain; [g]Laboratory of Clinical Chemistry, Maasstad Hospital, Rotterdam, the Netherlands; [h]Laboratory of Clinical Chemistry and Laboratory Medicine, Isala Hospital, Zwolle, the Netherlands; [i]Laboratory of Clinical Chemistry and Hematology, Medisch Spectrum Twente/Medlon BV, Enschede, the Netherlands; [j]Laboratory of Clinical Chemistry and Laboratory Medicine, Canisius Wilhelmina Hospital, Nijmegen, the Netherlands; [k]Laboratory of Clinical Chemistry and Laboratory Medicine, Máxima Medical Center, Eindhoven, the Netherlands; [l]Laboratory of Clinical Chemistry and Laboratory Medicine, Meander Medical Center, Amersfoort, the Netherlands; [m]Department of Clinical Chemistry and Haematology, Zuyderland Medical Center, Sittard/Heerlen, the Netherlands; [n]Department of Health Technology and Services Research, Technical Medical Centre, University of Twente, Enschede, the Netherlands.
*Address correspondence to this author at: Laboratory of Clinical Chemistry and Hematology, Jeroen Bosch Hospital, Henri Dunantstraat 1, 5223 GZ 's Hertogenbosch, the Netherlands. Tel +031(0) 0626281521; e-mail steef_kurstjens@hotmail.com.
[†]Joint last authors.
Received January 12, 2024; accepted May 13, 2024.
https://doi.org/10.1093/clinchem/hvae081

The clinical presentations of the conditions vary widely, ranging from non-anemic states or mild microcytic hypochromic anemia up to severe anemia requiring regular blood transfusions (3, 4).

According to the World Health Organization (WHO), hemoglobinopathies affect over 5.2% of the global population, and account for 3.4% of deaths in children below the age of 5 (5, 6). The prevalence is higher in malaria-endemic regions but is rising in traditionally non-endemic regions, such as Northern Europe and North America, due to increased migration (7–10). Many hemoglobinopathy carriers remain undiagnosed due to the low awareness of physicians in non-endemic regions, lack of symptoms in carriers, and limited access to advanced laboratory testing in developing countries (11, 12). Carrier detection, however, is crucial to identify couples at risk of having children affected with severe forms of hemoglobinopathy (13). Early diagnosis, which can be facilitated through screening and genetic counseling, is crucial in preparing at-risk couples.

Specific changes in complete blood count (CBC) parameters are used by laboratory specialists to recommend diagnostic testing for hemoglobinopathy to clinicians. These include reduced levels of hemoglobin (Hb) and mean corpuscular volume (MCV). The CBC changes seen in hemoglobinopathies can often mimic those of iron deficiency anemia (IDA), which can result in unnecessary iron supplementation.

To differentiate between hemoglobinopathies and IDA, various machine learning (ML) models or rule-based formulas, with varying degrees of accuracy, have been developed using CBC results to predict the risk of a potential hemoglobinopathy. The majority of these formulas specifically differentiate one distinct thalassemia (α or β) from IDA (e.g., 14–19). These formulas were not designed to differentiate structural hemoglobin variants, combinations of multiple hemoglobinopathies, or cases with a hemoglobinopathy and concomitant iron deficiency (20, 21).

The aim of this study was to develop and validate, using an extensive multicenter data set, a novel ML model and a logistic regression model to detect a broad spectrum of hemoglobinopathies that can be applied using routine CBC testing results.

Materials and Methods

Anonymized hemoglobinopathy diagnostic test results from 10 322 adults were extracted retrospectively over 12 years (2011 to 2022) from the laboratory information systems of 8 Dutch laboratories: Amphia Hospital (n = 1559), Isala Hospital (n = 1321), Jeroen Bosch Hospital (n = 927), Maasstad Hospital (n = 894), Máxima Medical Center (n = 719), Meander Medical Center (n = 984), Medlon BV (Medlon) (n = 3160), and Zuyderland Medical Center (n = 758). Along with the hemoglobinopathy test results, the data comprised CBC parameters including Hb, MCV, mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), red cell distribution width - coefficient of variation (RDWCV), red blood cell (RBC) count, and platelets (Plt).

DIAGNOSTIC METHODOLOGY

Analyzers used for measuring the hematological parameters and the hemoglobinopathy diagnostics vary according to laboratory (online Supplemental Table 1). Hemoglobinopathy diagnostics encompass the diagnostic analysis of structural hemoglobin variants and β-thalassemia using either high-performance liquid chromatography (HPLC) or capillary electrophoresis (CE). The DNA analysis of α-thalassemia was performed using either gene-associated polymorphism PCR (GAP-PCR) or strip PCR or multiplex ligation-dependent probe amplification (MLPA). In instances where no mutation was detected, yet clinical suspicion persisted, whole exome sequencing was conducted in Leiden University Medical Center (UMC) upon request. CBC parameters were measured using either an Advia 2120i (Siemens Healthineers) or Sysmex XN-9000 (Sysmex Corporation) (Supplemental Table 1). An overview of all CBC data (n = 8564) per laboratory is presented in online Supplemental Table 2. Differences in CBC parameters between laboratories can be accounted for due to large differences in hemoglobinopathy prevalence in the data sets across laboratories, ranging from 29.9% to 49.9%. As the data from the Jeroen Bosch Hospital and Maasstad Hospital contained no negative cases (Fig. 1), these data sets had a 100% prevalence.

MODEL DEVELOPMENT

The CBC parameters were used as input for the models. The clinical interpretation of the hemoglobinopathy test results by the specialist in laboratory medicine or the leading expert of Leiden UMC served as the ground truth for the model, with the primary objective of learning to distinguish positive from negative cases, maximizing clinical utility. Pathology-specific classifiers were trained on α-thalassemia, β-thalassemia, and combinations with the aim to differentiate subtypes from negative cases for comparison with the original positive–negative hemoglobinopathy model. However, the results indicated no substantial improvement of the pathology-specific models compared to the original positive–negative hemoglobinopathy classification model (online Supplemental Fig. 1). Since incorporating sex as a parameter did not affect the models' performance, it was excluded from our model.

Data from 7 out of 8 laboratories were used for model development. Cases that received blood

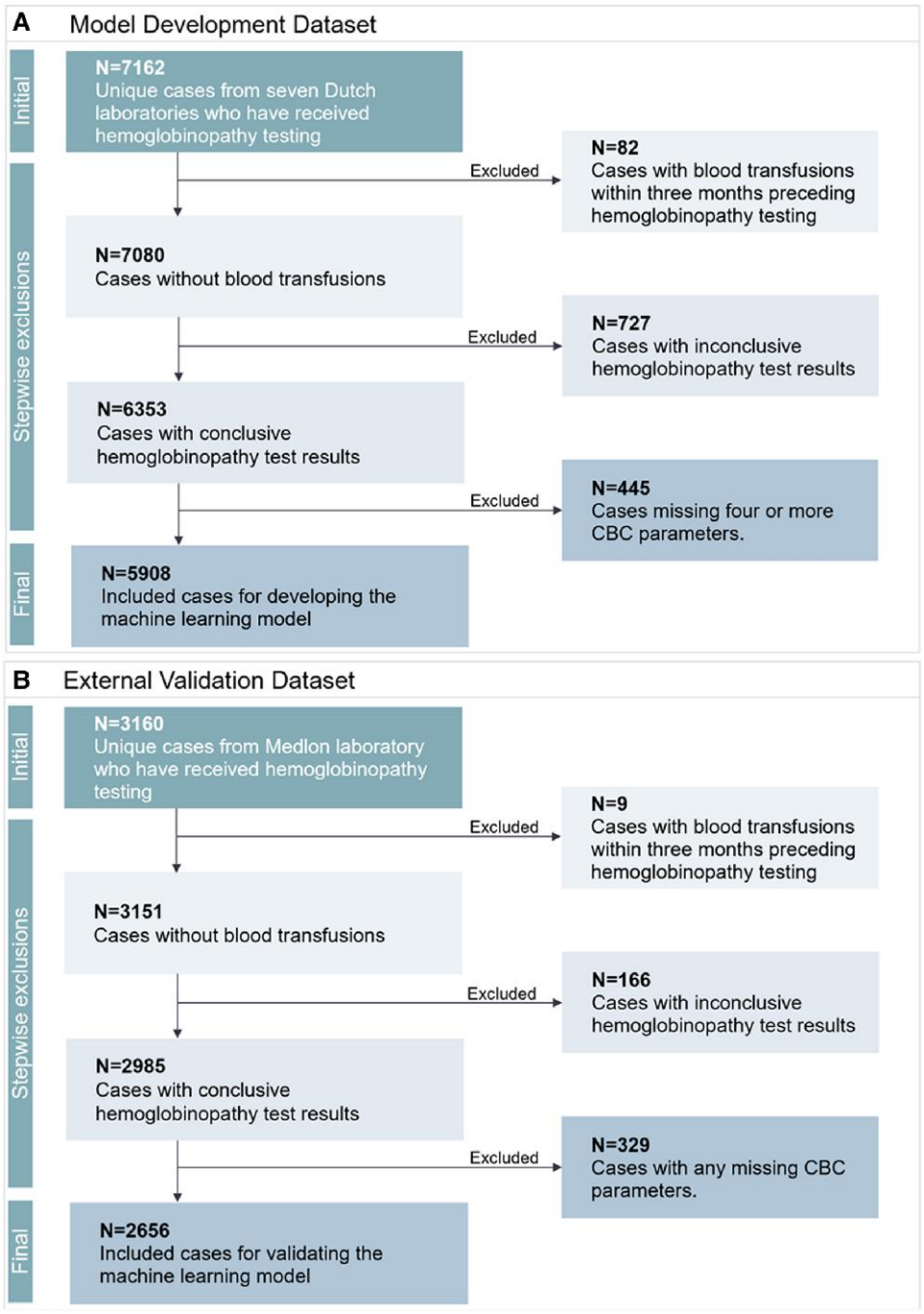

Fig. 1. Inclusions and exclusions in the model development and external validation data set. (A), The model development data set consisted of laboratories: Amphia Hospital (n = 1559), Isala Hospital (n = 1321), Jeroen Bosch Hospital (n = 927), Maasstad Hospital (n = 894), Máxima Medical Center (n = 719), Meander Medical Center (n = 984), and Zuyderland Medical Center (n = 758); (B), Medlon (n = 3160) was reserved for external validation. After exclusions, the development data set included 5908 cases, and the external validation data set comprised 2656 cases. Color figure available at https://academic.oup.com/clinchem.

transfusions in the preceding 3 months or with inconclusive hemoglobinopathy test results were excluded (Fig. 1A). Jeroen Bosch Hospital and Maasstad Hospital employ a restricted approach for hemoglobinopathy diagnostics, primarily relying on an MCV (<80 fL) cutoff to decide whether to perform α-thalassemia diagnostics. As this leads to biased data, cases from Jeroen Bosch Hospital and Maasstad Hospital with a negative test result were excluded. Cases missing 4 or more CBC parameters were also excluded (n = 445). Cases missing 3 or less CBC parameters were imputed using iterative imputation. This threshold was chosen to balance preservation of enough cases from each laboratory while minimizing missing parameters for imputation, resulting in a final model development data set of 5908 cases ($n_{positive}$ = 3249, $n_{negative}$ = 2659). The integrity of the imputation process was validated through pre- and post-imputation distribution comparisons, conducted via kernel density estimation plots, mean and interquartile range assessments, and individual case evaluations. Eighty percent of the data set (n = 4726) was used for training and tuning using repeated stratified 10-fold cross-validation, with the aims of preserving the same class distribution across the folds and repeating the cross-validation procedure multiple times using the mean performance result for tuning the models (22). Twenty percent was reserved for internal validation of the model's performance (n = 1182).

The XGB model was trained using an eXtreme Gradient Boosting algorithm with the XGBoost package (1.7.3). The model was trained to maximize the area under the receiver operating characteristic (AUROC). During training a learning rate (eta) of 0.1, 1000 boosting rounds and 20 early-stopping rounds and a tree-depth of 6 were used. Analyses were carried out using Python 3.8 (Python Software Foundation), with the packages Numpy (1.20), Pandas (0.28), Sklearn (1.1.1), XGBoost (0.24), and SHAP (0.41). A logistic regression-based model was developed using Sklearn (1.1.1). The formula of the logistic regression for calculating the hemoglobinopathy probability = 1/(1 + e^(−(−2.73516742 + −0.139101127*MCV [fL] + −0.133152018*RDWCV [%] + 3.38704562*MCH[fmol] + −0.998798*Hb[mmol/L] + 0.488415669*MCHC[mmol/L] + 1.74713625*RBC [count × 10^{12}/L] + 0.000129845*Plt[count × 10^{9}/L]))). To use the formula with traditional concentration units, the CBC data must be converted to SI units. As this study used a multivariable XGB model, the Transparant Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) checklist was used for increased transparency of the methodology (online Supplemental Table 3). XGB and logistic regression models are made publicly available via GitHub: doi:https://github.com/aschipper/hemoglobinopathies-AI and via Figshare: https://doi.org/10.6084/m9.figshare.25765302.

CLASSIFICATION OF HEMOGLOBINOPATHIES

In total, the data set comprised 106 distinct hemoglobinopathies, systematically categorized into 4 subclasses for the distinct evaluation of the models' discriminative performance against all negative cases in the external validation set: (a) α-thalassemia, (b) β-thalassemia, (c) structural hemoglobin variants, and (d) combinations of thalassemia and concomitant structural hemoglobin variants. α-thalassemia was subclassified as: hemoglobin H (Hb H) disease, α^0-heterozygote, α^+-homozygote, compound heterozygote, and α^+-heterozygote. β-thalassemia was subclassified as β-thalassemia and β−δ-thalassemia. Structural hemoglobin variants were subclassified as: hemoglobin E (Hb E) homozygote/heterozygote, hemoglobin C (Hb C) homozygote/heterozygote, hemoglobin D (Hb D) homozygote/heterozygote, hemoglobin S (Hb S) heterozygote, and sickle cell anemia. Sickle cell anemia encompasses Hb S homozygote, and compound heterozygote Hb S/Hb C, Hb S/β-thalassemia, Hb S/Hb E, and Hb S/Hb D. The "combinations" category predominantly comprised cases where α-thalassemia was present in combination with a structural hemoglobin variant.

MODEL VALIDATION

The complete data set from the Dutch laboratory Medlon was reserved for independent external validation (n = 2656, $n_{positive}$ = 1004, $n_{negative}$ = 1652) of the XGB and logistic regression models (Fig. 1B). Medlon carries out a comprehensive diagnostic evaluation for each hemoglobinopathy diagnostic request, which encompasses Sanger sequencing of both α and β globin genes. Cases with any missing CBC parameters were excluded (n = 329) (Fig. 1B).

THALASSEMIA VS IDA

The effectiveness of the XGB and logistic regression models in differentiating between thalassemia and IDA was evaluated on a previously published patient data set from Galdakao-Usansolo Hospital in Spain, which focusses on cases with microcytic anemia (14). CBC parameters were measured using Abbott Sapphire, Siemens Advia, and Beckman Coulter LH750 and LH780. IDA was considered present when serum ferritin was <15 μg/L, and/or transferrin saturation was <20%. The data set consisted of 2629 cases ($n_{thalassemia}$ = 1370, n_{IDA} = 1259) over a period of 9 years (2007 to 2015). As platelet values were missing in 1832 cases, these values were imputed using Sklearn's iterative imputer.

SCREENING FOR HEMOGLOBINOPATHIES IN CBC RESULTS

The use of the algorithms as a screening tool on all CBC requests in a healthcare setting was assessed on a data set

Table 1. Overview of CBC parameters presented as medians and ±IQR for each hemoglobinopathy (n = 8564).

Category	n	Hb[a]	MCH[a]	MCHC[a]	MCV[a]	Plt[a]	RBC[a]	RDWCV[a]
Negative	**4311**	**11.8 (10.2–13.5)**	**28.0 (24.2–29.0)**	**32.4 (31.1–33.5)**	**85 (79–90)**	**267 (221–325)**	**4.4 (3.9–4.8)**	**13.9 (12.9–16.1)**
Positive	**4253**	**11.8 (10.5–13.1)**	**23.4 (19.3–25.8)[b]**	**31.9 (30.9–33.0)[b]**	**74 (66–80)[b]**	**267 (220–320)**	**5.1 (4.5–5.7)[b]**	**15.3 (14.0–17.0)[b]**
α-thalassemia								
α^+-heterozygote	1101	11.9 (10.6–13.1)	25.3 (22.6–25.8)[b]	31.9 (30.9–32.7)[b]	79 (75–82)[b]	272 (223–324)	4.8 (4.3–5.2)[b]	14.5 (13.6–16.1)[b]
α^+-homozygote/compound heterozygote	339	11.8 (11.0–12.9)	22.5 (20.9–22.6)[b]	31.1 (30.3–31.7)[b]	72 (70–75)[b]	247 (213–296)	5.3 (4.8–5.7)[b]	15.0 (14.4–16.0)[b]
α^0-heterozygote	185	11.9 (11.1–12.7)	21.2 (19.3–20.9)[b]	31.1 (30.5–31.9)[b]	68 (66–70)[b]	269 (220–311)	5.6 (5.3–5.9)[b]	15.6 (14.8–16.7)[b]
Hemoglobin H disease	20	10.6 (9.0–13.2)	19.9 (16.1–22.6)[b]	29.7 (27.7–32.1)[b]	67 (60–73)[b]	270 (213–300)	5.6 (4.8–6.2)[b]	21.8 (20.7–24.4)[b]
β-thalassemia								
β-thalassemia	1192	11.4 (10.5–12.6)[b]	20.1 (17.7–20.9)[b]	31.4 (30.6–32.2)[b]	64 (62–68)[b]	262 (218–308)	5.7 (5.1–6.2)[b]	16.3 (15.4–17.7)[b]
β-δ-thalassemia	18	12.7 (11.6–13.2)	24.0 (20.9–25.8)[b]	31.9 (30.6–33.2)	71 (66–79)[b]	243 (205–294)	5.3 (4.7–5.9)[b]	16.3 (15.4–18.0)[b]
Structural hemoglobin variants								
Sickle cell anemia	120	9.2 (8.2–10.5)[b]	31.3 (25.8–32.2)[b]	34.6 (32.7–35.6)[b]	92 (81–98)[b]	427 (338–597)[b]	3.1 (2.5–3.7)[b]	18.3 (16.8–20.9)[b]
Hb S heterozygote	481	12.6 (10.8–14.0)[b]	28.2 (25.8–29.0)	33.4 (32.4–34.3)[b]	83 (78–88)[b]	266 (221–333)	4.6 (4.0–5.0)[b]	14.0 (13.2–15.5)
Hb E homozygote/heterozygote	203	12.3 (11.0–13.4)[b]	24.7 (22.6–24.2)[b]	32.7 (31.9–33.4)[b]	75 (72–78)[b]	293 (241–340)[b]	5.0 (4.5–5.5)[b]	14.3 (13.7–14.9)[b]
Hb C homozygote/heterozygote	133	12.4 (11.0–13.5)[b]	26.8 (24.2–27.4)	35.2 (33.8–35.9)[b]	78 (74–81)[b]	281 (234–329)	4.6 (4.1–5.1)[b]	14.8 (13.9–16.8)[b]

Continued

Table 1. (continued)

Category	n	Hb[a]	MCH[a]	MCHC[a]	MCV[a]	Plt[a]	RBC[a]	RDWCV[a]
Hb D homozygote/heterozygote	32	12.6 (11.0–14.7)[b]	27.6 (24.2–29.0)	33.5 (32.5–34.0)[b]	84 (79–87)	273 (220–309)	4.9 (4.3–5.0)[b]	13.5 (12.7–15.3)
Combinations								
Combinations	412	12.1 (10.8–13.4)[b]	24.0 (20.9–25.8)[b]	32.5 (31.4–33.7)[b]	74 (68–79)[b]	251 (204–296)[b]	5.1 (4.5–5.6)[b]	15.0 (13.9–16.5)[b]

The bold values in the table refer to inputs belonging to the positive and negative classes on which the models are developed.
[a]Hb, hemoglobin [g/dL] (SI conversion factor [g/dL to mmol/L]: 0.6206); MCH, mean corpuscular hemoglobin [pg] (SI conversion factor [pg to fmol]: 0.0621); MCHC, mean corpuscular hemoglobin concentration [g/dL] (SI conversion factor [g/dL to mmol/L]: 0.6206); MCV, mean corpuscular volume [fL]; Plt, platelet count [10^3/μL] (SI conversion factor [10^3/μL to count ×10^9/L]: 1.0); RBC, red blood cell [10^6/μL] (SI conversion factor [10^6/μL to count ×10^{12}/L]: 1.0); RDWCV, red cell distribution width - coefficient of variation [%].
[b]Median value differed significantly between positive and negative cases.

from the Jeroen Bosch Hospital. The data set consisted of all CBC results of January 2023, a total of 20 870 CBC results. Sixty-eight positive cases from 2022 to 2023 were artificially added to enrich the data set with positive cases, due to the low prevalence of hemoglobinopathies in the region. These were all positive cases, without any preselection (ensuring no selection bias), that underwent hemoglobinopathy diagnostics at Jeroen Bosch Hospital between April 2022 and June 2023 and were not part of the original model development data set. This resulted in a final prevalence of 0.59% in the data set. CBC results from children or cases with incomplete parameters were excluded (n = 969), resulting in a final data set of 19 969 cases ($n_{positive} = 118$, $n_{negative} = 19\,851$). "Negative" cases were classified as cases that had not been diagnosed with a hemoglobinopathy in the Jeroen Bosch Hospital. Therefore, the category of "negative" cases will contain an unknown (limited) number of false-negative cases. Additionally, the area under the precision-recall curves (AUPRCs) and Matthews correlation coefficients (MCCs) were generated to account for the large class imbalance in the data set and assess real-world utility scenarios.

STATISTICS

CBC parameters are presented as medians and interquartile ranges (±IQR) (Table 1 and online Supplemental Tables 2 and 4). Kruskal–Wallis tests were used to assess significant differences in medians between each hemoglobinopathy category and cases without a hemoglobinopathy. A value of $P < 0.05$ was considered statistically significant.

Confidence intervals (CI) for the AUROC and AUPRC values were established through cross-validation, employing 10-folds with 50 repeats. The 2.5th and 97.5th percentiles of the ranked list of 500 AUROC and AUPRC values were computed to determine the CI. The MCC was derived by identifying the optimal threshold that maximized MCC.

Results

Positive hemoglobinopathy cases exhibited distinct patterns compared to negative cases. These patterns included significantly lower MCV, MCH, and MCHC, and notably higher RDWCV and RBC levels than negative cases (Table 1). As Hb and RBC counts are sex-specific, they are presented as separate values for each sex (Supplemental Table 4).

MODEL PERFORMANCE

The primary objective of the models developed in this study was to differentiate negative cases from positive hemoglobinopathy cases based on routine CBC parameters. The internal validation of the XGB model achieved an AUROC of 0.90 (±0.01) compared to an

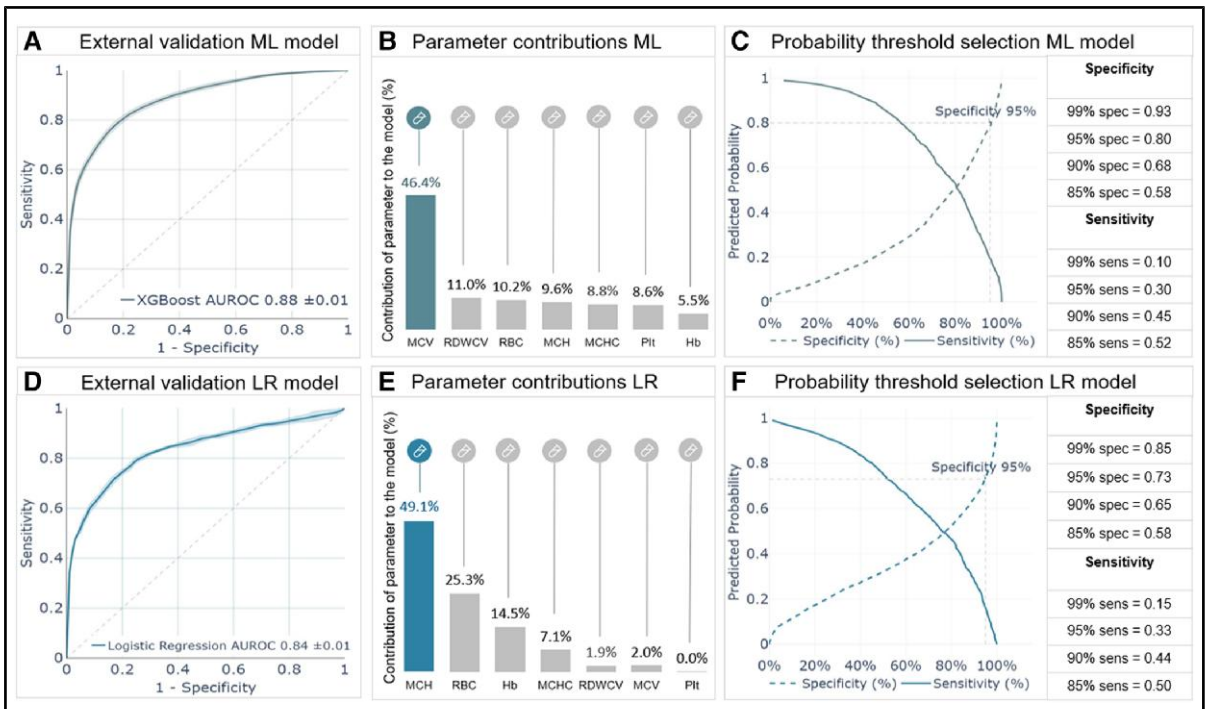

Fig. 2. Receiver operating characteristic plots of (A) the XGBoost (XGB) and (D) logistic regression (LR) model of the independent external validation set. Contributions of each individual parameter to the XGB model are SHapley Additive exPlanations (SHAP) values (B) and logistic regression coefficients (E) scaled and plotted as percentage contributions to the prediction. Plot with different probability thresholds based on prioritizing specificity or sensitivity for the (C) XGB and (F) logistic regression model. Abbreviations: MCV, mean corpuscular volume; RDWCV, red cell distribution width - coefficient of variation; MCH, mean corpuscular hemoglobin; Hb, hemoglobin; MCHC, mean corpuscular hemoglobin concentration; RBC, red blood cell count; Plt, platelet count. Color figure available at https://academic.oup.com/clinchem.

AUROC of 0.86 (±0.01) of the logistic regression (online Supplemental Fig. 2). On the independent external validation set, the XGB model achieved an AUROC of 0.88 (±0.01) and the logistic regression an AUROC of 0.84 (±0.01) (Fig. 2A and D). The contribution of each parameter to the models is presented in Fig. 2B and E, including the feature contributions to 2 random example cases, one negative and positive (online Supplemental Fig. 3). Both models generate a quantitative probability ranging from 0 to 1 that can be converted to a qualitative prediction of the absence or presence of hemoglobinopathy using a threshold prioritizing either specificity or sensitivity (Fig. 2C and F). A probability threshold of ≥0.80 for the XGB model and a probability threshold of ≥0.73 for the logistic regression are associated with a 95% specificity cutoff.

SUBCLASS ANALYSIS OF THE MODEL ON SPECIFIC HEMOGLOBINOPATHIES

In a subclassification analysis of α-thalassemia, the XGB model achieved AUROC scores of 0.98 for α^0-heterozygote, 0.95 for α^+-homo- and compound heterozygote, and 0.78 for α^+-heterozygote (Fig. 3A). In the case of β-thalassemia, the XGB model attained an AUROC of 0.97 (Fig. 3B). Subclass analysis on structural hemoglobin variants demonstrated AUROCs of 0.94 for Hb E, Hb C and Hb D homozygotes and heterozygotes, 0.91 for sickle cell anemia, and 0.70 for Hb S heterozygotes (Fig. 3C). For combinations of hemoglobinopathies, the XGB model reached an AUROC of 0.92 (Fig. 3D). Performances of specific hemoglobinopathies of the logistic regression model can be found in online Supplemental Fig. 4.

DIFFERENTIATING THALASSEMIA FROM IDA

We assessed the capability of the XGB and logistic regression models to distinguish between thalassemia and IDA using a patient data set previously published by Galdakao-Usansolo Hospital in Spain. Both the XGB and logistic regression model attained an AUROC of 0.95 (±0.01) in differentiating cases with IDA and

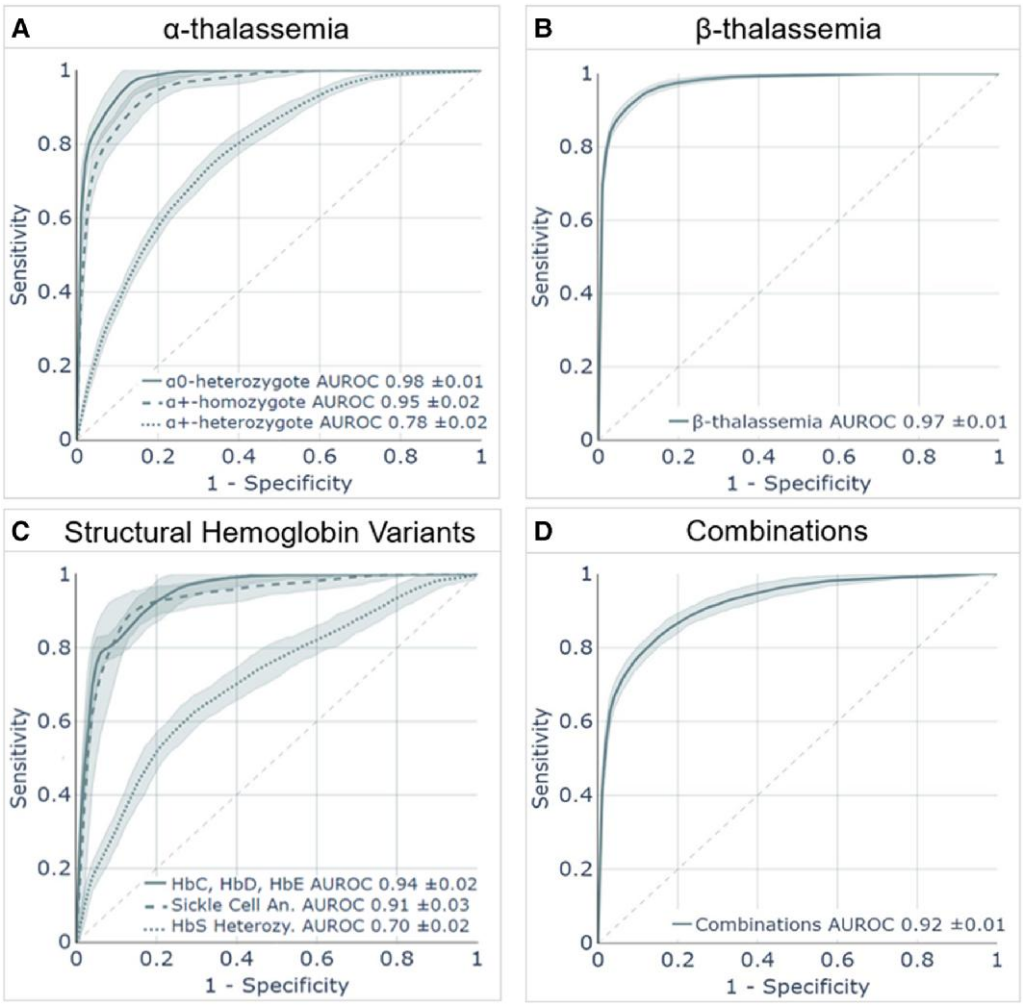

Fig. 3. Receiver operating characteristic plots for of the XGBoost (XGB) model for the subclasses of the external independent validation set. (A), α^0-heterozygote (n = 46), homozygote (including compound heterozygotes) (n = 68), and α^+-heterozygote (n = 304); (B), β-thalassemia (n = 306); (C), Hb E (hemoglobin E), Hb C (hemoglobin C), Hb D (hemoglobin D) (n = 35), sickle cell anemia (n = 44), and Hb S (Hemoglobin S) heterozygote (n = 79); and (D), combinations (combinations between thalassemia, and concomitant structural hemoglobin variants) (n = 118). Color figure available at https://academic.oup.com/clinchem.

thalassemia (Fig. 4A and D). Subclass analysis demonstrated that the XGB and logistic regression models distinguished α-thalassemia from IDA with an AUROC of 0.90 and 0.89, respectively. Both models reached an AUROC of 0.97 for distinguishing β-thalassemia from IDA (Fig. 4B and E). Median predicted probability scores of the XGB model for α-thalassemia (median = 0.93, IQR : 0.85 to 0.96) and β-thalassemia cases (median = 0.98, IQR: 0.96 to 0.99) significantly surpassed that for cases with an IDA (median = 0.54, IQR: 0.36 to 0.75) (Fig. 4C). The median predicted probability scores of logistic regression for cases with an IDA (median = 0.56, IQR: 0.45 to 0.65) was significantly lower compared to α-thalassemia (median = 0.83, IQR: 0.74 to 0.88) and β-thalassemia cases (median = 0.94, IQR: 0.89 to 0.98) (Fig. 4F).

EVALUATION OF THE MODEL AS A SCREENING TOOL ON CBC RESULTS

We evaluated the capacity of the XGB and logistic regression models as screening tools when implemented on a month of CBC results. In total, a data set of 19 969 CBC results from routine practice, containing 118 known positive cases, was used. The XGB model

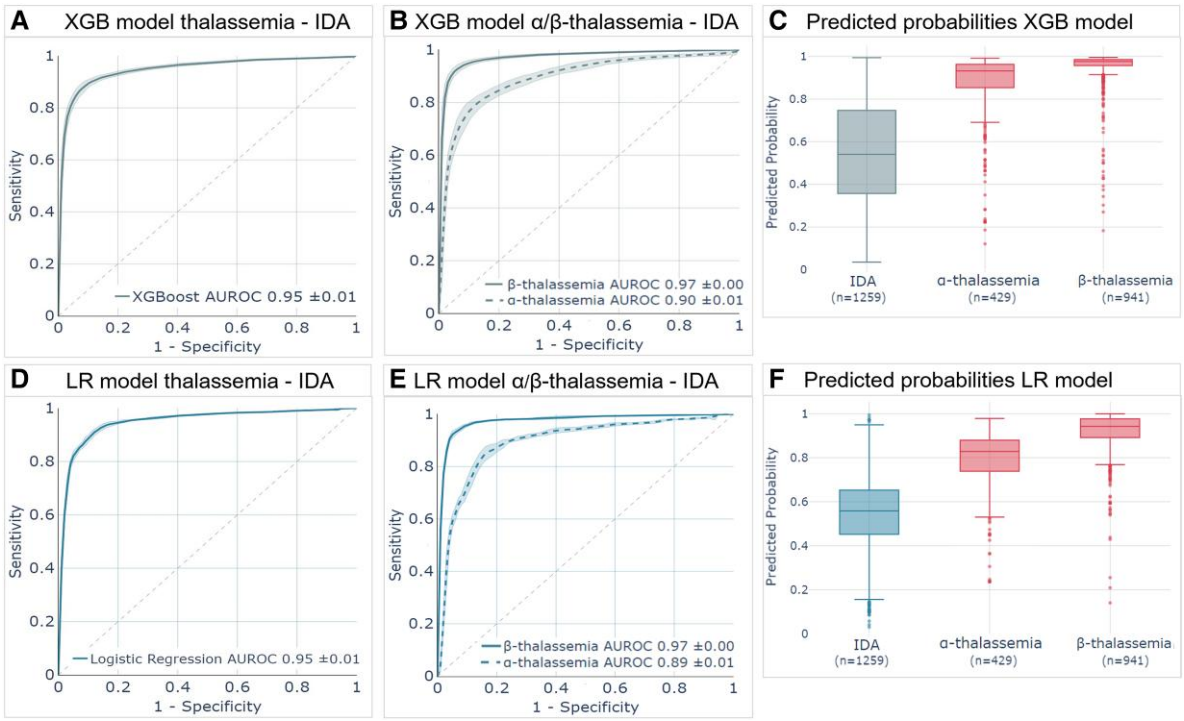

Fig. 4. Receiver operating characteristic plot of (A) the XGBoost (XGB) model and (D) logistic regression (LR) of the Spanish external validation data set differentiating thalassemia from IDA (iron deficiency anemia). Receiver operating characteristic plot of (B) the XGB model and (E) the logistic regression, differentiating α-thalassemia and β-thalassemia from IDA ($n_{\alpha\text{-thalassemia}} = 429$, $n_{\beta\text{-thalassemia}} = 941$, $n_{IDA} = 1259$). (C), XGB model predicted probabilities for IDA (median = 0.54, IQR: 0.36 to 0.75), α-thalassemia (median = 0.93, IQR: 0.85 to 0.96), and β-thalassemia (median = 0.98, IQR: 0.96 to 0.99); (F), Logistic regression predicted probabilities for IDA (median = 0.56, IQR: 0.45 to 0.65), α-thalassemia (median = 0.83, IQR: 0.74 to 0.88), and β-thalassemia (median = 0.94, IQR: 0.89 to 0.98). Color figure available at https://academic.oup.com/clinchem.

reached an AUROC of 0.97 (±0.01), and the logistic regression model 0.98 (±0.01) on differentiating negative from positive cases (Fig. 5A and D). When comparing sensitivities at several specificity thresholds, the logistic regression model showed higher sensitivities compared to the XGB model. At a specificity of 99.8%, the logistic regression model reached a sensitivity of 57%, whereas the XGB model had a sensitivity of 35% (Fig. 5B and E). Precision-recall curves revealed that the XGB model yielded a lower average precision of 0.55 (±0.06) compared to logistic regression's of 0.65 (±0.05). Notably, the logistic regression model exhibited greater uncertainty at high precision levels (0.75 to 1.00) (Fig. 5C and F). An overview of precision values at specific recall levels for both models is provided in online Supplemental Table 6. The maximum MCC for the XGB model was 0.51 at a probability threshold of 0.92, whereas, for logistic regression, it was 0.65 at a threshold of 0.78. Given that a significant number of false-positive cases from the XGB model were severely anemic patients from the intensive care unit (ICU), we do not advise using this algorithm for ICU patients.

Discussion

In this study, an ML algorithm was developed and extensively validated using routine CBC parameters to accurately predict a broad range of hemoglobinopathies, including thalassemias, various hemoglobin variants, and their diverse combinations, showcasing the ML algorithm's wide-ranging applicability in medical diagnostics. To address the significant challenge many laboratories encounter in integrating ML algorithms into their laboratory information systems, a logistic regression formula was concurrently developed, offering a more straightforward and feasible implementation in clinical care.

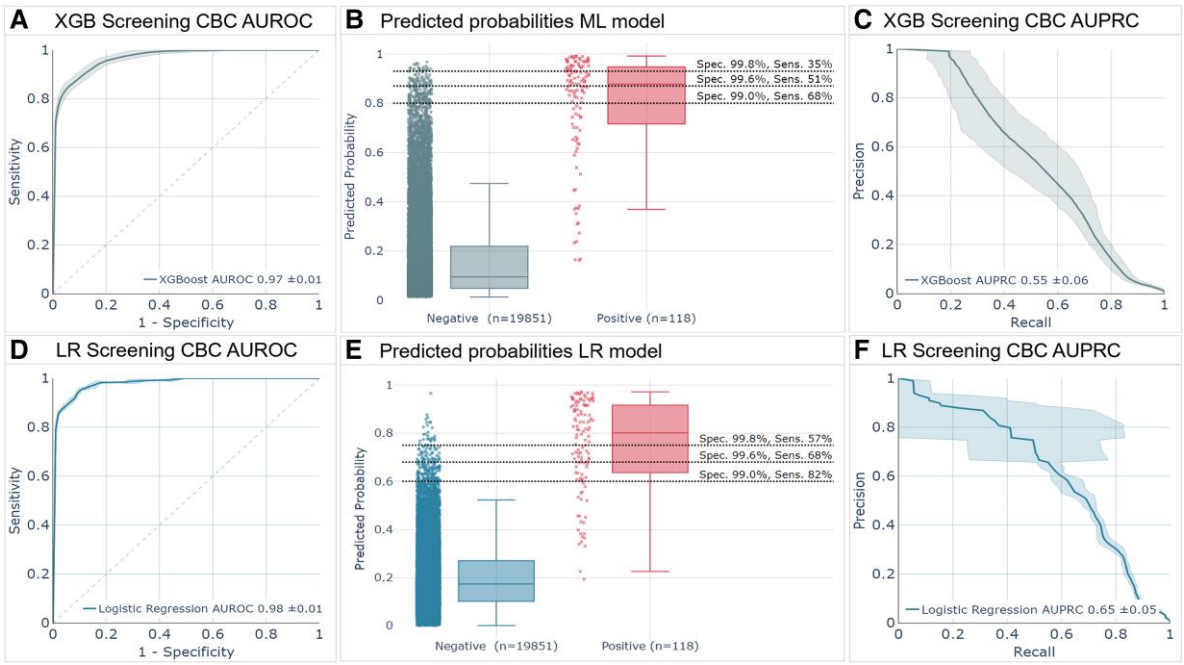

Fig. 5. Receiver operation characteristic plot of (A) the XGBoost (XGB) and (D) logistic regression (LR) models for distinguishing cases of the general population of Jeroen Bosch Hospital, considered negative, from known positive hemoglobinopathy cases. (B), XGB model predicted probabilities for positive cases (median = 0.85, IQR: 0.72 to 0.95) and negative cases (median = 0.10, IQR: 0.05 to 0.22); (E), LR model predicted probabilities for positive cases (median = 0.82, IQR: 0.64 to 0.92) and negative cases (median = 0.17, IQR: 0.10 to 0.27). The area under the precision-recall curve of (C) the XGB model and (F) the LR model. Color figure available at https://academic.oup.com/clinchem.

Numerous ML-based clinical decision support and case-finding systems rely heavily on laboratory data. These tools offer unique opportunities for laboratories to enhance the quality of healthcare. However, existing rule-based formulas and ML algorithms for identifying hemoglobinopathies have limitations. The majority of these rule-based formulas and ML models are tailored to specific subcategories or subgoals, such as identifying one specific thalassemia (19, 23–26) and distinguishing these from IDA (15, 17, 27–32). Moreover, these models are primarily designed using small single-center data sets lacking independent validation (15, 16, 18, 33–36). These factors limit the applicability and reliability of many of these formulas and models when applied outside their training context.

The deployment of ML models necessitates selecting an optimal threshold that prioritizes sensitivity or specificity, and the positive or negative predictive value. This choice is dependent upon local requirements and preferences, related to additional workload, cost-effectiveness, and the prevalence of hemoglobinopathies, which varies substantially among endemic and non-endemic regions. Considering these factors, the models presented in our study enable 2 potential clinical implementation strategies. Firstly, leveraging a high positive predictive value (PPV), the model can screen routine CBC tests to flag new potential hemoglobinopathy cases. Laboratory medicine specialists can then verify if these flagged patients have a known hemoglobinopathy diagnosis and, if not, recommend diagnostic testing for hemoglobinopathy. Secondly, employing its high negative predictive value (NPV), the model offers a cost-effective approach by advising against the pursuit of expensive α-thalassemia DNA diagnostics for patients who have both a negative HPLC/CE result combined with a low probability of hemoglobinopathy according to our model. This dual approach enhances early detection and reduces unnecessary testing. In this study the sensitivity and specificity for each threshold was provided, enabling laboratories to determine their desired cutoff value based on local preferences (Fig. 2C and F). Our simulation of the case-finding capacities of our models in a screening-based approach showed satisfactory results. Interestingly, the logistic regression model outperformed the XGB model (AUROC 0.98 vs 0.97,

and a MCC of 0.65 vs 0.51). This is likely due to the fact that the XGB model was also trained to identify sickle cell patients. In this screening analysis, there was only one sickle cell patient in the positive cases. Moreover, many of the false-positive cases in the XGB model were from ICU patients with severe anemia with high MCV, likely mimicking sickle cell CBC results. Therefore, use of the XGB algorithm for ICU patients is not advised.

Several limitations need to be considered. Firstly, the current model was exclusively developed and designed for adults, recognizing the substantial differences in CBC parameters between young children and adults. Secondly, the GAP-PCR was specifically designed to identify the common α-thalassemia deletions, and the strip-assay methods can detect approximately 90% of α-thalassemia gene mutations (8, 37). Consequently, instances classified as "negative" may still possess α-thalassemia mutations not covered by these diagnostic tests. The inclusion of such false-negative cases in the model development data set leads to an underperformance of the model.

On the other hand, this study exhibits several notable strengths. The XGB and logistic regression models were developed using extensive multicenter data collected over a span of 12 years, showcasing robust performance across various hematological analyzers. Furthermore, both models underwent thorough external validation. Moreover, the efficacy of the XGB model in distinguishing thalassemia from IDA was substantiated in a Spanish population, where it outperformed all other rule-based formulas, including Jayabose, Janel, Green and King, and Shine and Lal (14) (Supplemental Table 5).

In conclusion, this study effectively demonstrated the capability of an ML model to accurately predict a wide spectrum of hemoglobinopathies using routine CBC parameters. Moreover, a logistic regression model was developed, providing a more practical approach for implementation. Integration of either of these models into the laboratory information system facilitates automated detection of hemoglobinopathies based on routine CBC parameters.

Supplemental Material

Supplemental material is available at *Clinical Chemistry* online.

Nonstandard Abbreviations: CBC, complete blood count; XGB, eXtreme Gradient Boosting; IDA, iron deficiency anemia; AUROC, area under the receiver operating characteristic; Hb, hemoglobin; MCV, mean corpuscular volume; ML, machine learning; RBC, red blood cell count; MCC, Matthews correlation coefficient, ICU, intensive care unit.

Human Genes: *HBA1*, hemoglobin subunit alpha 1; *HBA2*, hemoglobin subunit alpha 2; *HBB*, hemoglobin subunit beta; *HBG1*, hemoglobin subunit gamma 1; *HBG2*, hemoglobin subunit gamma 2; *HBD*, hemoglobin subunit delta.

Author Contributions: *The corresponding author takes full responsibility that all authors on this publication have met the following required criteria of eligibility for authorship: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved. Nobody who qualifies for authorship has been omitted from the list.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form.*

Research Funding: None declared.

Disclosures: J. Krabbe, Abbott Cardiopulmonary Advisory Board. B. van Ginneken, shareholder for Thirona. R. Kusters, advisor to Sanquin.

Role of Sponsor: No sponsor was declared.

Acknowledgments: We are grateful to Henk Martens for his valuable input and insights. We would also like to thank Hans Hoffmann from H3L Consult for his contribution.

Data Availability: Information and data are available from the corresponding author upon reasonable request. XGB and logistic regression models are made publicly available via GitHub: doi:https://github.com/aschipper/hemoglobinopathies-AI and via Figshare: 10.6084/m9.figshare.25765302.

References

1. Centers of Disease Control and Prevention. Hemoglobinopathies: current practices for screening, confirmation and follow-up. Silver Spring (MD): Association of Public Health Laboratories; 2015.
2. Harteveld CL, Achour A, Arkesteijn SJG, ter Huurne J, Verschuren M, Bhagwandien-Bisoen S, et al. The hemoglobinopathies, molecular disease mechanisms and diagnostics. Int J Lab Hematol 2022;44:28–36.
3. Kohne E. Hemoglobinopathies: clinical manifestations, diagnosis, and treatment. Dtsch Arztebl Int 2011;108:532–40.
4. Taher AT, Weatherall DJ, Cappellini MD. Thalassaemia. Lancet 2018;391:155–67.
5. Modell B, Darlison M. Global epidemiology of haemoglobin disorders and derived service indicators. Bull World Health Organ 2008;86:480–87.
6. Williams TN, Weatherall DJ. World distribution, population genetics, and health burden of the hemoglobinopathies. Cold Spring Harb Perspect Med 2012;2:a011692.
7. Modell B, Darlison M, Birgens H, Cario H, Faustino P, Giordano PC, et al. Epidemiology of haemoglobin disorders in Europe: an overview. Scand J Clin Lab Invest 2007;67:39–69.
8. Harteveld CL, Higgs DR. Alpha-thalassaemia. Orphanet J Rare Dis 2010;5:13.
9. Kulozik AE. Editorial: hämoglobinopathien nehmen zu. Dtsch Arztebl Int 2010;107:63–4.
10. Angastiniotis M, Cannon L, Antoniou E, Brunetta AL, Constantinou G, Knoll EM, et al. Hemoglobin disorders in Europe: a systematic effort of identifying and addressing unmet needs and challenges by the Thalassemia International Federation. Thalass. Rep 2021;11:9803.
11. Goonasekera HW, Paththinige CS, Dissanayake VHW. Population screening for hemoglobinopathies. Annu Rev Genomics Hum Genet 2018;19:355–80.
12. van Vliet ME, Kerkhoffs JLH, Harteveld CL, Houwink EJF. Hemoglobinopathy

prevention in primary care: a reflection of underdetection and difficulties with accessibility of medical care, a quantitative study. Eur J Hum Genet 2022;30:790–4.
13. Traeger-Synodinos J, Harteveld CL, Old JM, Petrou M, Galanello R, Giordano PC, et al. EMQN best practice guidelines for molecular and haematology methods for carrier identification and prenatal diagnosis of the haemoglobinopathies. Eur J Hum Genet 2015;23:426–37.
14. Urrechaga E, Hoffmann JJML. Critical appraisal of discriminant formulas for distinguishing thalassemia from iron deficiency in patients with microcytic anemia. Clin Chem Lab Med 2017;55:1582–91.
15. Laengsri V, Shoombuatong W, Adirojananon W, Nantasenamat C, Prachayasittikul V, Nuchnoi P. ThalPred: a web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia. BMC Med Inform Decis Mak 2019;19:212.
16. Fu Y-K, Liu H-M, Lee L-H, Chen Y-J, Chien S-H, Lin J-S, et al. The TVGH-NYCU thalclassifier: development of a machine-learning classifier for differentiating thalassemia and non-thalassemia patients. Diagnostics 2021;11:1725.
17. Das R, Saleh S, Nielsen I, Kaviraj A, Sharma P, Dey K, Saha S. Performance analysis of machine learning algorithms and screening formulae for β-thalassemia trait screening of Indian antenatal women. Int J Med Inform 2022;167:104866.
18. Rustam F, Ashraf I, Jabbar S, Tutusaus K, Mazas C, Barrera AEP, de la Torre Diez I. Prediction of β-thalassemia carriers using complete blood count features. Sci Rep 2022;12:19999.
19. Phirom K, Charoenkwan P, Shoombuatong W, Charoenkwan P, Sirichotiyakul S, Tongsong T. DeepThal: a deep learning-based framework for the large-scale prediction of the α^{+}-thalassemia trait using red blood cell parameters. J Clin Med 2022;11:6305.
20. Traeger-Synodinos J, Harteveld CL. Advances in technologies for screening and diagnosis of hemoglobinopathies. Biomark Med 2014;8:119–31.
21. Jameel T, Baig M, Ahmed I, Hussain MB, Alkhamaly MD. Differentiation of beta thalassemia trait from iron deficiency anemia by hematological indices. Pak J Med Sci 2017;33:665–9.
22. LeDell E, Petersen M, van der Laan M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. Electron J Stat 2015;9:1583–607.
23. Al Agha AS, Faris H, Hammo BH, Al-Zoubi AM. Identifying β-thalassemia carriers using a data mining approach: the case of the Gaza Strip, Palestine. Artif Intell Med 2018;88:70–83.
24. Sadiq S, Khalid MU, Mui-Zzud-Din, Ullah S, Aslam W, Mehmood A, et al. Classification of β-thalassemia carriers from red blood cell indices using ensemble classifier. IEEE Access 2021;9:45528–38.
25. Feng P, Li Y, Liao Z, Yao Z, Lin W, Xie S, et al. An online alpha-thalassemia carrier discrimination model based on random forest and red blood cell parameters for low HbA2 cases. Clin Chim Acta 2022;525:1–5.
26. Barnhart-Magen G, Gotlib V, Marilus R, Einav Y. Differential diagnostics of thalassemia Minor by artificial neural networks model. J Clin Lab Anal 2013;27:481–6.
27. England JM, Fraser PM. Differentiation of iron deficiency from thalassaemia trait by routine blood-count. Lancet 1973;7801:449–52.
28. Mentzer WC Jr. Differentiation of iron deficiency from thalassaemia trait. Lancet 1973;7808:882.
29. Shine I, Lal S. A strategy to detect beta-thalassaemia minor. Lancet 1977;8013:692–4.
30. Jayabose S, Giamelli J, Levondoglu-Tugal O, Sandoval C, Ozkaynak F, Visintainer PL. Differentiating iron deficiency anemia from thalassemia minor by using an RDW-based index. J Pediatr Hematol Oncol 1999;21:314.
31. Huber A, Ottiger C, Risch L, Regenass S, Hergersberg M, Herklotz R. Thalassämie-Syndrome: klinik und diagnose. Swiss Med Forum 2004;4:947–52.
32. Zhang F, Yang J, Wang Y, Cai M, Ouyang J, Li J. TT@MHA: a machine learning-based webpage tool for discriminating thalassemia trait from microcytic hypochromic anemia patients. Clin Chim Acta 2023;545:117368.
33. Amendolia SR, Cossu G, Ganadu ML, Golosio B, Masala GL, Mura GM. A comparative study of K-nearest neighbour, support vector machine and multi-layer perceptron for thalassemia screening. Chemometr Intell Lab Syst 2003;69:13–20.
34. Borah MS, Bhuyan BP, Pathak MS, Bhattacharya P. Machine learning in predicting hemoglobin variants. Int J Mach Learn 2018;8:140–3.
35. Çil B, Ayyıldız H, Tuncer T. Discrimination of β-thalassemia and iron deficiency anemia through extreme learning machine and regularized extreme learning machine based decision support system. Med Hypotheses 2020;138:109611.
36. Das R, Datta S, Kaviraj A, Sanyal S, Nielsen P, Nielsen I, et al. A decision support scheme for beta thalassemia and HbE carrier screening. J Adv Res 2020;24:183–90.
37. Puehringer H, Najmabadi H, Law H-Y, Krugluger W, Viprakasit V, Pissard S, et al. Validation of a reverse-hybridization StripAssay for the simultaneous analysis of common alpha-thalassemia point mutations and deletions. Clin Chem Lab Med 2007;45:605–10.