Full text open access online (Since 2007)

Ethno Med, 17(3-4): 167-181 (2023)
PRINT: ISSN 0973-5070 ONLINE: ISSN 2456-6772 **DOI: 10.31901/24566772.2023/17.3-4.680**

Haemoglobinopathies: A Review on Statistical Modelling Perspective (Haemoglobinopathies: Statistical Modelling Techniques)

J. Kezia Angeline[1] and H. Gladius Jennifer[2]

School of Public Health, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Tamil Nadu, India
E-mail: [1]<johnkezangetsy@gmail.com>, [2]<gladiusj@srmist.edu.in>

KEYWORDS Haemoglobin Disorders. Monogenic Disorders. Prediction. Sickle Cell Disease. Thalassemia. Tribes

ABSTRACT Haemoglobinopathies indicate a group of monogenic disorders. The prevalence of Haemoglobinopathies differs based on geographical regions and is higher in tropical countries. Literature has identified that Haemoglobinopathies are higher in tribal and ethnic groups compared to the general population in India. This narrative review explores the methods to analyse data related to Haemoglobinopathies, identifies the research gaps and gains insights on model building and prediction of Haemoglobinopathies. This narrative review includes articles focusing on Haemoglobinopathies, Anaemia and other related disorders, published from 2008 onwards. Full-text and accessible articles found in relevant databases were included. Statistical and a few other techniques like Logistic Regression, Bayesian Modelling, Genetic Risk Scoring, Structural Equation Modelling, Multilevel Tests, Longitudinal Models, Machine Learning Algorithms and Genotyping methods were used to know the risk of acquiring Haemoglobinopathies and to identify important genetic variants. From the findings, it was observed that many studies were conducted to identify the prevalence of Haemoglobinopathies among the general population, but only few in the tribal population. Models based on machine learning algorithms were used for prediction involving haematological parameters and genetic variants, whereas the statistical techniques for prediction include demographic, nutritional, economic, cultural and social indicators.

INTRODUCTION

Haemoglobin (Hb), the red pigment found in red blood cells (RBCs), is in charge of transferring oxygen from the lungs to the body's organs and carbon dioxide from those organs back to the lungs. Any variation in the Hb structure leads to aberrant RBC function and illnesses such as thalassemia, sickle cell anaemia (SCA), etc. Haemoglobinopathies affect about 7 percent of people worldwide, in both developed and developing countries (Ganesh et al. 2023).

Genetic haemoglobin disorders that prevent haemoglobin from functioning normally are referred to as "haemoglobinopathies". They are classified as "thalassemia-related syndromes and abnormal haemoglobins (structural variants of haemoglobin)" (Kohne 2011).

Address for correspondence:
Dr H. Gladius Jennifer,
Associate Professor,
School of Public Health,
SRM Institute of Science and Technology,
SRM Medical College – New Block (5th Floor),
Kattankulathur, Chengalpattu,
Tamil Nadu, India
Phone: 7708628846
E-mail: gladiusj@srmist.edu.in

Disease Burden and Distribution: A Global Perspective

The disorder of Haemoglobinopathies includes both qualitative [abnormal haemoglobin (Hbs)] and quantitative [thalassemia] disorders of globin synthesis, which are the most commonly inherited diseases in people worldwide. In many developing nations, haemoglobinopathies are particularly prevalent in malaria-endemic regions, where it can pose a significant threat to public health (Kutlar 2007). Though sickle cell disorders are found to be more prevalent globally, thalassemia syndromes like α-thalassemia, β-thalassemia, and haemoglobin-E disease also have their prevalence, ranging from 2.5 to 15 percent in the eleven countries that make up the South-East Asia (SEA) Region of the World Health Organisation (WHO). Clinical manifestations can range from mild transfusion-dependent anaemia to severe anaemia, placing a heavy burden on medical and blood supplies (Region-

al Desk Review of Hemoglobinopathies with an Emphasis on Thalassaemia and Accessibility and Availability of Safe Blood and Blood Products as per These Patients' Requirement in South-East Asia under Universal Health Coverage (n.d.).

In tropical regions, there are various haemoglobinopathies with high gene frequencies. There are isolated cases of sickle cell anaemia in the Mediterranean region, the Middle East, and South-East Asia, in addition to the whole sub-Saharan region of Africa. Haemoglobin SC disease is more common in parts of Western Asia and North Africa. There are isolated cases of Haemoglobin S (HbS) beta-thalassemia in sub-Saharan Africa. Haemoglobin E levels are high in Bangladesh, Myanmar, East and South-East Asia, and the eastern half of the Indian regions. The mild forms of alpha-thalassemia are most prevalent in a broad tropical belt that extends from sub-Saharan Africa through the Mediterranean and Middle East to the Indian regions and all of East and South-East Asia (Williams and Weatherall 2012).

Distribution of Haemoglobinopathies and Haemoglobin Variants: An Indian Perspective

Of the 4800 cases screened to identify the structural variants of haemoglobins from the Department of Pathology and Thalassemia Control Cell of Hindu Rao Hospital in Delhi, 290 (6.04%) cases had abnormal hemoglobins. Traits like "HbE beta-thalassemia (HBE-BT), HbE trait (HBAE), beta thalassemia trait (BTT), HbS trait (HBAS), beta-thalassemia intermedia/major (BTI/M), delta beta thalassemia trait (delta BTT), HbE disease (HBEE), HbJ Meerut, and Hb Hope" were among the many variants reported in this study out of which BTT was the most prevalent trait (n=216,74) and HbJ Meerut (n=1, 0.34%. and Hb Hope (n = 1, 0.34%) were the least prevalent traits identified (Pant et al. 2014).

Among the 119,336 cases studied, over a ten-year period to find the prevalence of haemoglobinopathy and thalassemia in patients of hospitals with tertiary care facilities in West Bengal, abnormal haemoglobin patterns were identified in about 12.17 percent of patients. The trait of (beta) thalassemia was the most prevalent among those abnormal patterns, present in 4.60 percent of patients. 3.02 percent of patients had HbE trait, 1.66 percent had thalassemia major/intermedia, and 1.16 percent had Eβ-thalassemia. Other identified variants included those that cause "HbE and sickle-cell diseases, sickle thalassemia", traits such as "HbH, delta beta-thalassemia trait, HbQ-India, beta-thalassemia, HbD-Punjab, and double heterozygous state of HbS and HbE, and that of HbS and HbD, HbJ-Meerut, hereditary persistence of foetal haemoglobin (HPFH), and Hb Lepore" (Mondal and Mandal 2016).

Distribution of Haemoglobinopathies and Haemoglobin Variants: An Indian-Tribal Perspective

Haemoglobinopathies, including thalassemias, are distributed throughout India, with uneven distribution of genes. The distribution of alpha-thalassemia among the tribal people of India was scarce, patchy and incomplete (Sengupta 2008), whereas, pertaining to Sickle cell distribution, Beta-thalassemia and other Haemoglobinopathies were found to be concentrated more in the Eastern and Central parts of India, which showed various migration patterns (Goonasekera et al. 2018). Apart from these, the structurally abnormal variants of haemoglobin like "HbE, HbS and the Foetal Haemoglobin (HbF)" were also found to be distributed in various regions of India (Sengupta 2008).

Haemoglobinopathies, particularly, haemoglobin S(HbS) and haemoglobin E(HbE), and beta-thalassemia, have varied gene distributions and they pose serious health hazards to Indian tribal populations. Among various Indian tribal populations, those populations in the Andaman and Nicobar Islands, West Bengal, Odisha, and the North-East were the main populations affected by HbE (Ghosh et al. 2015).

For the purpose of identifying haemoglobinopathies and beta-thalassemia, 1204 samples from Assamese tribal workers in tea gardens were analysed using the Complete Blood Count (CBC) and High Performance Liquid Chromatography (HPLC) methods. The findings revealed that sickle cell anaemia and beta-thalassemia were very common among the selected population, with beta-thalassemia prevalence among the Munda ethnic group being 3.07 percent and sickle cell anaemia prevalence among the Lohar ethnic group being 4.73 percent, respectively (Teli et al. 2016).

To estimate the incidence of "Sickle Cell Anaemia (SCA), beta-thalassemia, and Glucose-6-Phosphate Dehydrogenase (G6PD) deficien-

cy" among the tribes of Rajasthan, a cross-sectional study was carried out in the Abu Road Block of Sirohi District, where screening for Sickle Cell Anaemia, beta-thalassemia, and G6PD deficiency was done through medical camps held in schools and villages. 7167 tribal members in total were screened, of whom 610 members tested positive for SCA, with 8.51 percent prevalence. G6PD deficiency was found to be prevalent in 2.88 percent and β-thalassemia in 7.25 percent (Mohanty et al. 2022).

Distribution of Haemoglobinopathies and Haemoglobin Variants: Perspectives from Tamil Nadu

5207 cases from Chennai were examined for beta-thalassemia heterozygous status, wherein 387 individuals were identified and subjected to molecular DNA analysis to find out the mutations causing beta thalassemia. The "IVS 1-5 (G-C) mutation" had the highest prevalence of the 30 different mutations that were identified through molecular characterisation of beta- thalassemia mutations. "IVS II-1 (G>T), CD 37 TGG-TGA, IVS II 781 (C-G), CD114 CTG-CCG, and PolyA (A-G)" were the five new, uncommon mutations that were discovered and first reported in India. HBB.c319DelC, a new beta-thalassemia mutation, was also discovered (Selvaraj et al. 2022).

To ascertain the incidence of haemoglobinopathies in South India, samples from Tamil Nadu and Andhra Pradesh were examined for four years (2018-2022). The regional laboratory in Chennai, which serves people of Tamil Nadu and Andhra Pradesh, received a total number of 17,678 cases during the period between January 2018 and July 2022 for screening of abnormal haemoglobinopathies or assessing anaemia. It was found that 1,194 samples out of the 17,678 cases had abnormal haemoglobin variants. HbD Homozygous was identified to have the least common trait with 0.08 percent and beta-thalassemia having the most common trait with 69.13 percent (824 cases), among all the abnormal haemoglobin variants (Kavitha et al. 2023).

Distribution of Haemoglobinopathies and Haemoglobin Variants: Perspectives from the Tribes of Tamil Nadu

There were 36 tribes residing in Tamil Nadu (Tribal Welfare Department, Tamil Nadu n.d.-a) with a population of 7,94,697 (India - A-11 Appendix: District Wise Scheduled Tribe Population (Appendix), Tamil Nadu 2011). Todas, Kotas, Irulas, Kurumbas, Paniyans, Kattunayakans were referred to as Particularly Vulnerable Tribal Groups (PVTGs) (Tribal Welfare Department, Tamil Nadu n.d.-b), and the name has been given so, because there might be a decrease in the population of these tribes or the population of these tribes were unchanged (Ganesh et al. 2021). Many of the tribes from Tamil Nadu dwell in the Eastern and Western Ghats and sporadic hilly areas bordering the hills and the plains. They are typically small and live in forests. They are fiercely committed to preserving their customs and culture. Although farming and cultivation are their main occupations, they also heavily rely on forest areas (Mohanty et al. 2013). In Tamil Nadu, the tribal communities in the Gudalur and Pandalur taluks of the Nilgiris district were found to be the most prevalent regions with sickle cell anaemia. Taking into account the current population of 25,000, studies conducted in this area by the Indian Council of Medical Research (ICMR) and the Nilgiris Adivasi Welfare Association (NAWA) estimate that 500 patients have sickle cell disease (TNHSP n.d.).

A multicentric study was carried out between the years of 2000 and 2005 among 14 primitive tribes from the Indian states of Tamil Nadu, Orissa, Gujarat, and Maharashtra in order to evaluate the spectrum of haemoglobinopathies. The "Irula, Kurumba, Moolu Kurumba, and Paniya tribes" in the Nilgiris in Tamil Nadu were chosen for this study to figure out the prevalence of beta-thalassemia and sickle-cell genes respectively. The "sickle-cell trait" was found to be the most prevalent (n = 819, 24.19%), with the least being one unidentified variant (0.03%), in a total of 3385 people who were screened. The sickle cell trait was the most common in Tamil Nadu's Paniya tribes (231 of 877). Based on the allele frequency distribution of beta-thalassemia and Haemoglobin S (HbS) (calculated from individuals who were not related), all tribal groups studied had the "HbS allele", with Tamil Nadu tribals having a higher frequency. Moolu Kurumbas of Tamil Nadu had a very high "HbS allele frequency" (Mohanty et al. 2013).

Studies have suggested that the prevalence of Haemoglobinopathies were considerably higher in the tribes when compared to the gen-

eral population. The reasons might be consanguineous marriages, migration patterns, lack of awareness on importance of family screening and genetic testing.

Molecular Methods to Rule Out Haemoglobinopathies

Variations in the globin genes, the alpha and the beta, especially the duplicated HBA1 and HBA2 and HBB, were responsible for most of the clinically relevant haemoglobinopathies. Over 1600 globin gene variants have been identified, of which about one-third are clinically significant. The "HbS variant" is the one that occurs clinically the most frequently, and it causes "SCD in homozygosity and in combination with a number of other common haemoglobin variants", such as "HbC, HbD, HbE, HbO-Arab and with thalassemia" (Harteveld et al. 2022).

Methods used for the molecular diagnosis of haemoglobinopathies include "dot-blot and Reverse Dot-Blot analysis (RDB), allele-specific oligonucleotide hybridisation, restriction enzyme analysis, gap-Polymerase Chain Reaction (gap-PCR) (for common deletions), allele-specific priming or Amplification Refractory Mutation System (ARMS), and Deoxyribonucleic Acid (DNA) sequencing". The most popularly used technique for detecting mutations is ARMS. To discover mutations that either add or remove restriction enzyme sites, restriction enzyme analysis could be used. Large deletions of the beta-globin gene can have an impact and can be identified using gap-PCR. "Automated DNA sequencing" is used to diagnose rare mutations in the beta-globin gene. Before the sequence can be recognised, a multi-step process involving "PCR amplification, cycle sequencing, and precipitation" is necessary. The subsequent step is to review the procedure and look for any changes (Verma et al. 2012).

Need of the Hour: Management and Control of Haemoglobinopathies Among the Tribal Population

Haemoglobinopathies are the diseases that are most frequently brought on by a single gene (Goonasekera et al. 2018). Thalassemia can be managed by screening the general population for carriers and conducting prenatal diagnostics on couples who have been discovered to be highly likely to give birth to a thalassemic child (Saxena and Phadke 2002). Population screenings for alpha- and beta-thalassemia, HbE disease, and sickle cell disease in newborns, adolescents, adults of reproductive age, and relatives of diagnosed cases are now being conducted by a number of governmental and international organisations (Goonasekera et al. 2018; Saxena and Phadke 2002).

The majority of tribal members lack basic literacy skills and are uninformed about haemoglobinopathies, including their prevalence, treatments that are available, and management techniques. Raising awareness among tribal members could be the first phase, which serves as the starting point for management strategies for haemoglobinopathies. The subsequent phase involves testing at four different levels, including extended family screening, carrier screening of boys and girls, genetic counselling for couples prior to conception and genetic testing of pregnant women who are visiting a medical facility (Sengupta 2008).

Aims/Objectives

In addition to identifying research gaps and providing insights into model building and Haemoglobinopathy prediction, this narrative review aims to investigate data analysis and a few oher techniques to modelling for Haemoglobinopathies.

METHODOLOGY

This narrative review dealing with application of Statistical Methods/Models along with a few other techniques in Haemoglobinopathies and related studies includes articles from the year January 2008 to October 2023. Most appropriate search engines like PubMed, Google Scholar and Scopus and keywords like Haemoglobinopathies, Modelling, Prediction and Tribes were used to search for and retrieve the relevant articles. Only the full-text articles that were available in journals with open access were included for the study. The data included in the articles were related to Haemoglobin, Anaemia and Haemoglobinopathies and consisted of both primary and secondary data, belonging to various regions worldwide, inclusive of the general popu-

lation as well as the tribes. Articles with abstracts, those with data other than Haemoglobin and Haemoglobinopathies and the ones with techniques apart from modelling were excluded. Modelling techniques like Logistic Regression, Genetic Risk Scoring and Prediction, Model building using Machine Learning Techniques, Structural Equation Modelling, Multilevel Tests, Longitudinal Models Bayesian Models and Genotyping methods were used for identifying the determinants and the predictors of Haemoglobinopathies.

Importance of Prediction Models

"Predictive modelling involves building models that can forecast outcomes in the future using mathematical or computational techniques." It is useful in several medical procedures, including clinical trials and clinical decision-making (Toma and Wei 2023). Based on the prediction horizon, prediction models are separated into two distinct groups of "prognostic models (time horizons are present) and diagnostic models (time horizons are absent)" (Belbasis and Panagiotou 2022).

RESULTS AND DISCUSSION

Logistic Regression Models

In epidemiologic studies, logistic regression permits simultaneous analysis of many explanatory variables besides minimising confounding variables. With many explanatory variables, logistic regression calculates odds ratios (OR). Multiple linear regressions are used, but the response variable can be binomial, multinomial, or ordinal (Sperandei 2014). Logistic regression requires the assumptions of "independence of errors, linearity for continuous variables, no multicollinearity, and no strongly influential outliers" (Stoltzfus 2011). A logistic regression model must be chosen for the study along with independent variables. The model building strategy and independent variables were to be considered simultaneously. Though direct (full, standard, or simultaneous), sequential (hierarchical), and step-wise (statistical) regression model building methods exist (Stoltzfus 2011).

In July-August 2012, 170 reproductive-age tribal women (15-45 years) from the tribal hamlets of Koraga and Marathi Naik in Udupi Taluk, Karnataka were recruited for a "community-based cross-sectional study" on the prevalence of anaemia. Demographics, diet, menstrual, marital, reproductive, and general health history were collected using an interview schedule, along with haemoglobin levels. Occupational differences were statistically significant in multivariate logistic regression. Non-working participants had an increased risk of anaemia [odds ratio of 2.924 (95% Confidence Interval (CI): 1.060–8.066), P value = 0.038] (Pattanshetty et al. 2013).

In order to ascertain the rate of delayed thalassemia diagnosis in the Iranian population, a cross-sectional study involving 1003 enrolled thalassemic patients was carried out in 2015. Utilising the univariate as well as multivariate ordinal logistic regression models, the factors related to a delay in diagnosis were assessed. The data collected included questions about birth date, demise date (when applicable), gender, thalassemia variations (major, intermedia) or sickle cell, educational level and employment of the parents. 64.9 percent of the 981 patients who had the delayed diagnosis were identified (response rate: 97.8%). According to a bivariate analysis, "delayed diagnosis" was associated with gender, level of maternal education, type of thalassemia, birth cohort, and status of the patient. Girls were more prone to have a delayed diagnosis than boys, as per the findings of multivariate ordinal logistic regression (adjusted OR = 1.32, 95% CI: 1.02, 1.71). In comparison to those with thalassemia intermedia, patients with thalassemia major had significantly lower odds of delayed diagnosis (0.58 times). For the birth cohorts from 1980 and earlier, 1981 through 1990, 1991 through 2000, and 2001 to the present, the odds ratios (ORs) for delayed diagnosis were 1.0, 1.52, 1.55, and 2.22 respectively ('p' for trend = 0.014). Patients who had passed away had a 1.95 times higher chance of getting a delayed diagnosis than those who were still alive (Hassanzadeh et al. 2016).

1,12,714 children from the "Indian National Fertility and Health Survey" between the years 2015-2016 who had available haemoglobin data were examined, which allowed researchers to assess the relative and combined contributions of factors related to childhood anaemia in India. To find connections between parent anaemia, dietary intake of kids, and household traits, "multinomial logistic regression models" were used.

The relationship between household characteristics, children's nutritional intake, and haemoglobin levels was also examined using a linear regression analysis. Children who had anaemia had odd ratios of 0.981 (0.961-1.001), vitamin A deficiency (0.813 (0.794-0.833), and lower maternal haemoglobin levels (1.992 (1.957-2.027). Comparisons of some important haemoglobin indices show that height and weight have a greater impact on the levels of haemoglobin than the nutritional status (Onyeneho et al. 2019).

In 2018, the Nigeria "Demographic and Health Survey (DHS)" data were used in a study to investigate the association between sickle cell disease and stunting, wasting, and underweight. 11,233 kids between the ages of 6 and 59 months who had successfully undergone genotyping for sickle cell disease made up the sample. According to sickle cell status, the study's goals were to describe the prevalence of undernutrition, evaluate its correlation with anthropometrically measured growth slowdown, and look into the mediating role of haemoglobin. For SCD genotyping, the "SickleSCAN" rapid diagnostic test was made use of. The Z scores were derived using the 2006 World Health Organisation's Child Growth Standards for the anthropometrics of weight and height with each other and each one with age. Logistic regression models were fitted to determine the association between SCD and stunting, weight loss, and underweight. The results pointed out that the prevalence of stunting was the highest and that of wasting was the lowest among children with SCD at 55.4 percent (54.5-56.4) and 9.1 percent, respectively. In comparison to children with normal haemoglobin, children with sickle cell disease had odds of stunting that were 2.39 times more [adjusted odds ratio (aOR) 2.39, 95% confidence interval (CI): 1.26-4.54]. SCD was much associated with being underweight (aOR 2.64, 95% CI: 1.25-5.98). All the three anthropometric indices, namely, stunting, wasting and underweight were significantly mediated by haemoglobin level. The regression analyses revealed a significant association between stunting and SCD (Islam et al. 2021).

The prevalence of thalassemia among the six major hill-tribes of Thailand was estimated through a genetic epidemiological study. The study had 1200 participants from 6 different tribes namely, "Akha, Lau, Hmong, Yao, Karen, and Lisu", who were students in the grades of 4 to 6 from 13 schools selected from Chiang Rai Province of Thailand. The required information was gathered through questions about age, gender, weight and level of grade. In addition, 15 questions about medical and family histories, like whether any such conditions existed before, whether any blood transfusions had been done, whether any ferrous sulphate had been taken in the past, and what tribe the parents belonged to, were also included. Thalassemia carriers made up 9.8 percent of the population. It was found that Karen tribe was more likely than Akha to have positive results for "Osmotic Fragility (OF) and/or DiChlorophenol Indophenol Precipitation (DCIP)" in the overall model of the logistic regression analysis (OR = 2.45, 95%CI = 1.20-4.98) used to analyse the relationships between tribes and positive results for OF and/or DCIP. Separating the genders revealed that Karen females were more likely than Akha males to have positive results for OF and/or DCIP (OR = 5.25, 95%CI = 1.04-6.11), while Yao males were more likely than Akha males to have the positive results (Apidechkul et al. 2021).

To compare anaemia trends in tribal and general caste Non-Pregnant Women of Reproductive Age Group (NPWRA) from 1998 to 2021, data from four rounds of the "National Family Health Survey (1998-1999, 2005-2006, 2015-2016, 2019-2021)" were utilised. The study found that tribal and general caste NPWRA had similar trends in anaemia (haemoglobin <12 g/dl) and calculated the extent of "decline" and "narrowing" due to underlying factors. The multivariable logistic regression showed that age, education level, number of children, place of residence, wealth, and NPWRA body mass index independently predicted the anaemia status. After controlling for independent determinants, the unadjusted beta coefficient of "rounds" (representing anaemia decline in log of odds) decreased from -0.314 (-0.377, -0.251) to -0.242 (-0.308, -0.176). After controlling the variables that were independently associated, the beta coefficient of the "caste*round" interaction term (representing the narrowing of tribal/general anaemia gap in log of odds of anaemia) attenuated from -0.311 (-0.382, -0.241) to -0.288 (-0.358, -0.217) (Ghosal et al. 2023).

In order to determine the prevalence of varied hemoglobinopathies and their associations with anaemia in the two sub-districts of Habiganj, Sylhet Division, Bangladesh, capillary blood samples from 900 women at the reproductive age-group and 395 children (aged 6-37 months) participating in the "Food and Agricultural Approaches to Reducing Malnutrition (FAARM)" trial were examined. The associations between each of the hemoglobinopathies, haemoglobin levels, and anaemia were examined using linear and logistic regression. The study's findings demonstrated that in non-pregnant women, lower haemoglobin levels (-0.5 g/dl) and higher odds of anaemia (OR: Odds Ratio 2.1) were strongly associated with any inherited blood disorder. Non-pregnant women with any form of beta-thalassemia had 1.1 g/dl lower haemoglobin concentration and more than 7 times the likelihood of having anaemia compared to women without any haemoglobin disorders. The average haemoglobin concentration was 0.4 g/dl lower in alpha-thalassemia patients who were not pregnant compared to those who did not have any haemoglobin disorders. The weak evidence suggested that the presence of inherited disorder of blood was found to be related to the lower concentrations of haemoglobin (-0.4 g/dl), and having beta-thalassemia with anaemia (OR 4.5) in children (Wendt et al. 2023).

Longitudinal Models

Longitudinal data analysis consists of data that is deficient, correlated, and accumulated over a patchy period of time. Using longitudinal data, researchers can assess a variety of aspects of a disease, such as the time of disease onset, individual and group patterns over a period of time, changes found in outcome(s) over time in relation to associated risk factors, and so forth (Garcia and Marder 2017).

The effects of a fortified snack on school-aged anaemia were examined in a Haitian cluster randomised trial from 2012 to 2013. A longitudinal study of 1,047 children aged 3-13 years measured haemoglobin (Hb), anthropometric measurements, and bioelectrical impedance. Socioeconomic and demographic factors, eating habits and infectious disease morbidities were measured at baseline and endline. Longitudinal regression with generalised least squares and logit models with random effects identified anaemia risk factors beyond the effect of intervention. Baseline anaemia affected 70.6 percent of children, with 2.6 percent having severe anaemia. Stunting was associated with mild and severe anaemia (adjusted OR: 2.47; 95% CI: 1.30-4.71). These findings emphasise the need to mitigate stunting, improve nutrition through supplementation and deworming, and investigate genetic factors contributing to anaemia (Iannotti et al. 2015).

Genetic Risk Scoring (GRS) and Prediction Models

Data from various markers are incorporated into a statistical model known as a risk prediction model. Common models include classification trees, Cox proportional hazards models, and logistic regression models. Each type of model generates a predicted risk for each individual using data from the model (Janes et al. 2008).

To develop a Single Nucleotide Polymorphism (SNP) model to predict foetal haemoglobin levels in Sickle Cell Anaemia, 841 Black CSSCD (NCT00005277) homozygous for the HbS gene or with HbS beta-thalassemia had their HbF levels measured for 181 patients with pulmonary hypertension and sickle cell disease receiving sildenafil therapy (Walk-PHaSST; NCT00492531), 77 patients from a study of pulmonary hypertension in children with sickle cell disease (PUSH; NCT00495638) and 127 sickle cell anaemia patients from C-Data were recruited. Boston University used Illumina Human610-Quad SNP arrays with 600,000 SNPs to genotype CSSCD, PUSH, C-Data, and Walk-PHaSST samples from the discovery and replication cohorts were included as the validation cohorts. After processing samples according to manufacturer instructions, "BeadStudio" Software produced genotypes using Illumina predefined clusters. Multiple GRS were trained using CSSCD genotype data. HBS1L-MYB intergenic region, BCL11A quantitative trait locus, and HBB gene cluster SNPs were included in the models. The HbF of sickle cell anaemia patients was predicted by an ensemble of 14 models with different genetic risk scores based on SNPs. In contrast to the discovery cohort, where predicted and observed HbF correlated 0.28 to 0.44, the ensemble of 14

models explained 23.4 percent of variability in HbF (Milton et al. 2014).

An international multi-centric study recruited 890 beta-thalassemia patients from various hospitals and clinics. The BCL11A, HBG2:g.-158C>T polymorphism, and HBS1L-MYB intergenic region modulate foetal haemoglobin production. To calculate a severity score, 54 genetic variants were identified at these five loci strongly associated with beta-thalassemia phenotype improvement. Through Cox proportional hazard analysis on a training set, the impact of these loci on the age of regular transfusions was analysed, a Thalassemia Severity Score was created, and validation was conducted on a testing set. The model had high discriminatory power (C-index=0.705; R square=0.343) and was validated on the testing set using transfusion-free survival probability ($p<0.001$) and transfusion dependency status (Area Under the Receiver Operating Characteristic Curve=0.774; $p<0.001$) (Danjou et al. 2015).

Methods Based on Machine Learning Techniques

The goal of machine learning helps machines to handle data more efficiently. Various ways to teach machines to learn without programming have been studied extensively. Numerous mathematicians and programmers use various methods to solve this massive data set problem (Mahesh 2018).

Better methods are needed to distinguish Iron Deficiency Anaemia (IDA) from Thalassemia Trait (TT). To develop an automated prediction model for IDA-TT separation, Thai adult HMA laboratory data was used. Hypochromic microcytic anaemia (HMA) data was divided into 146 TT and 40 IDA groups. Discriminant models were created using k-Nearest Neighbours(k-NN) decision tree, random forest (RF), artificial neural network (ANN), and support vector machine (SVM)". Comparing performance to thirteen discriminant formulas and indices. A web-based tool called 'ThalPred' used an SVM model to analyse seven blood parameters of Red Blood Cell (RBC), Haemoglobin (Hb), Haematocrit (HCT), Mean Corpuscular Volume (MCV), Mean Corpuscular Haemoglobin (MCH), Mean Corpuscular Haemoglobin Concentration (MCHC) and Red cell Distribution Width (RDW) . External accuracy, Red Blood Cell (RBC), Haemoglobin (Hb), Haematocrit (HCT), Mean Corpuscular Volume (MCV), Mean Corpuscular Haemoglobin (MCH), Mean Corpuscular Haemoglobin Concentration (MCHC) and Red cell Distribution Width (RDW) for ThalPred were 95.59, 0.87 and 0.98, respectively (Laengsri et al. 2019).

The haemoglobin variations in 752 complete blood count laboratory analyses of adult patients aged 18 and above from Lagos State University Teaching Hospital were categorised using machine learning's classification algorithms by using "MATLAB R2019a, Go (go1.16.5), GoLearn, Python (3.9.5), Scikit-Learn (1.1.0), and Google Sheets". Probabilistic models that best captured feature-class interaction was the study's goal. The confusion matrix, classification accuracy, precision, recall, specificity, F1 score, and Receiver Operating Characteristic (ROC) curve are used to evaluate models. Support Vector Machine (SVM), one of the five classification algorithms, had the highest accuracy (92.3%) and Area Under the Curve (AUC) (99.2%), while random forest had the lowest (84.7%). Decision tree, Naive Bayes, and k-nearest neighbours had 90.0 percent, 86.4 percent and 87.6 percent classification accuracy, respectively. The support vector machine algorithm predicts haemoglobin variants better using haematological parameters (Okandeji et al. 2022).

Recent research has suggested a precise method for detecting beta-Thalassemia using Complete Blood Count (CBC) red blood indices and supervised machine learning, by using a unified framework of two features selection techniques, Principal Component Analysis (PCA) and Singular Vector Decomposition (SVD). The data imbalance between beta-Thalassemia carriers and non-carriers were addressed using"Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic (ADASYN)" Along with modern machine learning, deep learning models were used in extensive experiments. The experimental results showed an accuracy score of 0.96 (Rustam et al. 2022).

Decision Tree Based Models

Decision trees are popular due to their simplicity, lack of ambiguity, and missing data robustness. Separate and continuous variables can be targets or independent. Symptom-based decision trees diagnose medical conditions. Decision trees define patients, clinical subtypes,

and conditions. Sampled study data can be used to create training and validation datasets. Decision tree models with the training dataset and the right tree size with the validation dataset are considered the best model (Song and Lu 2015).

Pakistan's Punjab Thalassemia Prevention and Programme (PTPP) has developed a three-layer framework to identify beta-thalassemia carriers. The dataset includes data from 5,066 patients who underwent screening in 2019. The framework uses Decision Tree, Naive Bayes, and Support Vector Machine learning techniques. The dataset is trained locally before being sent to a central server for aggregation via Federated Learning (FL). The model, trained using FL, achieved a 92.38 percent accuracy rate in distinguishing between beta-thalassemia carriers and those without the gene (Farooq and Younas n.d.).

Using the decision tree analysis and cut-off values of equations based on five blood cell characteristics, a new method accurately identified beta-thalassemia trait and Haemoglobin E carriers to speed up and lower the cost of large-scale screening programmes. Blood cell metrics and chromatographic data were used to identify thalassemia carriers and features. Standardised chromatographic screening of 2,984 people was examined. Of 289 beta-thalassemia carriers, 63 had haemoglobin E, 15 Punjab D, 4 HPFH (Hereditary Persistence of Foetal Haemoglobin) and 14 had a specific HbA2 level. The researchers tested the model with 108 samples. Three samples had HbE heterozygotes, 11 mild beta-thalassemia. Decision tree and equations using five blood cell parameters correctly identified haemoglobin E and beta-thalassemia trait carriers. This method found 11.65 percent haemoglobin E and 51.3 percent beta-thalassemia trait carriers. Chromatographic analysis was unnecessary after excluding 27 percent of non-thalassemia carriers (Samanta et al. 2021).

Structural Equation Modelling (SEM)

Structural equation modelling (SEM) quantifies and analyses latent-observed variable relationships. It examines causal relationships between variables which are linear, while accounting for measurement error. Its applications include simple variable relationship analyses and complex measurement equivalence analyses for first- and higher-order constructs (Beran and Violato 2010).

A study examined how psychosocial factors affected Thalassemia Major patients' quality of life in 12 Sicilian locations. Social support, life engagement, proactive personality, self-efficacy, quality of life, and objective and subjective responses to life demands were considered. Using structural equation modelling, it was identified that self-efficacy directly affected social support (β=.20, $p<.001$), quality of life (β=.23, $p<.001$), and life engagement (β=.18, $p<.001$) (Platania et al. 2017).

In a cross-sectional analytical study of 389 Transfusion-Dependent Thalassemia (TDT) patients receiving treatment in Tehran, Iran and thalassemia centres, the "PRECEDE (Predisposing, Reinforcing and Enabling Constructs in Educational Diagnosis and Evaluation) model" was utilised to predict Quality of Life (QoL). The collected data included socio-economic and demographic details, the Persian version of the six standardized questionnaires for measuring some of the factors of QoL in TDT patients based on the variables in the PRECEDE model, and a researcher-developed questionnaire to assess patients' familiarity with health and quality of life-enhancing behaviours and their enabling factors. Structural equation modelling with maximum likelihood estimation was done with "AMOS 23.0". Self-efficacy, anxiety-depression, knowledge, enabling factors, and reinforcing factors and perceived barriers, all correlated with QoL ($p=0.001$). Anxiety, depression, and perceived barriers were negative predictors of QoL in TDT patients, while a healthy lifestyle was found to be positive (Maheri et al. 2019).

A structural equation model with "AMOS 24.0" and maximum likelihood estimation was used to test the relationships between various coping mechanisms, uncertainty and hope with resilience among 312 adolescents and young adults with beta-thalassemia major in South Iran. According to the findings, resilience had a negative relationship with uncertainty and defensive coping while positively related with social support and hope. Social support and resilience were partly mediated through courageous coping (Rambod et al. 2023).

Multilevel Tests

Data obtained from a research design where observations from the same cluster under vari-

ous experimental conditions should be analysed using multilevel analysis, which is used to test the generalisability of the experimental effect across diverse settings and to pinpoint the cause of variation related to clusters in experimental effect, in addition to ensuring accurate statistical interpretation of the overall experimental effect (Aarts et al. 2015).

A "thermogravimetry-chemometrics (TGA/Chemometrics) test" for early thalassemia and sickle cell disease diagnosis was created from 30 microlitres of whole blood samples from 235 Romans, which had 120 healthy donors and 115 haemoglobinopathies patients. A multilevel test was created in two stages wherein Sickle Cell Anaemia (SCA) and thalassemia patients' whole blood samples were compared to healthy individuals using thermogravimetry coupled with chemometrics methods based on Principal Components Analysis were used to develop a prediction model, improve thermogravimetric profile correlation, and identify differences between healthy subjects and patients. Sickle Cell Disease (SCD) and thalassemia were separated in the second stage of optimisation by identifying and categorising anaemia. Principal Component Analysis (PCA) was used to evaluate sample separation. Partial Least Squares-Discriminant Analysis (PLS-DA) was used to calculate the final multilevel test's prediction model. "Pyris" recorded thermogravimetric data, and "Unscrambler" analysed it statistically. Healthy, SCA, and thalassemic samples were distinguished with high specificity and accuracy by the model. Before mean centreing pre-treatment, the first derivative of the thermogravimetric curves improved the model's outcome prediction. Blood samples were compared to healthy people's thermogravimetric profiles to find differences. A highly accurate and specific prediction model for anaemia and SCA/thalassemia was developed using PLS-DA. Second-level tests using molecular analysis of globin genes confirmed the mutations (Risoluti et al. 2020).

Bayesian Models

Bayesian statistics, based on Bayes' theorem, updates statistical model parameters using observable data. A likelihood function calculates the posterior distribution from the prior distribution, which represents background information, and observational data. The posterior can predict future events (van de Schoot et al. 2021).

To investigate the genetic relationships between single nucleotide polymorphisms (SNPs) and foetal haemoglobin (HbF) concentration in sickle cell anaemia patients, a novel Bayesian approach was used to model foetal haemoglobin concentration as a continuous variable and test the associations of SNPs with HbF in two independent groups of sickle cell anaemia patients of <24 years and ≥24 years. Patients having sickle cell anaemia were genotyped for SNPs in the cluster that resembles the beta-globin gene, quantitative trait loci (QTL) previously linked to foetal haemoglobin (HbF) concentration on chromosomes 6q, 8q, and Xp, and candidate genes that may affect HbF levels. HbF levels in patients below 24 years were correlated with five SNPs in TOX, two in the β-globin gene-like cluster, two in the Xp QTL, and one in chromosome 15q22. The association between four additional SNPs on chromosome 15q22 and HbF levels was found only in the larger patient group. In patients below 24 years, additional genes, such as those involved in nitric oxide metabolism, were linked to HbF levels. The results suggest that a number of genes may regulate how quickly HbF levels fall and then stabilise in individuals with sickle cell anaemia (Sebastiani et al. 2008).

A haematological survey was conducted on 5,045 blood donors at the Bhopal Memorial Hospital and Research Centre in central India from August 2007 to July 2009 to determine the prevalence of β-thalassemia trait. Using "Robust Bayesian Methods", the prevalence of β-thalassemia trait in the study population was found to be 9.59 percent (95% CI: 78-10.4%) (Chatterjee et al. 2010).

To determine the number and geographic distribution of Transfusion Dependent Thalassemia (TDT) patients in various counties of Fars Province in Southern Iran, a "Bayesian Spatial Analysis" on epidemiologic data was conducted. The data registry for this cross-sectional study was provided by the Organisation of Special Diseases of Shiraz University of Medical Sciences in Shiraz, Fars Province, Iran. It contained data on all TDT patients identified between 1955 and 2018 in all counties in the province. The Besag, York, and Mollie (BYM) model was used for mapping. The "Besag, York, and

Mollie (BYM) model" was one of the most popular hierarchical Bayesian models for mapping of disease that includes random effects resulting from unstructured and spatially structured heterogeneity into the log-linear model for the relative risk. A total of 1,581 TDT patients, including 271 patients with beta-thalassemia intermedia and 1,310 patients with beta-thalassemia major, were included. The overall level of relative risk for TDT, major thalassemia, and intermediate thalassemia was 0.76 (SD: 0.07, 95% CI: 0.61, 0.91), 0.78 (SD: 0.08, 95% CI: 0.62, 0.94), and 0.56 (SD: 0.10, 95% CI: 0.38, 0.75), respectively. The highest relative risk for TDT and beta-thalassemia major was found in Larestan County (in the Southern Fars Province). The county with the highest relative risk for β-thalassemia intermedia was Shiraz (Centre of Fars Province). Contrarily, whereas the one with the lowest relative risk was Abadeh in the Northern Fars Province (Haghpanah et al. 2021).

A "Bayesian tree-based method" was suggested for the discrimination using haematological parameters in order to establish a differential diagnosis method for separating beta-thalassemia trait from iron deficiency anaemia. The study used "Bayesian Logit Treed (BLTREED) and Classification And Regression Trees (CART)" models for differential diagnosis and included 907 patients with either beta-thalassemia trait (BTT) or iron deficiency anaemia (IDA), identified as the primary predictor for diagnostic discrimination is mean corpuscular volume (MCV). BLTREED performed better in terms of the area under the curve (AUC), while CART demonstrated higher sensitivity and negative predictive value. In comparison to CART, the BLTREED model demonstrated superior diagnostic accuracy for differentiating Thalassemia Trait (TT) from IDA and had a more interpretable tree size. The accuracy in differentiating between TT and IDA was demonstrated by the models' high true positive rate, true negative rate, positive predictive value, negative predictive value, and Youden's Index values in their predictive performance, which included BLTREED, CART1, and CART2. With an AUC of 0.9, indicating excellent differentiation, all three classification tree algorithms (BLTREED, CART1, and CART2) had good diagnostic accuracy for differentiating the patients (Jahangiri et al. 2021).

Genotyping

Genotyping is the process of determining an individual's genotype, or genetic makeup. Genotyping has applications in many domains, such as forensic investigations, research on humans and animals, risk assessment, cause identification for common diseases, and genealogy. Identifying the time correlation between exposure and result is a significant constraint in numerous epidemiological and clinical observational research projects. Studies on genetic associations have been helpful in resolving this problem. Using the ideas of Mendelian randomisation, genetic variants that change exposures are used to determine causal effects, producing a result that can be used to prevent potentially confounding relationships (Kockum et al. 2023).

In a cohort of 20,222 people, a study conducted in Southern China validated a Next-Generation Sequencing (NGS) assay for carrier testing of multiple genetic defects. Raw fastq data from the Illumina Pipeline software programme was processed via an internal pipeline named Gaea, which involved several bioinformatic analysis steps. Data filtering was done using an internal Perl script called "GaeaFastqQC" to remove low-quality reads or reads from sequencing adaptors. The map to the human reference genome (UCSC build hg19) was read using the BWA-based programme "GaeaBwaStreaming". Post-processing on the mapped reads involved duplicate marking, indel realignment, and base quality score recalibration. Using the UnifiedGenotyper algorithm, "GaeaGenotyper", a GATK-based programme, was used to identify Single Nucleotide Variants (SNVs): (missense, splice, nonsense, synonymous variants, etc.) and indels. To find structural variations, two different programmes were used, namely, "Breakpointer" and "BatCNV". These programmes employed different algorithms to pinpoint the precise breakpoints of Copy Number Variants (CNVs) as well as large gene deletions and duplications. Variants were interpreted in accordance with the guidelines of the American College of Medical Genetics (ACMG) and the literature, and then classified as pathogenic, likely pathogenic, of unknown clinical significance, likely benign, or

benign. In a cohort of 10,111 couples undergoing carrier testing, the NGS assay identified 186 at-risk couples, compared to 151 by standard methods. According to the study, next-generation sequencing (NGS) offers new insights into the genetic basis of hemoglobinopathies and may be a more accurate and efficient method of haemoglobinopathy screening and diagnosis (Shang et al. 2017).

A cross-sectional study of descriptive nature was conducted from March 2011 to July 2013 to examine the "spectrum of β-thalassemia mutations and sickle cell anaemia haplotypes" in Beja tribes and other minor groups in Port Sudan, Sudan. 209 subjects were tested for ²-thalassemia mutation and sickle cell haplotype, in which 29 (13.87%) had the -88(C→T) β-thalassemia mutation, while 27 (12.91%) had sickle cell anaemia, with 15 (55.6%) being heterozygous AS and 12 (44.4%) being homozygous SS. The phylogenetic tree suggested that tribal groups independently obtained the mutation. The selected samples were sequenced for sickle cell haemoglobin, where, "Hind III PCR" products showed no significant similarity to reference genomic strands on "BLAST" (Gibreel et al. 2018).

The number of articles identified for each of the modelling techniques is given in the Table 1:

Table 1: Modelling techniques and number of articles identified

S. No.	Technique	No. of articles
1	*Multivariate Modelling Techniques*	*12*
	Logistic Regression	7
	Logistic Regression with Longitudinal Model	1
	Multilevel Tests	1
	Structural Equation Modelling (SEM)	3
2	Genetic Risk Scoring (GRS) and Prediction Models	2
3	Methods Based on Machine Learning Techniques	3
4	Decision-tree Analysis Based Model	1
4.1	Decision-tree along with Federated Learning (FL) Techniques	1
5	*Bayesian Model*	*4*
	Bayesian Approach/Robust Bayesian Methods	2
	Bayesian Spatial Analysis	1
	Bayesian tree-based method	1
6	Genotyping	2
	TOTAL	25

Research Gaps Identified

From the literatures reviewed, it was observed that the statistical techniques and models that were utilised for prediction of determinants of Haemoglobinopathies were scarce in studies from India and the tribes. Most of the techniques used dealt with risk factors and determinants, whereas, those involving genes and variants were predicted using machine learning models. Studies on Haemoglobinopathies in tribes regarding modelling and prediction are needed to identify the role of genetic variants and other determinants in causing the disease.

CONCLUSION

The analysis of these articles has provided knowledge about the prevalence of haemoglobinopathies and the related disorders. This narrative review found that logistic regression technique was used extensively to build a model with significant predictors of Haemoglobinopathies, followed by Bayesian methods, which used approaches based on decision-trees also. Haemoglobinopathies have received less attention in the aspect of research in India and among the tribes. Apart from social, demographic, cultural and geographic factors, genetic factors also play a major role in causing Haemoglobinopathies.

RECOMMENDATIONS

More number of studies are required in India, especially in Tamil Nadu for building prediction-based models. The literature reviewed has also suggested the validation of the study findings based on model building and prediction in different study settings. Such studies will help the Government and other such organisations to take necessary action to prevent the recurrence of such diseases in the future.

LIMITATIONS

Although the results of a few studies that used statistical methods to predict the likelihood of Haemoglobinopathies and create prediction-based models have been discussed in this narrative review, more research in a wide range of contexts is required to fully understand the practical application of these methods.

ACKNOWLEDGEMENT

The author expresses her sincere gratitude to her Research Supervisor, her parents, family members and friends and well-wishers for their continuous guidance and support.

DECLARATION OF CONFLICTING INTERESTS

The author(s) declare no conflict of interest.

FUNDING

The author(s) did not receive any source of funding.

REFERENCES

Aarts E, Dolan CV, Verhage M, van der Sluis S 2015. Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. *BMC Neuroscience*, 16(1). https://doi.org/10.1186/s12868-015-0228-5

Abaker Gibreel MO, El Saeed Elkarsani M, Mohammed Munsour M, El Taher HB 2018. Genotyping of individuals with hemoglobinopathies in Beja Tribes and other minor groups in Port Sudan, Eastern Sudan. *Medical Laboratory Journal*, 12(6): 40-45.

Ampolu Kavitha, Vyas Kavita, Mukilarasi 2023. Incidence of hemoglobinopathies in South Indian population – 4 year study. *Annals of Pathology and Laboratory Medicine*, 10(5): A47-52.

Apidechkul T, Yeemard F, Chomchoei C et al. 2021. Epidemiology of thalassemia among the hill tribe population in Thailand. *PLOS ONE*, 16(2): e0246736.

Belbasis L, Panagiotou OA 2022. Reproducibility of prediction models in health services research. *BMC Research Notes*, 15(1). https://doi.org/10.1186/s13104-022-06082-4

Beran TN, Violato C 2010. Structural equation modeling in medical research: A primer. *BMC Research Notes*, 3(1). https://doi.org/10.1186/1756-0500-3-267

Chatterjee N, Amit Kumar Mishra, Soni R et al. 2010. Bayesian estimates of the prevalence of β-Thalassemia trait in voluntary blood donors of Central India: A survey. *Hemoglobin*, 34(6): 548-560.

Danjou F, Francavilla M, Anni F et al. 2015. A genetic score for the prediction of beta-thalassemia severity. *Haematologica*, 100(4): 452-457.

Farooq M, Younas H n.d. Beta Thalassemia Carriers Detection Empowered Federated Learning. From <https://arxiv.org/ftp/arxiv/papers/2306/2306.01818.pdf> (Retrieved on 28 November 2023).

Ganesh B, Rajakumar T, Acharya S et al. 2021. Particularly vulnerable tribal groups of Tamil Nadu, India: A sociocultural anthropological review. *Indian Journal of Public Health*, 65(4): 403.

Ganesh B, Thangarasu Rajakumar, Subhendu Kumar Acharya et al. 2023. Prevalence of Haemoglobinopathies Among Malayali Tribes of Jawadhu Hills, Tiruvannamalai District, Tamil Nadu, India: A Community-Based Cross-Sectional Study. Preprints with the LANCET. From <https://doi.org/10.2139/ssrn.4628188> (Retrieved on 27 November 2023).

Garcia TP, Marder K 2017. Statistical approaches to longitudinal data analysis in neurodegenerative diseases: Huntington's Disease as a model. *Current Neurology and Neuroscience Reports*, 17(2). https://doi.org/10.1007/s11910-017-0723-4

Ghosh K, Colah RB, Mukherjee MB 2015. Haemoglobinopathies in tribal populations of India. *The Indian Journal of Medical Research*, 141(5): 505–508.

Goonasekera HW, Paththinige CS, Dissanayake VHW 2018. Population screening for hemoglobinopathies. *Annual Review of Genomics and Human Genetics*, 19(1): 355–380.

Harteveld CL, Achour A, Arkesteijn SJG et al. 2022. The hemoglobinopathies, molecular disease mechanisms and diagnostics. *International Journal of Laboratory Hematology*, 44(S1): 28-36.

Hassanzadeh J, Mirahmadizadeh A, Karimi M et al. 2016. Delay in diagnosis of hemoglobulinopathies (Thalassemia, Sickle Cell Anemia): A need for management of thalassemia programs. *Iranian Journal of Pediatrics*, 27(2). https://doi.org/10.5812/ijp.6740

Iannotti LL, Delnatus JR, Odom et al. 2015. Determinants of anemia and hemoglobin concentration in Haitian school-aged children. *The American Journal of Tropical Medicine and Hygiene*, 93(5): 1092-1098.

India - A-11 Appendix: District wise scheduled tribe population (Appendix), Tamil Nadu - 2011. Censusindia.gov.in. From <https://censusindia.gov.in/nada/index.php/catalog/43032> (Retrieved on 25 January 2023).

Islam MR, Moinuddin M, Ahmed A et al. 2021. Association of sickle cell disease with anthropometric indices among under-five children: evidence from 2018 Nigeria Demographic and Health Survey. *BMC Medicine*, 19(1). https://doi.org/10.1186/s12916-020-01879-1

Jahangiri M, Rahim F, Saki N et al. 2021. Application of Bayesian Decision Tree in hematology research: Differential diagnosis of β-Thalassemia trait from iron deficiency anemia. *Computational and Mathematical Methods in Medicine*, 2021: 1–10.

Janes H, Pepe MS, Gu W 2008. Assessing the value of risk predictions by using risk stratification tables. *Annals of Internal Medicine*, 149(10): 751–760.

Jyoti Ghosal, Bal M, Manoranjan Ranjit et al. 2023. To what extent classic socio-economic determinants explain trends of anaemia in tribal and non-tribal women of reproductive age in India? Findings from four National Family Heath Surveys (1998-2021). *BMC Public Health*, 23(1). https://doi.org/10.1186/s12889-023-15838-x

Kockum I, Huang J, Pernilla Stridh 2023. Overview of genotyping technologies and methods. *Current Protocols*, 3(4). https://doi.org/10.1002/cpz1.727

Kohne E 2011. Hemoglobinopathies. *Deutsches Aerzteblatt Online*, 108(31-32). https://doi.org/10.3238/arztebl.2011.0532

Kutlar F 2007. Diagnostic approach to hemoglobinopathies. *Hemoglobin*, 31(2): 243–250.

Laengsri V, Shoombuatong W, Adirojananon W et al. 2019. ThalPred: A web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia. *BMC Medical Informatics and Decision Making*, 19(1). https://doi.org/10.1186/s12911-019-0929-2

Maheri M, Rohban A, Sadeghi R et al. 2019. Predictors of quality of life in transfusion-dependent thalassemia patients based on the PRECEDE Model: A Structural Equation Modeling Approach. *Journal of Epidemiology and Global Health*. https://doi.org/10.2991/jegh.k.191001.001

Mahesh B 2018. Machine learning algorithms - A review. *International Journal of Science and Research (IJSR)*, 381-386.

Masoume Rambod, Saeed Hamidizadeh, Mohammad Rafi Bazrafshan 2023. Risk and protective factors for resilience among adolescents and young adults with beta-thalassemia major. *BMC Psychology*, 11(1). https://doi.org/10.1186/s40359-023-01268-2

Milton JN, Gordeuk VR, Taylor JG et al. 2014. Prediction of fetal hemoglobin in sickle cell anemia using an ensemble of Genetic Risk Prediction Models. *Circulation-Cardiovascular Genetics*, 7(2): 110-115.

Mohanty D, Mukherjee MB, Colah RB et al. 2013. Spectrum of hemoglobinopathies among the primitive tribes. *Asia Pacific Journal of Public Health*, 27(2): NP562–NP571.

Mohanty SS, Parihar S, Huda RK 2022. Prevalence of sickle cell anemia, β-thalassemia and glucose-6-phosphate dehydrogenase deficiency among the tribal population residing in the Aravali hills of Sirohi region of Rajasthan state. *Clinical Epidemiology and Global Health*, 13: 100916.

Mondal S, Mandal S 2016. Prevalence of thalassemia and hemoglobinopathy in eastern India: A 10-year high-performance liquid chromatography study of 119,336 cases. *Asian Journal of Transfusion Science*, 10(1): 105.

Okandeji AA, Odeyinka OF, Sogbesan AA et al. 2022. A comparative analysis of haemoglobin variants using machine learning algorithms. *Nigerian Journal of Technology (NIJOTECH)*, 41(4): 789-796.

Onyeneho NG, Ozumba BC, Subramanian SV 2019. Determinants of childhood anemia in India. *Scientific Reports*, 9(1): 16540.

Pant L, Kalita D, Singh S 2014. Detection of abnormal hemoglobin variants by HPLC Method: Common problems with suggested solutions. *International Scholarly Research Notices*, 2014: 1-10.

Pattanshetty S, Chandrasekaran V, Kamath R 2013. Prevalence of anemia among tribal women of reproductive age in Udupi Taluk, Karnataka. *Journal of Family Medicine and Primary Care*, 2(4): 345.

Platania S, Gruttadauria S, Citelli G 2017. Associations of Thalassemia Major and satisfaction with quality of life: The mediating effect of social support. *Health Psychology Open*, 4(2): 205510291774205.

Regional Desk Review Of Haemoglobinopathies With An Emphasis On Thalassaemia And Accessibility And Availability Of Safe Blood And Blood Products As Per These Patients' Requirement In South-East Asia Under Universal Health Coverage. n.d. From <www. who. int. https://www.who.int/publications/i/item/9789290228516> (Retrieved on 28 January 2023).

Risoluti R, Caprari P, Gullifa G 2020. An Innovative Multilevel Test for hemoglobinopathies: TGA/Chemometrics Simultaneously identifies and classifies sickle cell disease from thalassemia. *Frontiers in Molecular Biosciences*, 7.

Rustam F, Ashraf I, Jabbar S 2022. Prediction of beta-Thalassemia carriers using complete blood count features. *Scientific Reports*, 12(1). https://doi.org/10.1038/s41598-022-22011-8

Samanta A, Chaudhuri PK, Das U et al. 2021. A new approach based on erythrocyte indices to exclude normal populations from chromatography based thalassemia screening programs with very high fidelity. *Iranian Journal of Blood and Cancer*, 13(4): 107-118.

Saxena A, Phadke SR 2002. Thalassaemia control by carrier screening: The Indian scenario. *Current Science*, 83(3): 291-295.

Sebastiani P, Wang L, Nolan VG 2008. Fetal hemoglobin in sickle cell anemia: Bayesian modeling of genetic associations. *American Journal of Hematology*, 83(3): 189-195.

Selvaraj B, Subramanian G, Ramanathan SK et al. 2022. Studies on molecular spectrum of beta thalassemia among residents of Chennai. *AIMS Molecular Science*, 9(3): 107-135.

Sengupta M 2008. Thalassemia among the tribal communities of India. *The Internet Journal of Biological Anthropology*, 1(2). https://doi.org/10.5580/9a1

Sezaneh Haghpanah, Naeimehossadat Asmarian, Omid Reza Zekavat et al. 2021. Bayesian spatial modeling of transfusion-dependent β-thalassemia incidence rate in Fars Province, Southern Iran. *Spatial and Spatio-Temporal Epidemiology*, 36: 100389-100389.

Shang X, Peng Z, Ye Y et al. 2017. Rapid targeted next-generation sequencing platform for molecular screening and clinical genotyping in subjects with hemoglobinopathies. *EBioMedicine*, 23: 150-159.

Song YY, Lu Y 2015. Biostatistics in psychiatry (26) Decision tree methods: Applications for classification and prediction. *Shanghai Arch Psychiatry*, 27(2): 130–135.

Sperandei S 2014. Understanding Logistic Regression Analysis. *Biochemia Medica*, 24(1): 12-18.

Stoltzfus JC 2011. Logistic Regression: A brief primer. *Academic Emergency Medicine*, 18(10): 1099-1104.

Teli AB, Deori R, Saikia SP 2016. Haemoglobinopathies and β-Thalassaemia among the Tribals Working in the Tea Gardens of Assam, India. *Journal Of Clinical And Diagnostic Research*, 10(12). https://doi.org/10.7860/jcdr/2016/22010.9002

TNHSP. n.d. Tnhsp.org. From <https://tnhsp.org/tnhsp/tribal-healthcare.php> (Retrieved on 4 October 2023).

Toma M, Wei OC 2023. Predictive modeling in medicine. *Encyclopedia*, 3(2): 590–601.

Tribal Welfare Department - Tamilnadu. n.d.-a. From <www.tntribalwelfare.tn.gov.in. https://www.tntribalwelfare.tn.gov.in/Tribes-of-Tamilnadu.php> (Retrieved on 15 October 2023).

Tribal Welfare Department - Tamilnadu. n.d.-b. From <www.tntribalwelfare.tn.gov.in.https://www tntribalwelfare. tn.gov.in/tribal_research_ centre.php> (Retrieved on 15 October 2023).

van de Schoot R, Depaoli S, King R et al. 2021. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1): 1-26.

Verma IC, Saxena R, Kohli S 2012. Hemoglobinopathies in India-Clinical and laboratory aspects. *Clinics in Laboratory Medicine*, 32(2): 249-262.

Wendt AS, Brintrup J, Waid, JL 2023. Thalassemia and hemoglobinopathy prevalence in a community-based sample in Sylhet, Bangladesh. *Orphanet Journal of Rare Diseases*, 18(1): 192.

Williams TN, Weatherall DJ 2012. World distribution, population genetics, and health burden of the hemoglobinopathies. *Cold Spring Harbor Perspectives in Medicine*, 2(9): a011692–a011692.

Paper received for publication in October, 2023
Paper accepted for publication in Novembe, 2023