MINISTRY OF EDUCATION AND TRAINING MINISTRY OF HEALTH
NATIONAL INSTITUDE OF MALARIOLOGY PARASITOLOGY AND ENTOMOLOGY

CLINICAL EPIDEMIOLOGY AND THALASSEMIA GENE MUTATIONS CAUSING CONGENITAL ANEMIA IN COUPLES UNDERGOING EXAMINATION AT THE NATIONAL HOSPITAL OF OBSTETRICS AND GYNECOLOGY, RESULTS OF ARTIFICIAL INTELLIGENCE APPLICATION IN PRENATAL SCREENING

Chuyên ngành: Dịch tễ học
Mã số: 9720117

SUMMARY OF DOCTORAL DISSERTATION IN MEDIDINE

Hà Nội - 2023

The work was completed at the postgraduate training
National institute of Malariology Parasitology and Entomology

Scientific Supervisors

1. Assoc. Prof. Dr. Nguyen Tuan Hung
2. Assoc. Prof. Dr. Nguyen Thi Trang

Argument 1: Assoc. Prof. Dr. Doan Huy Hau
Affiliation: Vietnam Military Medical Academy
Argument 2: Assoc. Prof. Dr. Le Xuan Hung
Affiliation: Hanoi University of Public Health
Argument 3: Prof. Dr. Tran Quoc Kham
Affiliation: Ministry of Health

The dissertation has been successfully defended before the Institute-level Dissertation Evaluation Committee, in a meeting held at the National institude of Malariology Parasitology and Entomology, at 8:00 AM on February 01, 2024.

The dissertation can be found at:

- National ibrary
- National institute of Malariology Parasitology and Entomology Library

INTRODUCTION

Congenital thalassemia, a monogenic disorder, is the most common genetic disorder worldwide. In 2018, approximately 5.2% of the population carried the thalassemia gene, with 1.1% of couples at risk of having offspring with hemoglobin disorders and 2.7/1000 cases of affected pregnancies[1]. In Vietnam, around 7-10% of the population carries the gene, with 20,000 individuals affected by severe thalassemia, resulting in over 2,000 affected births annually[2]. Notably, the disease exhibits geographical and ethnic distribution patterns.

The cause of thalassemia is mutations in the genes governing the synthesis of globin chains, leading to an imbalance in globin chain types and abnormalities in hemoglobin, reducing the quality of red blood cells. Thalassemia is preventable and treatable. To prevent the disease, screening is essential to accurately identify the mutation type and clinical characteristics of the disease. With the rapid development of artificial intelligence, clinical decision support systems have garnered considerable research interest. In thalassemia gene carrier screening, artificial intelligence contributes to improving the quality and accuracy of detection, classification, and prenatal diagnosis through simultaneous analysis of multiple hematological indices, combined with family history and obstetric history. Additionally, it enhances screening value by analyzing big data, automating result reporting, and reducing the impact of the expertise and experience of physicians on result evaluation.

Based on these observations, we conducted a study titled "C**linical Epidemiology and Thalassemia gene mutations causing congenital anemia in couples undergoing examination at the national hospital of obstetrics and gynecology**, **results of artificial intelligence application in prenatal screening"** with the following objectives:

1. ***Describe epidemiological and clinical characteristics, Thalassemia gene mutations causing congenital anemia in couples undergoing examination and prenatal csreening at the National hospital of Obstetrics and Gynecology (NHOG) during the period 2012-2022.***
2. ***Evaluate the results of applying expert system software and machine learning software in screening for congenital anemia at the National hospital of Obstetrics.***

RESEARCH CONTENT

1. ***Describe certain clinical epidemiological characteristics in couples undergoing examination and prenatal screening at the National hospital of Obstetrics and Gynecology during the period 2012-2022;***
2. ***Commenting on Thalassemia gene on gene mutations related to congenital anemia in the patient group;***
3. ***Evaluating the sensitivity and specificity of expert system software and machine learning software in screening for congenital hemolytic disease.***

RESEARCH SCOPE

The research focuses on analyzing and evaluating clinical epidemiological characteristics, Thalassemia gene mutations causing congenital anemia, and the application of two artificial intelligence software, namely expert system software and machine learning software, in screening for Thalassemia gene mutations in pregnant women and their husbands examined at the National hospital of Obstetrics and Gynecology during the period from 2012 to 2022.

NEW CONTRIBUTIONS OF THE THESIS

1. Provides a detailed description of the clinical epidemiological characteristics and distribution of congenital hemolysis gene variants within the pregnant community and their partners at the National hospital of Obstetrics and Gynecology from 2012 to 2022.

2. Identify common genetic characteristics, contributing to a better understanding of the genetic variant landscape and aiding in healthcare decisions for pregnant women.

3. Introduce, for the first time in Vietnam, the utilization of artificial intelligence software in screening individuals carrying congenital hemolysis genes. Compare the sensitivity and specificity of expert knowledge-based systems and machine learning software, aiming to delineate strengths and weaknesses for optimizing the screening process and making effective healthcare decisions for pregnant women and their fetuses.

SCIENTIFIC AND PRACTICAL SIGNFICANCE OF THE THESIS

Scientific significance:

- Provides detailed information on the prevalence and distribution of Thalassemia gene variants in pregnant women and their partners undergoing prenatal screening, expanding knowledge about the genetic carrier status and incidence of congenital hemolysis in Vietnam.
- Describes the clinical epidemiological characteristics of participating couples in the screening, laying the foundation for further research on fetal and maternal health.

Practical value:

- Healthcare Decision Support: Offers detailed information on the congenital hemolysis gene variant status of pregnant women and their partners, aiding healthcare decisions, particularly during the prenatal period.
- Optimization of Screening Process: Evaluates the outcomes of artificial intelligence software, contributing to the optimization of the screening process and guiding the selection of the most suitable technology in practical settings.

THESIS STRUCTURE

The thesis, comprising 115 pages, is organized into the following sections: Introduction (2 pages); Overview (32 pages); Research objects and methods (16 pages); Results (32 pages); Discussion (30 pages); Conclusion (2 pages); Recommendation (1 page).

The thesis includes 38 tables, 12 figures and references to 120 sources.

Chapter 1: OVERVIEW

1.1. Epidemiology of Thalassemia disease in the world and Vietnam

Thalassemia disease is prevalent in most countries around the world, especially the Mediterranean region, North Africa, India, China and Southeast Asia, including Vietnam[3]. α-thalassemia is the most common mutation among Thalassemia types and is found in almost all populations, particularly in individuals of Asian descent, certain regions of China, and Southeast Asia, with a carrier rate of approximately 22.6%[4]. The carrier rate of β-thalassemia is around 1.5% of the world's population (approximately 80 - 90 million carriers). This rate is notably high in many countries in the Mediterranean region, such as Saudi Arabia (10%), Greece (8%), and Italy (4.8%)[5,6].

In Vietnam, the carrier rate of Thalassemia varies from 3.5% to 28%, depending on ethnicities and geographical regions[7–9]. In 2017, the overall carrier

rate across different ethnicities nationwide was 13.8%. Specifically, the highest α^0-thalassemia carrier rates were observed in certain ethnic groups such as Xinh Mun (18.5%), Lu (16.4%), and Muong (16.1%)[10]. The carrier rate of α^+-thalassemia was high in Raglay (85.2%), Ta Oi (63.3%), and Bru Van Kieu (62%)[10]. β-thalassemia is more prevalent among ethnic minorities in Northern Vietnam, such as the Muong ethnic group. The carrier rate of β^0-thalassemia is elevated in some ethnic groups like Giay (11.5%), San Chay (9.8%), and La Ha (9.5%)[10].

1.2. Application of Artificial Intelligence in Prenatal Thalassemia Screening

There are two artificial intelligence software applications utilized in prenatal Thalssemia screening: machine learning software and expert knowledge-based software.

1.2.1. Expert Knowledge-Based software application

In prenatal Thalassemia screening, expert knowledge-based software is constructed from a knowledge base containing rule sets in IF-THEN format based on hematological and blood biochemical indices (blood formula, ferritin, serum iron, and hemoglobin component analysis). Experts provide the knowledge system with insights into the correlation between these indices, reflecting the high or low risk of carrying Thalassemia genes for the pregnant woman, her spouse, and the fetus. Consequently, the system can draw conclusions when fed with new data from actual patients.

1.2.2. Machine Learning software application

The role of the machine learning software algorithm is to build a mathematical model from a dataset containing both input and desired output. In this context, input data and hematological and blood biochemical indices resemble those in the expert knowledge-based software. The output data indicate the high or low risk of carrying Thalassemia genes for the pregnant woman or her spouse.

1.3. Evaluation of the value of two Artificial Intelligence applications in prenatal Thalassemia screening

Each system, before being implemented, requires meticulous control and evaluation to ensure accuracy and optimal effectiveness.

Table 1.1. Conclusion of the Thalassemia screening system using Artificial Intelligence

		Labeling results	
		Mutation	No mutation
Conclusion of the system	High risk	True positive (TP)	False positive (FP)
	Low risk	False negative (FN)	True negative (TN)

To evaluate the value of the system, it is necessary to rely on indices including accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. These indices are calculated using the following formulas:

General accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$

Sensitivity = $\frac{TP}{TP + FN}$

Specificity = $\frac{TN}{TN + FP}$

Positive predictive value = $\frac{TP}{TP+FP}$

Negative predictive value = $\frac{TN}{TN+FN}$

Chapter 2: OBJECT AND METHOD OF RESEARCH

2.1. Research subjects

- Pregnant women and their husbands attending prenatal care and screening at the National hospital of Obstetrics and Gynecology at any gestational age, without distinction based on the order of pregnancy.
- Pregnant women and their husbands with results from peripheral blood cell analysis or a family history of Thalassemia/detection of Thalassemia gene carriers.
- Pregnant women and their husbands at high risk (anemia test results/family history related to Thalassemia) undergo hemoglobin electrophoresis, genetic testing for conclusive results. Cases with low risk or no detected gene mutations at the current time need postnatal assessment.

2.2. Location and time of research

- *Research location:* National hospital of Obstetrics and Gynecology
- *Research period:* From 2012 to 2022

2.3. Study design

Objective 1: Conduct a study describing a series of cases involving couples undergoing thalassemia screening using both retrospective (2012-2020) and prospective (2020-2022) approaches.

Objective 2: Conduct experimental research on the group undergoing examination and treatment at the Central Obstetrics Hospital, assessing screening outcomes using two Affiliate Intelligence software programs.

2.4. Sample size and sampling method

- The sample size for objective 1 is calculated using the sample size formula to estimate a proportion in the population: $n = Z^2_{(1-\frac{\alpha}{2})} \times \frac{p(1-p)}{(p\times\varepsilon)^2} = Z^2_{(1-\frac{\alpha}{2})} \times \frac{(1-p)}{p\times\varepsilon^2}$

In fact, 2,584 subjects, 1,292 couples, were collected for prenatal check-ups at the National hospital of Obstetrics and Gynecology that met the inclusion and exclusion criteria.

- Sample size for objective 2 is calculated using the formula:

TP, FN, TN, FP are the numbers of true positives, false negatives, true negatives and false positives, respectively.

$$TP + FN = \frac{Z_\alpha^2 \times p_{se} \times (1-p_{se})}{\omega^2} \quad ; \quad TN + FP = \frac{\boldsymbol{Z_\alpha^2 \times p_{sp} \times (1-p_{sp})}}{\boldsymbol{\omega^2}}$$

Sample size to evaluate sensitivity is calculated according to the formula: $n_1 = \frac{TP+FN}{p}$

Sample size to evaluate specificity was calculated according to the formula: $n_2 = \frac{TN+FP}{1-p}$

Substituting the above two formulas, the result is $n_1 = 207$ and $n_2 = 43$. In fact, the sample size collected is 244 subjects.

2.5. Research content

The selection of medical records for pregnant women and their husbands undergoing Thalassemia screening at the Central Obstetrics Hospital from 2012 to 2022 will adhere to specific criteria. Appropriate selection and exclusion criteria will involve hematological, blood biochemical, and hemoglobin electrophoresis tests, including the following indices: RBC, HGB, HCT, MCH, MCV, MCHC, RDW, serum ferritin, serum iron, hemoglobin electrophoresis, and the Thalassemia history of pregnant women.

Subjects identified as having a high risk of carrying Thalassemia genes will undergo genetic testing using the Strip Assay technique to analyze genetic characteristics. Subsequently, selected research subjects will be input into both the expert system software and machine learning software for result analysis. This process aims to determine the risk of Thalassemia, Thalassemia subtype, and draw conclusions about the hematological status.

2.6. Research variables and indices

- Target 1 variables and indices:

+ General information of wives and husbands: Age, gestational age, ethnicity.

+ Group of variables for blood indices: RBC, HGB, HCT, MCV, MCH, MCHC, RDW.

+ Group of variables for biochemical tests: Serum iron, serum ferritin.

+ Group of variables for hemoglobin electrophoresis results: HbA1, HbA2, HbE, HbF, Hb Bart's, Other Hb.

+ Group of variables determining thalassemia genes using Strip Assay software.

- Target 2 variables and indices: Risk of being affected, conclusion of carrying α or β genes provided by the software, conclusion of anemia provided by the software, classification of red blood cell size, values of sensitivity, specificity, positive predictive value, negative predictive value, and accuracy of the software.

2.7. Research process

Step 1: Collect medical records meeting selection criteria.

Step 2: Enter data, including test results of pregnant women and their husbands, comprising RBC, HGB, HCT, MCH, MCV, MCHC, RDW, serum ferritin, serum iron, hemoglobin electrophoresis, results of determining thalassemia genes using the Strip Assay technique, and history of thalassemia, miscarriage, or family members carrying thalassemia genes.

Step 3: Use artificial intelligence software to analyze results. The AI software system used in this study includes machine learning software and expert system software developed by the research team of Prof. Dr. Tran Danh Cuong, Prof. Dr. Nguyen Thi Trang, and colleagues from the state-funded project "Research on the Application of Artificial Intelligence in Screening for Some Common Congenital Abnormalities in Vietnam" under the Industry 4.0 Program for the period 2019-2025, which was successfully tested in 2023.

Step 4: Evaluate the software based on sensitivity, specificity, and accuracy indices. Compare with the results of genetic mutation testing (gold standard) currently in use.

Chapter 3: STUDY RESULTS

3.1. Some clinical epidemiological characteristics, genetic variants of congenital thalassemia in couples attending and undergoing antenatal screening at the National hospital of Obstetrics and Gynecology during the period 2012-2022.

3.1.1. Thalassemia gene carrier rate

Prenatal screening research at the National hospital of Obstetrics and Gynecology from 2012 to 2022 included 1292 couples with the following results:

Table 3.1. Percentage of Thalassemia gene carriers screened at the NHOG (n=2584)

	Wife (n_1 = 1292)		Husband (n_2 = 1292)		Total n = 2584	
	Quantity	Rate (%)	Quantity	Rate (%)	Quantity	Rate (%)
Alpha thalassemia	230	17.8	228	17.65	458	17.72
Beta thalassemia	21	1.62	20	1.55	41	1.59
Combined α and β Thalassemia	9	0.7	8	0.62	17	0.66
Alpha/HbE	8	0.62	2	0.16	10	0.39
Beta/HbE	0	0	1	0.08	1	0.04

HbE	6	0.46	14	1.16	20	0.77
Non-genetic carrier	48	3.72	56	4.33	104	4.03
No need for genetic testing	970	75.08	963	74.54	1933	74.8
Total	1292	100	1292	100	2584	100

Results showed that among 970 pregnant women and 963 husbands identified as low-risk, not requiring diagnostic Thalassemia gene testing. For the remaining cases that underwent Thalassemia gene diagnostic testing, in both groups, the most prevalent gene carrier rate was α-thalassemia, accounting for 17.8% (230/1292) in pregnant women and 17.65% (228/1292) in husbands. Following that, the carrier rate for β-thalassemia in both pregnant women and husbands was 1.62% and 1.55%, respectively. Less common genes included those associated with a combination of α and β-thalassemia and α/HbE. Additionally, it was observed that 6 pregnant women and 14 husbands carried the HbE gene.

Table 3.2. Percentage of both spouses carrying the Thalassemia gene screened at the NHOG

Genetically modified	Quantity	Rate (%)
Alpha thalassemia	210	16.25
Beta thalassemia	20	1.5
Combined α and β Thalassemia	8	0.6
Total	238	18.35

There are 210 couples who both carry the α-thalassemia gene, accounting for 16.25%, 20 couples who both carry the β-thalassemia gene, accounting for 1.5%, and 8 couples who both carry the α and β thalassemia genes (0.6%).

Table 3.3. Distribution of α-thalassemia gene mutations in pregnant women and their husbands coming for screening at the NHOG

Subject / Mutation type	Wife		Husband	
	Quantity	Rate (%)	Quantity	Rate (%)
$-\alpha^{3.7}$	6	2.38	5	2.1
$-\alpha^{4.2}$	4	1.59	5	2.1
$--^{SEA}$	235	93.25	221	92.86
$--^{THAI}$	3	1.2	1	0.42
Anti $-\alpha^{3.7}$	2	0.79	4	1.68

α2 INT CD [T>C]	0	0	1	0.42
α2 CD125 [CTG>CCG](HbQs)	2	0.79	0	0
α2 CD142 [TAA>CAA](HbCs)	0	0	1	0.42
Other mutations	0	0	0	0
Total	252	0	238	0

Individuals may carry one or multiple α-thalassemia mutations simultaneously. There were 252 α-thalassemia mutations identified in pregnant women and 238 α-thalassemia mutations in husbands. The most common mutation in these couples was heterozygous --SEA, accounting for 93.25% in pregnant women and 92.86% of the total mutations in husbands. Following that were mutations like heterozygous $-\alpha^{3.7}$, $-\alpha^{4.2}$, α^{2} CD142 [TAA>CAA], anti $\alpha^{3.7}$, and others. Particularly, in husbands, four cases of Anti $-\alpha^{3.7}$ in heterozygous form were detected. All remaining mutation alleles were in heterozygous form.

Table 3.4. Distribution of β-thalassemia gene mutations in pregnant women and their husbands coming for screening at the NHOG

Subject / Mutation type	Wife		Husband	
	Quantity	Rate (%)	Quantity	Rate (%)
-28 [A>G]	1	3.45	0	0
Cap+1 [A>C]	1	3.45	0	0
cd 17 [A>T]	10	34.48	10	27.78
cd 26 [G>A] HbE	6	20.69	15	41.67
IVS 1.1 [G>T]	1	3.45	1	2.78
cd 41/42 [-TTCT]	5	17.24	8	22.22
cd 71/72 [+A]	3	10.34	2	5.56
cd 95 [+A]	0	3.45	0	0
IVS 2.654 [C>T]	1	3.45	0	0
Other mutations	0	0	0	0
Total	28	100	36	100

Among the various mutations detected in both spouses, the two most common mutations were CD17 [A>T], CD26 [G>A] HbE, followed by CD41/42 [+A] and CD71/72 [+A]. Mutations -28 [A>G], Cap+1 [A>C], cd 95 [+A], and IVS 2.654 [C>T] were found in the wife, with one case for each mutation type.

3.1.2. Some epidemiological, clinical, and paraclinical characteristics of people carrying the gene for congenital hemolytic disease

Table 3.5. Regional distribution of α-thalassemia mutations in pregnant women coming for screening at the NHOG (n=247)

Region / Mutation	The Red River Delta, Quantity (%)	Northeast, Quantity (%)	Northwest, Quantity (%)	North Central, Quantity (%)	Other region, Quantity (%)
$-\alpha^{3.7}$	1 (1.1)	2 (2.6)	0 (0)	2 (4.6)	0 (0)
$-\alpha^{4.2}$	3 (3.4)	1 (1.3)	0 (0)	0 (0)	0 (0)
--SEA	81 (91)	70 (92.2)	35 (100)	42 (93)	2 (100)
--THAI	2 (2.2)	1 (1.3)	0 (0)	0 (0)	0 (0)
Anti-$\alpha^{3.7}$	0 (0)	2 (2.6)	0 (0)	0 (0)	0 (0)
α2cd125[CTG>CCG]	2 (2.2)	0 (0)	0 (0)	1 (2.4)	0 (0)
Total (247)	89 (100)	76 (100)	35 (100)	45 (100)	2 (100)

The most common type of α-thalassemia mutation is -SEA, with prevalence in the Red River Delta region at 91%, Northeast region at 92.2%, Northwest region at 100%, and North Central region at 93%. The next commonly found strains are $-\alpha^{3.7}$ and $-\alpha^{4.2}$. The least common are –THAI and α^2 cd125 [CTG>CCG], each with three cases. Among these, –THAI comprises 2 cases in the Red River Delta region and 1 case in the Northeast region, while α^2 cd125 [CTG>CCG] includes 2 cases in the Red River Delta region and 1 in the North Central region. Anti-$\alpha^{3.7}$ has 2 cases, constituting 2.6%, in the Northeast region.

Table 3.6. Regional distribution of β-thalassemia mutations in pregnant women coming for screening at the NHOG (n=247)

Region / Mutation	The Red River Delta, Quantity (%)	Northeast, Quantity (%)	Northwest, Quantity (%)	North Central, Quantity (%)	Other region, Quantity (%)
-28 [A>G]	1 (10)	0 (0)	0 (0)	0 (0)	0 (0)
Cap+1 [A>C]	1 (10)	0 (0)	0 (0)	0 (0)	0 (0)

cd 17 [A>T]	2 (23.3)	5 (45.4)	0 (0)	3 (75)	0 (0)
cd 26 [G>A] HbE	2 (23.3)	1 (9.1)	2 (66.7)	1 (25)	0 (0)
IVS 1.1[G>T]	0 (0)	1 (9.1)	0 (0)	0 (0)	0 (0)
cd 41/42 [TTCT]	1 (10)	2 (18.2)	1 (33.3)	0 (0)	0 (0)
cd 71/72 [+A]	2 (23.3)	1 (9.1)	0 (0)	0 (0)	0 (0)
IVS 2.654[C>T]	0 (0)	1 (9.1)	0 (0)	0 (0)	0 (0)
Total (27)	9 (100)	11 (100)	3 (100)	4 (100)	0 (100)

In the Red River Delta region, the most common β-thalassemia mutations are cd 17 [A>T], cd 26 [G>A] HbE, and cd 71/72 [+A], each with 2 cases (23.3% prevalence), while mutations -28 [A>G] and Cap+1 [A>C] are only found in this region. In the Northeast region, the most prevalent mutation is cd 17 [A>T] with 5 cases (45.4%) and cd 41/42 [-TTCT] with 2 cases (18.2%). Mutations IVS 1.1 [G>T] and IVS 2.654 [C>T] are exclusively found in this region, each with 1 case.

In the Northwest region, only two types of mutations are found: cd 26 [G>A] HbE with 2 cases (66.7%) and cd 41/42 [-TTCT] with 1 case (33.3%). In the Central Coast region, there are only two mutations: cd 17 [A>T] with 3 cases (75%) and cd 26 [G>A] HbE with 1 case (25%).

Table 3.7. Thalassemia gene carrier rate by ethnicity in pregnant women screened at the NHOG (n=1292)

Nation	Carring α thalassemia genes Quantity, (%)	Carring β thalassemia genes Quantity, (%)	Total
The Kinh	148 (16.28)	18 (1.98)	166 (12.85)
The Muong	32 (26)	1 (0.8)	33 (2.55)
The Tay	20 (21.5)	4 (4.3)	24 (1.86)
The Thai	17 (24.64)	2 (2.9)	19 (1.47)
The Nung	12 (36)	1 (3.33)	13 (0.85)
The Dao	5 (25)	0	5 (0.33)

The San Diu	7 (35)	0	7 (0.46)
Others	6 (26.1)	1 (4.35)	7 (0.46)
Total	247 (90.15)	27 (9.85)	274 (21.2)

Analysis of the carrier rates of α and β thalassemia across ethnic groups reveals that the ethnic groups with the highest carrier rates for α thalassemia are Nung (36%), San Diu (35%), Muong (26%), and Dao (25%). The carrier rates for β thalassemia are high in ethnic groups such as Tay (4.3%), Nung (3.3%), and Thai (2.9%). The Kinh ethnic group, despite having the largest population, has relatively low carrier rates for both α and β thalassemia, accounting for 16.28% and 1.98%, respectively.

Table 3.8. Ethnic distribution of α-thalassemia mutations in pregnant women undergoing screening at NHOG (n=247)

Nation / Mutation type	The Kinh, Quantity (%)	The Muong Quantity (%)	The Tay, Quantity (%)	The Thai, Quantity (%)	The Nung, Quantity (%)	The Dao, Quantity (%)	The San Diu, Quantity (%)	Other nations, Quantity (%
$\alpha^{3.7}$	3 (2.025)	0	0	1 (5.9)	0	1 (20)	0	0
$\alpha^{4.2}$	3 (2.025)	1 (3.1)	0	0	0	0	0	0
--SEA	137 (92.6)	30 (93.8)	19 (95)	15 (88.2)	12 (100)	4 (80)	7 (100)	6 (100)
--THAI	2 (1.34)	0	0	1 (5.9)	0	0	0	0
Anti-$\alpha^{3.7}$	1 (0.67)	1 (3.1)	0	0	0	0	0	0
α2 CD125 [CTG>CCG] (HbQs)	2 (1.34)	0	1 (5)	0	0	0	0	0

Other mutations	0	0	0	0	0	0	0	0
Total (n=247)	148 (100)	32 (100)	20 (100)	17 (100)	12 (100)	5 (100)	7 (100)	6 (100)

Among the various α-thalassemia mutations in pregnant women carrying the gene, the most common is the --SEA mutation, found in all ethnic groups above 80%. The highest prevalence is among the Nung and San Diu ethnic groups (100%), followed by the Thai, Muong, and Kinh ethnic groups, all with a gene carrier rate above 90%. The two mutations with the lowest rates are anti-$\alpha^{3.7}$ and HbQs. The anti-$\alpha^{3.7}$ mutation is present in 01 pregnant woman from the Kinh ethnic group (0.67%) and 01 person from the Muong ethnic group (3.1%). The HbQs mutation is observed in 02 cases, one from the Kinh ethnic group, accounting for 1.34%, and one from the Tay ethnic group (5%). Other mutations have not been detected in any cases. Following the --SEA mutation, the subsequent more prevalent mutations are $\alpha^{3.7}$ (in 5 pregnant women), $\alpha^{4.2}$ (in 4 individuals), --THAI, and HbQs (3 people).

Table 3.9. Ethnic distribution of β-thalassemia mutations in pregnant women coming for screening at the NHOG (n = 27)

Nation	The Kinh, Quantity (%)	The Muong Quantity (%)	The Tay, Quantity (%)	The Thai, Quantity (%)	The Nung, Quantity (%)	The Dao, Quantity (%)	The San Diu, Quantity (%)	Other nations, Quantity (%)
-28[A>G]	1 (5.6)	0	0	0	0	0	0	0
Cap+1 [A>C]	1 (5.6)	0	0	0	0	0	0	0
cd17[A>T]	7 (38.6)	1 (100)	0	1 (50)	1 (100)	0	0	0
Cd26 [G>A]HbE	5 (27.8)	0	0	1 (50)	0	0	0	0
IVS 1.1 [G>T]	0	0	1 (25)	0	0	0	0	0
cd 41/42 [-TTCT]	0	0	3 (75)	0	0	0	0	1 (100)

cd 71/72 [+A]	3 (16.8)	0	0	0	0	0	0	0
IVS 2.654 [C>T]	1 (5.6)	0	0	0	0	0	0	0
Tổng (N=27)	18	1	4	2	1	0	0	1

Among the 1292 pregnant women undergoing thalassemia screening, after conducting diagnostic genetic testing for thalassemia in those with high risk, 8 types of β-thalassemia mutations were identified. Among these, the cd 17 [A>T] mutation is the most common. Ethnic groups with the highest prevalence of this mutation are the Muong and Nung people (100%), followed by the Thai people (50%), and the Kinh people (38.6%). The cd 26 [G>A] HbE mutation has the second-highest prevalence, accounting for 50% in the Thai ethnic group and 27.8% in the Kinh ethnic group. The cd 41/42 [-TTCT] mutation has the third-highest prevalence, observed in the Tay ethnic group (75%) and one case in another ethnic group. The remaining mutations have equal prevalence, with each ethnic group having one individual with those mutations.

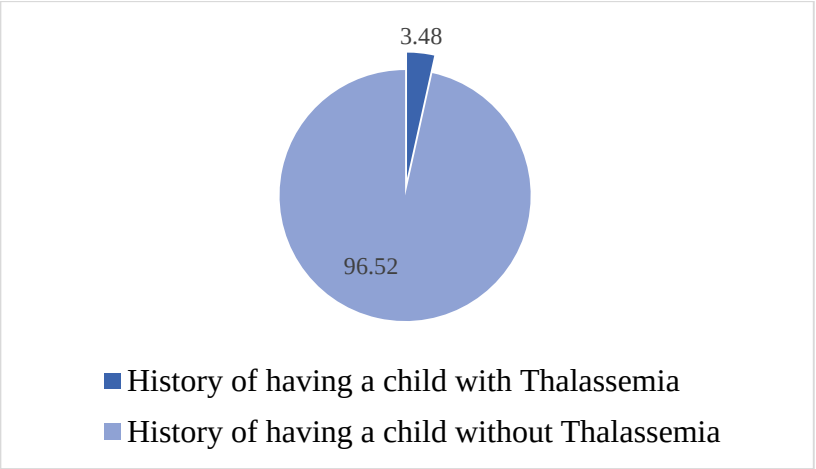

Figure 3.1. History of having a child with the Thalassemia gene of couples participating in screening at the NHOG

Among 1292 couples who came for screening for Thalassemia, the number of couples with a history of having children carrying the Thalassemia gene was 3.48%, the number of couples without a history of having children with the Thalassemia gene was 96.52%.

Table 3.10. Average hematological indices in wifes carrying α-thalassemia genes (n=242)

Blood formula index	Mutation in 1 gene (n_1=7)	Mutations in 2 genes (n_2=230)	Mutations in 3 genes (n_3=7)	α thalassemia /HbE (n_4=8)	p
RBC (T/l)	4.7 ± 0.32	4.94 ± 0.49	4.8 ± 0.81	4.76 ± 0.46	0.33
HGB (g/l)	118 ± 9.55	105.65 ± 8.86	89.56 ± 11.22	97.63 ± 17.27	0.00
HCT (l/l)	0.35 ± 0.02	0.33 ± 0.03	0.31 ± 0.02	0.31 ± 0.05	0.00
MCV (fL)	75.67 ± 4.90	67.65 ± 4.33	66.04 ± 8.45	64.63 ± 6.12	0.00
MCH (pg)	25.18 ± 2.53	21.45 ± 1.22	18.88 ± 2.35	20.38 ± 2.50	0.00
MCHC (g/l)	332.22 ± 15.48	318.19 ± 11.86	287.22 ± 22.04	317.38 ± 21.67	0.00
RDW (%)	13.80 ± 1.93	15.92 ± 3.06	22.59 ± 3.80	15.94 ± 3.97	0.00

Among the 274 pregnant women carrying α-thalassemia genes, there were 7 cases with mutations in 1 gene, 230 cases with mutations in 2 genes, 7 cases with mutations in 3 genes, and 8 cases with α-thalassemia/HbE compound mutations. When comparing the hematological indices among different α-thalassemia genotypes, it was observed that, for the MCV index, the group of pregnant women with mutations in 1 α gene had the highest average value (75.67 ± 4.90), while the group with compound genotypes α-thalassemia/HbE had the lowest (64.63 ± 6.12). The differences were statistically significant with $p < 0.05$.

Regarding the RDW index, the group with mutations in 1 gene had the lowest value (13.80 ± 1.93), followed by the group with mutations in 2 genes (15.92 ± 3.06) and the group with compound genotypes α-thalassemia/HbE (15.94 ± 3.97), while the highest value was observed in the group with mutations in 3 genes (22.59 ± 3.80) with statistically significant differences ($p < 0.05$).

There was no statistically significant difference in the average RBC index among the groups of pregnant women carrying α-thalassemia genes ($p > 0.05$).

Table 3.11. Average hematological indices in husbands carrying α-thalassemia genes (n=238)

Blood formula index	Mutations in 1 gene (n_1=14)	Mutations in 2 genes (n_2=220)	Mutations in 3 genes (n_3=2)	α thalassemia /HbE (n_4=2)	p
RBC (T/l)	5.63 ± 0.39	6.44 ± 0.43	6.55 ± 1.17	6.35 ± 0.41	0.00

HGB (g/l)	143 ± 13.95	137.13 ± 8.18	112.5 ± 10.25	136 ± 5.66	0.00
HCT (l/l)	0.44 ± 0.04	0.44 ± 0.03	0.39 ± 0.01	0.42 ± 0.03	0.01
MCV (fL)	78.59 ± 7.23	67.96 ± 4.63	60.33 ± 9.72	66.3 ± 0.00	0.00
MCH (pg)	25.44 ± 2.64	21.33 ± 1.46	17.4 ± 1.47	21.45 ± 0.49	0.00
MCHC (g/l)	323.53 ± 12.61	314.24 ± 10.88	291.25 ± 23.64	323 ± 7.07	0.00
RDW (%)	13.91 ± 2.77	16 ± 3.81	21.18 ± 2.50	16.4 ± 2.12	0.01

Among the 238 husbands carrying α-thalassemia genes, there were 14 cases with a mutation in 1 gene, 220 cases with a mutation in 2 genes, 2 cases with a mutation in 3 genes, and 2 cases with a mutation in α-thalassemia/HbE. The group of husbands with mutations in 3 genes had the highest average RBC count (6.55 ± 1.17), followed by the group with mutations in 2 genes (6.44 ± 0.43), and the lowest was the group with mutations in 1 gene (5.63 ± 0.39), with statistically significant differences ($p < 0.05$).

Regarding the MCH and MCH indices, the group of husbands with mutations in 1 gene had the highest average values (25.44 ± 2.64 and 323.53 ± 12.61, respectively), followed by the group with a combined α-thalssemia/HbE genotype (21.45 ± 0.49 and 323.00 ± 7.07, respectively) and the group with mutations in 2 genes (21.33 ± 1.46 and 314.24 ± 10.88, respectively). The lowest values were observed in the group with mutations in 3 genes (17.40 ± 1.47 and 291.25 ± 23.64, respectively), with statistically significant differences ($p < 0.05$).

The group of husbands with mutations in 1 gene had the lowest average RDW index (13.91 ± 2.77), followed by the group with mutations in 2 genes (16 ± 3.81), and the group with a combined α-thalssemia/HbE genotype (16.4 ± 2.12). The highest values were observed in the group with mutations in 3 genes (21.18 ± 2.50), with statistically significant differences ($p < 0.05$).

Table 3.12. Average blood formula indexes in pregnant women carrying the β-thalassemia gene

Genotype / Index	$\beta\beta^+$	$\beta\beta^0$	$\beta^0\beta^0$	HbE disease	p
RBC (T/l)	4.78 ± 0.4	4.89 ± 0.64	5.44	4.49 ± 0.39	0.05

HGB (g/l)	108.25 ± 10.5	99.48 ± 10.64	104	112.4 ± 9.84	0.00
HCT (l/l)	0.33 ± 0.03	0.31 ± 0.04	0.32	0.33 ± 0.03	0.11
MCV (fL)	69.60 ± 5.59	63.93 ± 4.58	59.4	73.83 ± 6.61	0.00
MCH (pg)	22.71 ±1.71	20.48 ±1.5	19.1	25.07 ± 2.17	0.00
MCHC (g/l)	326.75 ± 8.96	320.49 ± 13.09	322	340.1 ± 11.05	0.00
RDW (%)	15.07 ± 0.82	16.96 ±3.21	17	14.48 ± 1.5	0.01

The group of pregnant women with β-thalassemia/HbE compound genotype had the highest average HGB and MCHC levels (112.4 ± 9.84, 340.1 ± 11.05, respectively), followed by the $\beta\beta^+$ genotype group (108.25 ± 10.5, 326.75 ± 8.96), lower in the $\beta^0\beta^0$ genotype group (104; 322), and the lowest in the $\beta\beta^0$ genotype group (99.48 ± 10.64, 320.49 ± 13.09) with statistically significant differences ($p < 0.05$).

The average MCV and MCH indices in the groups of pregnant women carrying β-thalassemia genes were the highest in the β-thalassemia/HbE compound genotype group (73.83 ± 6.61, 25.07 ± 2.17), and the lowest in the $\beta^0\beta^0$ genotype group (59.4; 19.1), with statistically significant differences ($p < 0.05$).

In the groups of pregnant women carrying β-thalassemia genes, the $\beta^0\beta^0$ genotype group had the highest average RDW index (17), followed by the $\beta\beta^0$ genotype group (16.96 ± 3.21), lower in the $\beta\beta^+$ genotype group (15.07 ± 0.82), and the lowest in the group with the β-thalassemia/HbE compound genotype (14.48 ± 1.5) with statistically significant differences ($p < 0.05$).

The differences in the average RBC and HCT indices between the groups of pregnant women carrying β-thalassemia genes were not statistically significant ($p < 0.05$).

Table 3.13. Average blood formula indexes in husbands of β-thalassemia types

	$\beta\beta^0$	β-thalassemia/HbE	p
RBC (T/l)	6.31 ± 0.62	6 ± 0.48	0.11
HGB (g/l)	128.43 ± 12.73	142.93 ± 12.11	0.00
HCT (l/l)	0.41 ± 0.04	0.43 ± 0.04	0.10
MCV (fL)	66.03 ± 8.2	71.8 ± 7.74	0.04
MCH (pg)	20.37 ± 1.41	23.98 ± 2.87	0.00

MCHC (g/l)	315.67 ± 16.62	333.93 ± 15.77	0.00
RDW (%)	17.7 ± 5.5	14.69 ± 1.55	0.06

The average peripheral blood cell indices indicate microcytic and hypochromic red blood cells in groups of husbands carrying β-thalassemia genes, with an average MCV < 80 fl and MCH < 28 pg.

The group of husbands with the β-thalassemia/HbE compound genotype has higher average values for HGB, MCV, MCH, and MCHC, which are 142.93 ± 12.11, 71.8 ± 7.74, 23.98 ± 2.87, and 333.93 ± 15.77, respectively, compared to the group with the heterozygous genotype $\beta\beta^0$ (128.43 ± 12.73, 66.03 ± 8.2, 20.37 ± 1.41, 315.67 ± 16.62), with statistically significant differences ($p < 0.05$).

The average RDW index in the group of husbands with the β-thalassemia/HbE compound genotype is 14.69 ± 1.55, lower than the group with the heterozygous genotype $\beta\beta^0$ (17.7 ± 5.5), with no statistically significant difference ($p > 0.05$).

There is no statistically significant difference in the average RBC and HCT indices between the groups of husbands carrying β-thalassemia genes ($p > 0.05$).

Table 3.14. The rate of pregnant women carring disease genes according to screening methods

	Non-genetic carrier	α	β	α /HbE	Hbe	Combined α and β Thalassemia	Total
MCV < 85 or MCH < 28	26 (41.94)	28 (45.16)	6 (9.68)	1 (1.61)	1 (1.61)	0	62 (100)
MCV < 85 or MCH < 28 and iron deficiency	6 (5.66)	86 (81.13)	4 (3.77)	5 (4.72)	2 (1.89)	3 (2.83)	106 (100)
MCV < 85 or MCH < 28 and nomal iron levels	14 (9.27)	115 (76.16)	11 (7.28)	2 (1.32)	3 (1.99)	6 (3.97)	151 (100)
MCV < 85 or MCH < 28 and iron overload	0	1 (100)	0	0	0	0	1 (100)

Among 62 pregnant women with MCV < 85 fL or MCH < 28 pg, 28 pregnant women (45.16%) carry α-thalassemia genes, while 26 individuals (41.94%) do not carry any disease genes. The group with β-thalassemia and HbE genes consists of a total of 7 cases. No cases with coexisting genes were identified. For those with MCV < 85 fL or MCH < 28 pg along with iron deficiency, the majority of cases were associated with α-thalassemia genes, accounting for 81.13%. Pregnant women without thalassemia genes represented 5.66%. In cases with MCV < 85 fL or MCH < 28 pg and normal iron status, α-thalassemia still dominated with 76.17%. Pregnant women without thalassemia genes constituted 9.27%. All cases (100%) with MCV < 85 fL or MCH < 28 pg along with iron excess carried α-thalassemia genes.

Table 3.15. The rate of husband carring disease genes according to screening methods

	Non-genetic carrier	α	β	α /HbE	Hbe	Combined α and β	Total

						Thalassemia	
MCV < 85 or MCH < 28	19 (23.75)	46 (57.50)	8 (10)	1 (1.25)	3 (3.75)	3 (3.75)	80 (100)
MCV < 85 or MCH < 28 and iron deficiency	6 (8.96)	52 (77.61)	2 (2.99)	1 (1.49)	4 (5.97)	2 (2.99)	67 (100)
MCV < 85 or MCH < 28 and nomal iron levels	4 (2.67)	126 (84)	10 (6.67)	0	7 (4.67)	3 (2)	150 (100)
MCV < 85 or MCH < 28 and iron overload	0	3 (100)	0	0	0	0	3 (100)

Among 80 husbands with MCV < 85 fL or MCH < 28 pg, 46 individuals (57.5%) carry Thalassemia genes, while 19 individuals (23.75%) do not carry any disease genes. The group of husbands with β-thalassemia genes accounts for 10%, and the groups with HbE and combined α and β genes both have a rate of 3.75%.

For those with MCV < 85 fL or MCH < 28 pg along with iron deficiency, the majority of cases are associated with α-thalassemia genes, with a high rate of 77.61%. Husbands without Thalassemia genes represent 8.69%. In cases with MCV < 85 fL or MCH < 28 pg and normal iron status, α-thalassemia still dominates with 84%. Husbands without Thalassemia genes constitute 2.67%.

All cases (100%) of husbands with MCV < 85 fL or MCH < 28 pg along with iron excess carry α-thalassemia genes. There are no cases without disease genes.

3.2. Results of Applying Expert System Software and Machine Learning Software in Screening for Congenital Anemia

The research subjects, who underwent genetic testing using the Strip Assay software to identify thalassemia-causing genes, were selected for data analysis using two artificial intelligence software programs: the expert knowledge-based system and the machine learning software. The results are presented as follows:

3.2.1 Results from the expert knowledge-based system software

Table 3.16. Distribution of α and β thalassemia gene carrier rate was screened

Disease phenotype		Quantity (n)	Rate (%)
Non-genetic carrier		43	17.6
Genetic carrier	α -thalassemia	161	66.0
	β -thalassemia	40	16.4
Total		244	100

(RBC, HGB, HCT, MCV, MCH, MCHC, RDW, Fe, Ferritin)

After analyzing the genes of 244 individuals, 201 individuals (82.4%) were found to carry genes, with 161 individuals (66.0%) carrying α-thalassemia genes and 40 individuals (16.4%) carrying β-thalassemia genes. No one carried both α-thalassemia and β-thalassemia genes. The number of individuals not detected with mutations using the current testing method is 43 individuals (17.6%). The accuracy rate of the current method is 50.5%, and the error rate is 49.5%.

Table 3.17. Screening value of HCG software compared to expert diagnosis in the medical record

	Expert Conclusion	HCG software	N = 202		Sensitivity	Specificity
High risk	184	179	TP = 178	FP = 1	96.7%	94.4%
Low risk	18	23	TN = 17	FN = 6		

Among the 244 pregnant women and their husbands assessed for risk using the expert system, 42 cases were undetermined due to not undergoing ferritin testing. There were 8 false-positive cases, all of which had at least one of the three abnormalities in the blood formula, including anemia, small red blood cells, and pale red blood cells, but without iron deficiency. Additionally, two of these cases were concluded to be "At risk of carrying the normal gene of a carrier or thalassemia carrier (α-thalassemia carrier) or β-thalassemia deletion," meaning there is a risk of carrying the gene, but the genetic test results did not detect mutations. These 8 cases could be true false positives, but it could also be due to the limitations of the test not identifying all mutations occurring in these patients. One false-negative case had a completely normal blood formula, with no iron deficiency, but the genetic test results showed that the patient carried the $-\alpha^{3.7}$ mutation of α-thalassemia. After calculating the indices, the expert system software showed a high screening value with a specificity of 94.4% and sensitivity of 96.7% compared to the expert's diagnosis in the medical record.

Table 3.18. Diagnostic performance of the HCG software with genetic test results (The gold standard is the accurate diagnosis of gene carrier/non-carrier)

	Genetic testing	HCG software	N = 202		Sensitivity	Specificity
Genetic carrier	172	179	TP = 171	FP = 8	99.4%	73.3%

Non-genetic carrier	30	23	TN = 22	FN = 1		

With 202 cases included in the trial, when evaluating the additional diagnostic value of the expert system software compared to genetic testing, the software demonstrated a specificity of 73.3% and a sensitivity of 99.4%.

Table 3.19. Evaluation of the results of the expert knowledge system in diagnosing the possibility of carrying the Thalassemia gene

Chỉ số hiệu lực của biện pháp	Tỷ lệ (%)
Sensitivity	99.4
Specificity	73.3
Accuracy	95.5
Positive predictive value	95.5
Negative predictive value	96.6

3.2.2. Results of machine learning software

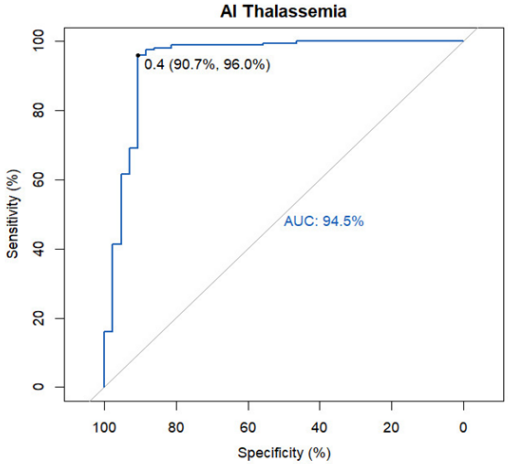

Figure 3.2. ROC curves on the probability prediction of machine learning models

After determining the probability of carrying the gene from machine learning models, the cut-off threshold will be calculated to classify gene carriers and non-carriers in a way that maximizes sensitivity and specificity. The results show an AUC of 0.945, indicating that the model performs well at the cut-off probability level of 0.42. With this cut-off threshold, the assessment of the risk of carrying the gene is presented in the table below.

Table 3.20. Screening value of the machine learning model compared to expert diagnosis in the medical record

Prediction outcome	Expert Conclusion	Machine learning	N = 244		Sensitivity	Specificity
High risk	199	197	TP = 197	FP = 2	95.74 %	100%
Low risk	45	45	TN = 45	FN = 0		

Among a total of 244 cases evaluated for the comprehensive peripheral blood analysis results, the expert concluded 199 high-risk cases and 45 low-risk cases. The machine learning software concluded 197 high-risk cases and 45 low-

risk cases. Thus, there were 2 false negatives. Compared to the doctor's conclusion in the medical records, the AI software demonstrated a sensitivity of 95.74% and specificity of 100%. However, it is necessary to compare the diagnostic results of the gene mutation test to assess the diagnostic value of the AI software in screening individuals carrying Thalassemia genes.

Table 3.21. Diagnostic value of machine learning models compared to genetic testing

Predicion outcome	Genetic testing	Machine learning	N = 244		Sensitivity	Specificity
Genetic carrier	201	197	TP = 193	FP = 4	96.0%	90.7%
Non-genetic carrier	43	45	TN = 39	FN = 8		

With 244 test cases, when assessing the diagnostic value of the machine learning software compared to genetic testing, the software achieved a specificity of 90.7% and sensitivity of 96.0%.

CHAPTER 4: DISCUSSION

4.1. Some clinical epidemiological characteristics, genetic variants of congenital thalassemia in couples attending and undergoing antenatal screening at the National hospital of Obstetrics and Gynecology during the period 2012-2022.

The prenatal Thalassemia screening study conducted at the National hospital of Obstetrics and Gynecology from 2012 to 2022 yielded screening results for 1292 couples. The highest rates of α-thalassemia were 17.8% in pregnant women and 17.65% in husbands, while β-thalassemia rates were 1.63% and 1.55%, respectively. Less common mutations included α and β thalassemia coexistence in pregnant women and husbands at 0.7% and 0.62%, and α/HbE and HbE disease, each with rates below 1%. The gene carrier rates for Thalassemia subtypes in this study were similar to some previous studies on hematological characteristics and Thalassemia gene carrier rates in Vietnam, such as those by Luc Thi Hiep (2015) and Nguyen Thi Van Vy (2020). These studies consistently reported α-thalassemia as the most prevalent (ranging from 16% to 18%), while the β-thalassemia carrier rate in our study was lower than in the other two studies. The rate of carriers for both α and β thalassemia in all three studies was below 1%. Analyzing the overall gene carrier rate for couples revealed that out of 210 couples, both carriers of α-thalassemia accounted for 16.25%, 20 couples both carriers of β-thalassemia

accounted for 1.5%, and 8 couples both carriers of α and β thalassemia accounted for 0.6%.

The most common α-thalassemia gene mutations were ---SEA, with carrier rates of 93.25% in wives and 92.86% in husbands. Following that were mutations like $-\alpha^{3.7}$, $-\alpha^{4.2}$, and ----THAI, with carrier rates ranging from 0.42% to 2.38% in both wives and husbands. Some other mutations with rates below 1% included Anti $-\alpha^{3.7}$, α^2 CD125 [CTG>CCG], and α^2 CD142 [TAA>CAA] (HbCs). The ---SEA mutation is the most prevalent type in Vietnam, as demonstrated in several studies by Nguyen Khac Han Hoan (2013), Vu Thi Bich Huong (2016), Bach Quoc Khanh (2017), and Vu Hai Toan (2018)[10–13]. Common β-thalassemia mutations in this study were CD17 [A>T] and CD26 [G>A] HbE, accounting for carrier rates of 34.48% in wives and 41.67% in husbands. The order of prevalence for β-thalassemia differed from some studies in Vietnam but remained among the most common mutations in the country.

The prevalence of Thalassemia gene carriers varies significantly based on geographical and ethnic factors. Ethnic minorities residing in mountainous regions, such as the Muong, E de, Tay, Thai, Stieng, tend to have higher Thalassemia gene carrier rates compared to the Kinh ethnic group living in the lowland areas. In this study, the Red River Delta and Northeast Vietnam were the regions with the highest rates of α and β Thalassemia gene carriers, ranging from 30.8% to 40.7%. Specifically, the highest α-thalassemia gene carrier rate was in the Northwest region (94.59%) and North Central region (91.83%). The α-thalassemia gene carrier rates were high among various ethnic groups, with Dao and San Diu having the highest rates (100%), followed by Muong (96%), Nung (92.31%), and Kinh (89.16%). The β-thalassemia gene carrier rates among ethnic groups were as follows: Tay (16.67%), Thai (10.53%), Nung (7.69%), Kinh (10.84%), and Muong (4%).

Analyzing the carrier rates of α and β thalassemia genes by ethnicity revealed that the ethnic groups with the highest carrier rates for α-thalassemia were Nung (36%), San Diu (35%), Muong (26%), and Dao (25%). The carrier rates for β-thalassemia were high among Tay (4.3%), Nung (3.3%), and Thai (2.9%). The Kinh ethnic group had carrier rates of 16.28% for α-thalassemia and 1.98% for β-thalassemia. Although the Kinh ethnic group had the highest overall proportion, their carrier rates were lower than those of other ethnic groups. This can be explained by the fact that some ethnic minority groups live in remote areas, and their lifestyle and cultural practices tend to involve consanguineous marriages within the community. Therefore, ethnic groups with higher Thalassemia gene carrier rates are more likely to have couples with both partners carrying the gene marrying each other, leading to an increased risk of having children with Thalassemia.

Through blood tests, biochemical analyses, and electrophoresis of Thalassemia gene carriers among pregnant women and their husbands, it was observed that pregnant women carrying Thalassemia genes exhibited characteristics of small and hypochromic red blood cells, indicated by MCV below 80 fl and MCH below 28 pg. In pregnant women carrying α-thalassemia genes, the average HGB, MCV, MCH, and MCHC indices were higher in the single gene mutation group than in the double gene mutation group, and the triple gene mutation group had the lowest values. The differences were statistically significant ($p = 0.00$). The average RDW index in the single gene mutation group was the lowest, followed by the double and triple gene mutation groups, with statistically significant differences ($p < 0.05$). Similar results were observed in husbands carrying Thalassemia genes.

4.2. Results of Applying Expert System Software and Machine Learning Software in Screening for Congenital Anemia

Our research utilized two widely developed screening software applications in Vietnam, namely, the Expert System Knowledge-Based Software and Machine Learning Software, yielding the following outcomes:

4.2.1. Results of the Expert System Knowledge-Based Software (HCG) in Congenital Thalassemia Screening

The Expert System Knowledge-Based Software is one of the decision support applications for healthcare professionals in prenatal screening. The results of applying this software in Thalassemia screening at the Central Obstetrics and Gynecology Hospital (COGH) revealed that out of 244 individuals analyzed for genetic traits, 201 individuals (82.4%) carried Thalassemia genes, with α-thalassemia and β-thalassemia rates at 66% and 16.4%, respectively. Blood formula analysis of these subjects showed variations, with 189 cases having small, hypochromic red blood cells (with or without anemia), accounting for 77.46% of the cases.

The Expert System Knowledge-Based Software also provided conclusions about the iron status of 192 research subjects, with 3 cases (1.2%) identified as iron-deficient. Additionally, 52 cases could not be evaluated for iron status due to insufficient serum ferritin data. A comparison between expert conclusions and HCG software results revealed that out of 202 cases with assessed risk levels, high-risk cases according to expert conclusions and HCG software were 184 and 179 cases, respectively. There were 8 false-positive cases and 1 false-negative case. After calculating various indices, the HCG software demonstrated high screening value with a sensitivity of 96.7% and specificity of 94.4%. Comparing the HCG software with genetic test results, 179 cases carrying Thalassemia genes were identified compared to the genetic test result of 172 cases.

The Expert System Knowledge-Based Software achieved a diagnostic accuracy of 95.5%, with positive predictive and negative predictive values of 95.5% and 96.6%, respectively. These values surpass the current method employed at the Central Obstetrics and Gynecology Hospital (50.5%).

The Expert System Knowledge-Based Software for prenatal Thalassemia screening requires a well-established and regularly updated knowledge base by genetics, hematology, and obstetrics experts to ensure accuracy and relevance. However, this process demands substantial time and financial resources for data collection, processing, and updates, and may involve disagreements among experts when constructing rule sets. A limitation compared to machine learning methods is that expert system software cannot autonomously learn and enhance prediction results with additional data. Human supervision is necessary for learning adjustments and improving prediction processes.

4.2.2. Results of machine learning software in congenital Thalassemia screening

Out of 201 cases with Thalassemia genes, the Expert System Knowledge-Based Software detected 171 cases, accounting for 85.07%, leaving 30 cases undetected. Meanwhile, the machine learning model identified 193 cases with only 8 cases undetected. The highest sensitivity and specificity, achieved at the cut-off value of 0.42, were 96.0% and 90.7%, respectively.

Compared to the expert system, the machine learning model has the advantage of learning from a large amount of real-world data, thereby providing rules closely aligned with reality and good coverage, though it cannot explain why those rules exist.

Several global studies also use blood formula indices for analysis, with the potential addition of leukocyte and platelet indices depending on the classification purpose, along with the selection of the most crucial indices. However, the differentiator is that our machine learning Thalassemia screening software includes serum iron indices in the model training process to enhance the ability to distinguish between Thalassemia and iron-deficiency anemia.

Throughout the research, we identified some limitations:

- Data was collected only from pregnant women undergoing prenatal screening at the Central Obstetrics and Gynecology Hospital, which may not represent the entire pregnant population in Vietnam.
- The subjects diagnosed prenatal through genetic testing were mostly high-risk cases based on blood formula, and only a few low-risk pregnant women were genetically tested, leading to limitations in situations where the machine learning model could learn.
- The test data set was relatively small, with only 244 cases, and collected solely at the National hospital of Obstetrics and Gynecology. Therefore, the results cannot be universally applied across the country.

- Hence, we propose the following solutions:
- Further research should be conducted on diverse groups of pregnant women to assess result diversity.
- Establish a specific process for periodic data updates to enhance the machine learning software's results when widely applied in healthcare facilities.

CONCLUSION

1. Epidemiological and Clinical Characteristics, Genetic variants of congenital Thalassemia in couples attending prenatal acreening at National hospital of Obstetrics and Gynecology during the period 2012-2022.

Among various types of Thalassemia, α and β thalassemia mutations were the most prevalent, with the rate of carrying α-thalassemia gene in pregnant women being 17.8% and in husbands being 17.65%, while carrying β-thalassemia gene in pregnant women and husbands was 1.62% and 1.55%, respectively. The rate of both couples carrying α-thalassemia gene was 16.25%, β-thalassemia was 1.5%, and carrying both α and β-thalassemia genes was 0.6%. The prevalence of Thalassemia gene carriers among couples is 18.35%

The most common α-thalassemia mutation in couples was the --SEA mutation, accounting for 93.25% in pregnant women and 92.86% in husbands. The most common β-thalassemia mutations were CD17 [A>T], CD41/42 [+A], CD71/72 [+A], with the CD26 [G>A] HbE mutation having a relatively high prevalence in husbands (41.67%).

Among the various hematological parameters in individuals carrying α-thalassemia gene, the MCV and MCH indices in the single-gene mutation group have the highest values, while the lowest values are observed in the compound gene type. The RDW index exhibits a range from low to high in the single-gene mutation groups, two-gene mutation groups, compound α-thalassemia/HbE genotype, and three-gene mutation groups. Regarding the hematological indices in individuals carrying β-thalassemia gene, the highest average MCV and MCH values are found in the compound β-thalassemia/HbE genotype group, and the lowest values are observed in the β0β0 genotype group. The RDW index shows a decreasing trend in the β0β0, ββ0, and ββ+ genotype groups, with the lowest values observed in the compound β-thalassemia/HbE genotype group.

Differences in the rates of iron deficiency, normal iron, and iron excess between the normal group and couples with Thalassemia gene did not show statistical significance.

2. Results of Applying Expert System Software and Machine Learning Software in Screening for Congenital Anemia

Both artificial intelligence software applications used in Thalassemia gene screening exhibited very high sensitivity and specificity. The expert system's

sensitivity and specificity were 96.7% and 94.4%, respectively, while the machine learning software achieved 95.74% sensitivity and 100% specificity.

The machine learning software demonstrated higher utility as it incorporated multiple hematological indices for analysis. However, the expert system had higher validation due to its robust foundation, especially in handling specific cases.

RECOMMENDATION

Expand Thalassemia screening: Continue implementing Thalassemia screening for all pregnant women at all Obstetrics and Gynecology hospitals nationwide, especially at grassroots healthcare facilities and regions with a high population of ethnic minorities.

Data collection and model training: Collect additional data that encompasses diverse scenarios to facilitate the training of machine learning models. This should include cases with both low and high Thalassemia gene risk, along with genetic test results obtained through high-resolution techniques, such as next-generation gene sequencing, to maximize mutation detection. Additionally, the knowledge base should be adjusted and updated with the latest guidelines to ensure the software achieves the highest accuracy.

Testing decision support sortware: Conduct testing of decision support software in Thalassemia prenatal screening on different study groups before deciding on widespread implementation. This will help evaluate the software's performance across various populations and healthcare settings, ensuring its effectiveness and reliability.

LIST OF PUBLISHED RESEARCH WORKS RELATED TO THE DISERTATION

1. **Nguyễn Bá Tùng**, Trần Danh Cường, Nguyễn Thị Trang và cộng sự (2023). Ứng dụng trí tuệ nhân tạo trong tư vấn sàng lọc trước sinh Thalassemia. *Tạp Chí Học Việt Nam*, **526(2)**.
2. **Nguyễn Bá Tùng**, Nguyễn Thị Trang và Nguyễn Tuấn Hưng và cộng sự (2023). Một số đặc điểm dịch tễ, lâm sàng và cận lâm sàng ở thai phụ mang gen bệnh tan máu bẩm sinh đến khám tại Bệnh viện Phụ sản Trung ương, 2012 – 2022. *Tạp Chí Học Việt Nam*, **531**(**1B**).
3. Nguyễn Thị Hồng Hạnh, **Nguyễn Bá Tùng**, Trần Danh Cường và cộng sự (2023). Đánh giá giá trị của kỹ thuật di truyền huyết sắc tố trong sàng lọc người mang gen Thalassemia. *Tạp Chí Dược Học Quân Sự*, **48**(**8**), 26–36.