Research on NIPT Risk Modeling and Optimization Based on Statistical Analysis and Machine Learning

Xiaoxi Li[1,a], Yiyang Zhong[1,b], Zihan Xu[2,c]

[1]School of Computing and Data Science, Xiamen University Malaysia, Sepang, Selangor, Malaysia

[2]School of Economics and Management, Xiamen University Malaysia, Sepang, Selangor, Malaysia

[a]SWE2309521@xmu.edu.my, [b]SWE2309548@xmu.edu.my, [c]FIN2409033@xmu.edu.my

Keywords: NIPT; Spearman Correlation Analysis; K-Means Clustering; Monte Carlo Algorithm; Machine Learning

Abstract: Accurate assessment of fetal chromosome concentrations in non-invasive prenatal testing (NIPT) is crucial for prenatal screening. This paper focuses on the analysis of factors affecting sex chromosome concentration, and establishes a mathematical model based on clinical data to study the important factors affecting the accuracy of detection. The first step was to explore the relationship between fetal Y chromosome concentration and gestational age, BMI and other indicators, and the nonlinear relationship was revealed by Spearman correlation analysis and random forest model, and the influence of various factors was quantified by PDP analysis and ternary polynomial fitting. The second step was to minimize the potential risk of pregnant women, K-means clustering was carried out according to BMI indexes, a risk model was constructed, and the best detection time was determined by quantile regression and comprehensive risk function. In the third step, multifactorial was introduced, Gaussian mixed model clustering and Cox proportional hazard model were used to optimize the risk function in combination with a variety of optimization strategies, and the Monte Carlo simulation gave the optimal screening time of BMI in the four groups (22 weeks for the first two groups, 17 and 18 weeks for the last two groups). This paper constructs a complete NIPT detection data analysis and optimization system, and innovatively combines statistical analysis and machine learning methods to solve the key problems of clinical testing.

1. Introduction

Non-invasive prenatal testing (NIPT) is a non-invasive prenatal screening technique that screens for chromosomal abnormalities by analyzing fetal cell-free DNA in the blood of pregnant women [1]. This technique relies on high-throughput sequencing and bioinformatics analysis to determine whether the fetus has aneuploidy abnormalities by calculating the relative concentration ratio of the target chromosome [2]. NIPT has the advantages of non-invasiveness, safety, and high accuracy, and has become an important means of clinical prenatal screening [3]. However, the accuracy of NIPT is influenced by various factors [4]. Studies have shown that fetal Y chromosome concentration is a key internal control indicator for evaluating the reliability of detection, and it is usually required that the male Y chromosome concentration is not less than 4% [5]. This concentration was significantly correlated with individual characteristics such as gestational age and body mass index (BMI) of pregnant women [6]. In current clinical practice, unified detection time nodes are generally used, which can easily lead to two types of risks: first, the concentration of Y chromosome is not up to standard when detected too early, resulting in distorted results; The second is that the detection is too late, delaying the intervention of abnormal fetuses [7].

This paper studies the key problems in NIPT detection and establishes a mathematical model based on the provided clinical data to solve three core problems. The first step was to explore the correlation between fetal Y chromosome concentration and gestational age, BMI and other indicators, and to construct a mathematical model and test its significance, so as to provide a theoretical basis for subsequent optimization. The second step is to take the BMI of male pregnant

  DOI: 10.25236/iwmecs.2025.008

women as the main influencing factor, and reasonably group them to determine the optimal NIPT detection time for each group to minimize potential risks and analyze the impact of detection errors on the results. The third step is to comprehensively consider multiple factors such as height, weight, age and detection errors, group male pregnant women and determine the best detection time to minimize potential risks, and analyze the impact of errors. Through these three steps, the systematic research aims to provide more personalized and precise solutions for clinical NIPT testing, improving the accuracy and clinical practicability of the test.

2. Model creation, solution and discussion

2.1. Model establishment

2.1.1. Y chromosome concentration regression model

In order to explore the complex relationship between fetal Y chromosome concentration and gestational age, BMI and other indicators, the data were preprocessed, including outlier treatment and standardization. Spearman correlation analysis is used to preliminarily explore the monotonic trend between variables, which does not require variables to obey the normal distribution, and is more robust to outliers:

$$p_{Y,X_j}^{(S)} = \frac{\sum_{i=1}^{n}(R_Y(i)-\overline{R}_Y)(R_{X_j}(i)-\overline{R}_{X_j})}{\sqrt{\sum_{i=1}^{n}(R_Y(i)-\overline{R}_Y)^2\sum_{i=1}^{n}(R_{X_j}(i)-\overline{R}_{X_j})^2}} \tag{1}$$

where $R(i)$ is the rank of the i observation, and $\overline{R}$ is the mean of the corresponding rank. In order to further control the confounding factors, the partial Spearman correlation coefficient is calculated.

On the basis of correlation analysis, machine learning algorithms such as random forest, gradient elevator and neural network are used to construct the prediction model, and the optimal model is determined by comparing MAE, RMSE and R^2 indicators through 5-fold cross-validation. The random forest model makes full use of the sample features by integrating multiple decision trees:

$$\hat{f}_{x_j}(x_j) = \frac{1}{n}\sum_{i=1}^{n} f(x_j, x_c^{(i)}) \tag{2}$$

Among them, f is the trained random forest model, x_j is the feature to be analyzed, and $x_c^{(i)}$ is the true value of other features.

Based on the results of PDP analysis, the nonlinear relationship between each feature and Y chromosome concentration was quantified by fitting the ternary polynomial model by least squares method:

$$\hat{y}_{pdp} = \beta_0 + \beta_1 \cdot x_j + \beta_2 \cdot x_j^2 + \beta_3 \cdot x_j^3 + \varepsilon \tag{3}$$

2.1.2. Risk optimization model based on BMI grouping

With the goal of minimizing the potential risk of pregnant women, the K-means clustering algorithm was used to objectively group pregnant women according to BMI value. The effect of different cluster numbers is evaluated by the contour coefficient, and the optimal group number is determined to be 3.

Low BMI group: [26.62, 31.40], Medium BMI group: [31.41, 34.57], High BMI group: [34.58, 39.35]

A comprehensive risk model was constructed to quantify the detection risk at different gestational age and BMI levels. First, define the success rate function:

$$P_{success}(GA, BMI) = \frac{1}{1+e^{-(GA-\mu_{GA})/\sigma_{GA}}} \cdot \exp\left(-\frac{(BMI-\mu_{BMI})^2}{2\sigma_{BMI}^2}\right) \tag{4}$$

The first logical function describes the effect of gestational age on the success rate, and the second Gaussian function describes the effect of BMI on the success rate.

The risk weight function is defined as:

$$W_{risk}(GA, BMI) = W_0 + \alpha \cdot \exp\left(-\frac{(GA-GA_{\text{opt}})^2}{2\sigma_r^2}\right) \cdot \frac{BMI}{BMI_{\text{ref}}} \tag{5}$$

Among them, W_0 is the basal risk, α is the adjustment coefficient, and GA_{opt} is the ideal gestational age for detection.

The comprehensive risk function consists of several components:

$$group\_risk = w_1 \cdot R_{\text{early}} + w_2 \cdot R_{\text{weighted}} + w_3 \cdot R_{\text{error}} + w_4 \cdot R_{\text{delay}} + w_5 \cdot R_{\text{consistency}} \tag{6}$$

Each component measures early risk, weighted early risk, error risk, delay penalty, and intra-group consistency penalty.

2.1.3. Multi-factor comprehensive risk optimization model

Considering the influence of height, weight, age and other factors, the Gaussian mixed model (GMM) was used for multivariate cluster analysis. GMM assumes that the data is a mixture of multiple Gaussian distributions:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k) \tag{7}$$

The optimal number of groups is determined by AIC and BIC criteria, and the model is highly interpretable and moderately complex when $k = 4$.

The Cox proportional risk model was used to analyze the time to event data of "Y concentration of 4%":

$$h(t \mid X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \tag{8}$$

where $h(t \mid X)$ is the risk function of the individual at gestational age t, $h_0(t)$ is the baseline risk function, and X is the covariate.

Constructing a multi-component risk function to quantify the cost of missed, false, and premature screening:

$$J_g(w) = R_{FN}(w) + R_{FP}(w) + R_{\text{Early}}(w) \tag{9}$$

$$R_{FN}(w) = \alpha \cdot (1 - p_g(w)) \cdot \exp[0.15(w-15)] \tag{10}$$

$$R_{FP}(w) = \beta \cdot p_g(w) \cdot \exp[(-0.1(w-12)] \tag{11}$$

$$R_{\text{Early}}(w) = \lambda \cdot \max(0, 16-w)^2 \tag{12}$$

Three strategies are used to optimize the risk function by using grid search, derivative-free optimization and golden section method to ensure the robustness of the solution.

2.2. Model Solution and Results

2.2.1. Step 1 model solution results

Spearman correlation analysis showed that there was a significant relationship between Y chromosome concentration and gestational age and BMI:

Table 1 Comparison of performance between gestational age and BMI

Correlation	ρ	95%CI permutation	test p-value
Y chromosome concentration vs gestational age	0.069	[0.010,0.128]	0.022
Y chromosome concentration vs BMI	-0.144	[-0.202,-0.085]	0.000

As shown in Table 1, the results of machine learning model comparison show that random forests have the best performance:

Table 2 Machine learning model results show comparison

Model Name	Mean Absolute Error	Root Mean Square Error	Determination coefficient
Random forest	0.0207	0.0262	0.3008
Gradient hoist	0.0217	0.0277	0.2269
Neural Networks	0.0246	0.0300	0.0868

As shown in Table 2, the third-order polynomial fitting results based on the random forest model quantify the influence of various factors:

$$y_{BMI} = -0.6063 + 0.06056x - 0.001746x^2 + 0.0000163x^3 \tag{13}$$

$$y_{GA} = 0.03776 + 0.005283x - 0.0002889x^2 + 0.000006218x^3 \tag{14}$$

Gestational age showed a positive driving effect, BMI showed a negative nonlinear effect, and age had a weak effect.

2.2.2. Step 2 model solution results

K-means clustering obtained three BMI groups and corresponding optimal detection time, and the Monte Carlo simulation results showed that the risk control of each group was good:

Table 3 Recommended gestational age for multivariate clustering

Grouped	Best gestational age	Mean risk	Risk 95% confidence interval	Success rate
G1	14.0	0.38	0.34-0.46	0.34-0.46
G2	12.0	0.41	0.39-0.44	0.39-0.44
G3	10.0	0.46	0.44-0.49	0.44-0.49

As shown in Table 3, all three groups performed well, and the best detection window for different BMI groups was within the range of 10-14 weeks, with good overall risk control and high success rate.

2.2.3. Step 3 model solution results

GMM clustering divided the BMI of male pregnant women into four categories, and the results of the Cox proportional hazards model showed that the latent category LC1 had a significant impact on the risk of events (HR=2.20, 95%CI: 1.11-4.38, p=0.02), and the weight was marginally negatively correlated (HR=0.97, p=0.05).

Table 4 The three optimization methods obtained the optimal screening gestational age for each group

Group	Sample size	Grid search optimal	Continuous optimization optimal	Golden ratio optimal	Average optimal
G1	24	22.0	18.0	18.0	19.33
G2	147	22.1	22.0	22.0	22.03
G3	38	17.0	17.0	17.0	17.00
G4	58	22.1	22.0	22.0	22.03

As shown in Table 4, the risk function decomposition analysis showed that the dominant risk components of different groups were different, with the risk of missed detection in Group 3 (71.4%) and the risk of false detection in Group 2 and Group 4 (about 80-90%). The Monte Carlo test verified that the optimal screening gestational age in each group was highly concentrated in 1000 replicates under the four-category measurement error setting, with an average bias of about +4.0

weeks, but the risk level remained stable, suggesting that the strategy had good robustness.

2.3. Results and discussion

This study systematically addresses key issues in NIPT detection through three steps. In the first step, a quantitative relationship model between Y chromosome concentration and influencing factors was established, and it was found that gestational age had a significant positive effect on Y chromosome concentration, while BMI showed a negative effect, which was consistent with clinical observation and provided an important basis for the optimization of detection time. The stochastic forest model shows optimal performance in predicting Y chromosome concentration, and its nonlinear modeling ability better captures the complex characteristics of biomedical data.

The second step provides personalized detection time recommendations based on BMI grouping, and the robustness of the protocol is verified by risk modeling and Monte Carlo simulation. The low, medium, and high BMI groups are recommended to be tested at 14, 12, and 10 weeks, respectively, a result that challenges the traditional concept of uniform testing time and reflects the concept of personalized medicine. The construction of the risk function comprehensively considers various factors in clinical practice, so that the model output has clear clinical guidance significance.

In the third step, the best screening time for pregnant women with different characteristics was determined by GMM clustering and Cox risk model. The best screening gestational age for the first two groups was 22 weeks, and the last two groups were 17 and 18 weeks, respectively, reflecting the testing time requirements of different population characteristics. Multi-strategy validation of the optimization algorithm ensures the reliability of the understanding, and the Monte Carlo simulation confirms the robustness of the scheme against measurement errors.

Overall, the NIPT detection data analysis and optimization system constructed in this study innovatively combines statistical analysis with machine learning methods, from single-factor to multi-factor, from linear to nonlinear, and gradually solves the key problems in clinical testing. The model fully considers various risk factors in practical applications, and provides a scientific basis and actionable personalized plan for clinical practice.

3. Conclusion

This paper focuses on the key issues in non-invasive prenatal testing (NIPT) and constructs a complete risk modeling and optimization system through three systematic research steps. The first step is to explore the relationship between fetal Y chromosome concentration and gestational age, BMI and other indicators, and the complex nonlinear relationship between various factors and Y chromosome concentration is revealed by Spearman correlation analysis and random forest regression model. The study found that gestational age had a significant positive driving effect on Y chromosome concentration, while BMI showed a significant negative effect, and the effect of pregnant women's age was relatively weak. This step quantifies the influence of each factor through machine learning algorithm comparison and polynomial fitting, which lays a solid theoretical foundation for subsequent detection time optimization.

In the second step, with the goal of minimizing the potential risk of pregnant women, the K-means clustering group was carried out based on BMI index, and a comprehensive risk function model was established to determine the optimal detection time for each group. The study divided pregnant women into three BMI groups: low, medium and high, and recommended testing at 14, 12 and 10 weeks respectively. Through quantile regression and Monte Carlo simulation, it is verified that the detection error has little impact on each group, and the detection results are stable and reliable. The innovation of this step lies in quantifying clinical risk into an optimizeable objective function, providing a scientific basis for personalized testing time recommendations.

In the third step, the Gaussian mixed model is used for multivariate cluster analysis, the Cox proportional risk model is used to construct the risk function, and the optimal detection time point is solved by multiple optimization algorithms. The study determined the optimal screening time for pregnant women with different BMI characteristics in four groups, with the best screening gestational week for the first two groups being 22 weeks, and the last two groups being 17 weeks

and 18 weeks, respectively. The Monte Carlo simulation shows that the random error has little effect on the detection results, and the systematic error has a significant influence, but the overall results remain stable. This step represents an advancement from univariate to multivariate analysis, enhancing the clinical applicability and robustness of the model.

The main contribution of this study is to construct a complete NIPT detection data analysis and optimization system, which innovatively combines statistical analysis and machine learning methods to solve key problems in clinical testing. The proposed risk assessment model based on multi-index fusion not only provides a new method for the assessment of Y chromosome concentration, but also provides a new idea for the formulation of personalized plans for prenatal screening. The modeling framework and method system of this study have good versatility, and can be generalized to the optimization and analysis of other prenatal screening indicators, which is of great value for improving the level of birth defect prevention and control. Future research can further expand the sample size, optimize model parameters, and apply this framework to a wider range of prenatal screening scenarios.

Acknowledgements

Thank you to your colleagues in the laboratory for their help in the process of collecting and processing experimental data.

References

[1] Hartwig, T. S., Ambye, L., Sorensen, S., & Jørgensen, F. S. (2017). Discordant non-invasive prenatal testing (NIPT)-a systematic review. Prenatal diagnosis, 37(6), 527-539.

[2] Mackie, F. L., Hemming, K., Allen, S., Morris, R. K., & Kilby, M. D. (2017). The accuracy of cell-free fetal DNA-based non-invasive prenatal testing in singleton pregnancies: a systematic review and bivariate meta-analysis. BJOG: an international journal of obstetrics and gynaecology, 124(1), 32–46.

[3] Zheng, Y., Wan, S., Dang, Y., Song, T., Chen, B., & Zhang, J. (2020). Clinical experience regarding the accuracy of NIPT in the detection of sex chromosome abnormality. The journal of gene medicine, 22(8), e3199.

[4] Bianchi, D. W., Parker, R. L., Wentworth, J., Madankumar, R., Saffer, C., Das, A. F., ... & CARE Study Group. (2014). DNA sequencing versus standard prenatal aneuploidy screening. New England Journal of Medicine, 370(9), 799-808.

[5] Norwitz, E. R., & Levy, B. (2013). Noninvasive prenatal testing: the future is now. Reviews in Obstetrics and Gynecology, 6(2), 48-62.

[6] Gil, M. M., Accurti, V., Santacruz, B., Plana, M. N., & Nicolaides, K. H. (2017). Analysis of cell-free DNA in maternal blood in screening for aneuploidies: updated meta-analysis. Ultrasound in Obstetrics & Gynecology, 50(3), 302-314.

[7] Gregg, A. R., Skotko, B. G., Benkendorf, J. L., Monaghan, K. G., Bajaj, K., Best, R. G., ... & Watson, M. S. (2016). Noninvasive prenatal screening for fetal aneuploidy, 2016 update: a position statement of the American College of Medical Genetics and Genomics. Genetics in Medicine, 18(10), 1056-1065.