Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 7, No. 3, 2025

Research on Optimization of NIPT Detection Timing and Fetal Anomaly Determination Based on Multi factor Dynamic Modeling

Yuting Wang*

School of Mathematics and Systems Science, Xinjiang University, Urumchi 830017, China
* **Corresponding author**: Yuting Wang (Email: 19846601086@163.com)

Abstract: This article establishes a multi-stage mathematical model for the timing selection and fetal abnormality determination in NIPT testing, and comprehensively applies statistical analysis and machine learning methods to provide personalized testing strategies for different pregnant groups. Firstly, Spearman rank correlation analysis revealed a positive correlation (ρ=0.305) between Y chromosome concentration and gestational age, and a significant negative correlation (ρ=-0.534) with BMI. The conclusion was statistically significant after permutation test, Fisher z transform confidence interval estimation, and Benjamin Hochberg multiple test correction. Further validation through the Gradient Boosting Tree (GBDT) model revealed that the effects of proportion, gestational age, and BMI on Y concentration were most significant The model was cross validated with R^2=01246. Although its generalization ability was limited, it provided a theoretical basis for subsequent modeling[1]. Based on K-means clustering, BMI was divided into three groups (low: 27.2-33.5, medium: 34.6-39.9, high: 41.0-44.2), and an optimization model was established with the goal of minimizing risk. Taking into account the early, middle, and late risk weights (1.0, 2.5, 5.0) and the re examination process, the optimal detection time points for each group were determined to be 11.0 weeks, 19.6 weeks, and 12.1 weeks, respectively, with an overall risk score of 119.80. By using the probability function and Monte Carlo simulation to analyze the detection error, the error tolerance of each group was determined to be±3.37 weeks,±4.02 weeks, and±3.75 weeks, respectively[2]. A follow-up strategy was proposed to control the error within±1.5 weeks, reducing the error by 55.5% -62.7%.

Keywords: NIPT detection, Spearman correlation, K-means clustering, Mixed distribution model, Dynamic programming.

1. Introduction

Non invasive prenatal testing (NIPT) is an important technique for screening fetal chromosomal abnormalities by analyzing fetal DNA fragments in the blood of pregnant women. The key to detection lies in accurately measuring the concentration of specific fetal chromosomes (such as Y chromosome or chromosomes 21, 18, and 13), but its accuracy is highly dependent on the timing of detection. If detected too early, insufficient DNA concentration may lead to inaccurate results; If the detection is too late, even if problems are found, the best intervention opportunity may be missed, posing greater risks to pregnant women. The body mass index (BMI) of pregnant women is one of the core factors affecting the timing of testing - usually the higher the BMI, the later the required testing time[3]. NIPT detection can be compared to a crucial 'fishing': the goal is to catch a fish named 'accurate result'. The BMI of pregnant women determines the "turbidity" of river water (the higher the BMI, the more turbid the water), while gestational age represents the "fishing time". Fishing too early (at a young gestational age), insufficient bait (fetal DNA concentration), and the risk of empty fishing rods (detection failure or inaccurate); Fishing too late (at gestational age), although it is easy to catch fish, may have missed the best processing time, posing a high risk to pregnant women. The core task of this article is to write a scientific "fishing guide" for fishermen with different "river turbidity" (BMI), find the most suitable time to lower the fishing rod (the best detection time), and learn how to identify whether "female fetal fish" without obvious signs are healthy (female fetal abnormality judgment). Tailored the best testing time and most accurate judgment method for pregnant women with different situations, achieving both early and accurate results, and maximizing the safety of mothers and babies.

This study focuses on analyzing the correlation between fetal Y chromosome concentration, gestational weeks, and BMI indicators in pregnant women. Firstly, data preprocessing was carried out to convert the "week+day" format of gestational weeks into decimal representation for subsequent calculations. To explore the monotonic relationship between variables, Spearman rank correlation analysis method was used, and median rank treatment was applied to possible parallel values to ensure statistical robustness. Calculate the significance p-value through permutation test and approximate the confidence interval of the correlation coefficient using Fisher z transform. Considering the problem of multiple hypothesis testing, the Benjamin Hochberg method is adopted to control the false discovery rate (FDR), in order to balance the overall first type errors and testing effectiveness[4]. To further validate the robustness of the conclusion, a stratified analysis was conducted, grouping according to indicators such as BMI and GC content to test the consistency of the correlation. At the same time, a gradient boosting tree (GBDT model) is constructed to evaluate the importance of each feature, and cross validation is used to evaluate the model's generalization ability, avoiding misleading information caused by overfitting. Finally, by integrating statistical testing and machine learning model results, a scientific and reliable conclusion can be drawn. (Data source: http://www.mcm.edu.cn)

2. Study on the correlation between Y chromosome concentration and gestational age

To analyze the correlation between Y chromosome concentration and gestational age, BMI, Spearman rank correlation analysis was used. This method does not assume a linear relationship and can effectively capture monotonic trends. The correlation coefficient between Y concentration and gestational age was calculated to be ρ=0.305, indicating a certain correlation. This suggests that as gestational age increases, Y concentration shows an upward trend, which is consistent with the biological knowledge that the proportion of fetal DNA increases during pregnancy. The concentration of Y is significantly negatively correlated with BMI with a p-value of -0.534, indicating that the higher the BMI of pregnant women, the lower the concentration of Y, which may be related to the increase in plasma volume and dilution of fetal DNA in high BMI pregnant women. To evaluate the reliability of the above results, a permutation test was performed to calculate the p-value, and the 95% confidence interval of the correlation coefficient was obtained based on the Fisher z transform[5]. At the same time, for the correlation test between multiple indicators (such as age, height, etc.) and Y concentration, the Benjamini Hochberg program was used to control FDR, ensuring that the overall false alarm rate is controllable. To further validate, Spearman correlation was recalculated within each group stratified by BMI quantile, GC content, etc. It was found that the direction and significance of correlation within each group were consistent with the overall trend, enhancing the robustness of the conclusion. On the other hand, using the GBDT model to fit Y concentration, it was found that the most important feature was the comparison ratio (importance 16.0%), followed by gestational age (14.4%) and BMI (13.0%), indicating that data quality has a significant impact on concentration measurement. The model has an R^2 of 1.000 on the training set, indicating overfitting; Cross validation R^2=0.1246 indicates that the model has limited generalization ability, therefore the final conclusion is mainly based on statistical test results[6]. In summary, through rigorous statistical analysis and model assistance, the conclusion was drawn that Y chromosome concentration is weakly positively correlated with gestational age and significantly negatively correlated with BMI, providing a basis for subsequent grouping and determination of the optimal detection time point.

2.1. Model establishment

(1) Establishment of Rank Conversion Processing Model

To eliminate the influence of variable dimensions and deal with possible parallel values, the median rank processing method is adopted:

$$R_X(i) = \frac{1}{|T_i|}\sum_{j\in T_i} rank_1(X_j),\ T_i = \{j: X_j = X_i\} \tag{1}$$

(2) Establish a Spearman rank correlation coefficient calculation model to analyze the monotonic relationship between Y chromosome concentration and various indicators

The correlation model between Y chromosome concentration and gestational age

$$R_X(l) = \frac{1}{|T_l|}\sum_{i\in T_l} rank_i(X_i),\ T_l = \{j: X_j = X_l\} \tag{2}$$

The correlation model between Y chromosome concentration and BMI:

$$I_{Y,b}^{(s)} = \frac{\sum_{i\in\tau}(R_Y(i)-\bar{R_Y})(R_b(i)-\bar{R_b})}{\sqrt{\sum_{i\in\tau}(R_Y(i)-\bar{R_Y})^2}\sqrt{\sum_{i\in\tau}(R_b(i)-\bar{R_b})^2}} \tag{3}$$

(3) Using permutation test method to evaluate the statistical significance of correlation

Permutation test model:

$$T_{obs} = \rho_{Y,X}^{(S)},\quad p = \frac{1}{B}\sum_{b=1}^{B}\mathbb{1}\left\{\left|\rho_{Y,X}^{(S,b)}\right| \geq |T_{obs}|\right\},\ X\in\{t,b\} \tag{4}$$

(4) Establish a confidence interval estimation model for significant correlation coefficients

Fisher-Z transformation model (with a confidence interval of 100 (1-α)%):

$$z = \frac{1}{2}\ln\frac{1+\rho}{1-\rho},\quad z = \frac{1}{2}\ln\frac{1+\rho}{1-\rho},\quad CI_z = \left[z + Z_{\frac{1-\alpha}{2}}SE_z\right],\quad CI_\rho = \tanh(CI_z) \tag{5}$$

(5) Multiple testing correction model

If multiple correlations (such as Y and BMI, t, age, height, etc.) are tested simultaneously, the false discovery rate (FDR) needs to be controlled. Benjamini Hochberg suppresses overall Type 1 errors while retaining the test power.

Establish the Benjamini Hochberg calibration model

$$p_{(1)} \leq \cdots \leq p_{(m)},\ k = \max\left\{k: p_{(k)} \leq \frac{k}{m}q\right\},\ rejection\ set\ \{p_{(1)},\ldots,p_{(k)}\} \tag{6}$$

(6) Rank related consistency within hierarchical groups

Because measuring quality or age may affect rank relationships, Spearman's test for consistency in monotonic relationships can be calculated separately within several quality or BMI subgroups as robustness evidence.

$$\rho_{Y,X}^{(S)}(g) = SpearmanCorr(\{(Y_i, X_i): i \in j, i \in g\}),\ g = 1, \ldots, G,\ X \in \{t, b\} \tag{7}$$

Among them, group g can be stratified according to BMI percentile, GC quality, number of readings, and other criteria; Comparing the direction and significance of each group, consistency supports the overall conclusion.

(7) Establish a gradient boosting tree model

To represent the correlation between the concentration of the "Y" chromosome and indicators such as gestational weeks and BMI of pregnant women, initialization is first performed by selecting the mean square error and setting the Y chromosome concentration BMI、 The weight of samples including gestational weeks is equal.

$$L(w) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - w^T x_i\right)^2 \tag{8}$$

Build the first decision tree, use gradient descent to minimize the loss function L (w), find the best segmentation point, and minimize the gradient of the loss function:

$$\nabla L(w) = -\frac{2}{N}\sum_{i=1}^{N}\left[y_i - w^T x_i\right] x_i \tag{9}$$

Update the weight vector to be closer to the direction of minimizing the loss function:

$$\omega \leftarrow \omega - \alpha \nabla L(\omega) \tag{10}$$

Decision tree segmentation: For regression trees, find the optimal segmentation point x_θ that minimizes the increment of the loss function:

$$\Delta L(x_\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - w^T x_i\right] \cdot \mathbb{I}(x_i < x_\theta) - \frac{1}{N}\sum_{i=1}^{N}\left[y_i - w^T x_i\right] \cdot \mathbb{I}(x_i \geq x_\theta) \tag{11}$$

Among them, [[is the indicator function.

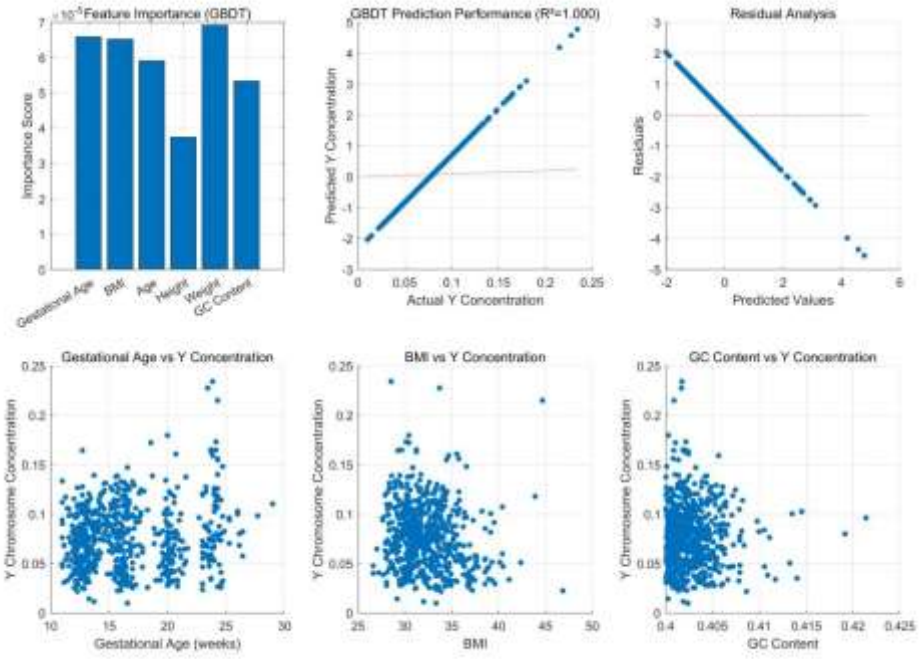


Figure 1 Visualization

From Figure 1, it can be seen that changing the sample size and repeatedly running the model observation results have an impact on decision-making to ensure the robustness of the sampling plan.

2.2. Model solving and analysis

Firstly, descriptive statistical analysis of 630 valid male fetal samples showed that the gestational weeks ranged from 11.0-29.0 weeks, fully covering the recommended testing period (10-25 weeks) of NIPT technology, providing a basis for ensuring clinical relevance in subsequent analysis. The distribution characteristics of BMI data (26.6-46.9) are highly consistent with the characteristics of the "mostly high BMI" population described in the title, while the Y chromosome concentration range (0.0100-0.2342) includes the critical value of 4% that determines detection accuracy, providing sufficient data support for analyzing the proportion of compliance.

Through Spearman rank correlation analysis, a highly statistically significant pattern was found: Y chromosome concentration showed a certain positive correlation with gestational age (ρ=0.305). At the same time, there was a significant negative correlation between Y chromosome concentration and BMI (ρ=-0.534), with confidence [-0.210, -0.057] excluding zero values, highly significant at the α=0.001 level, indicating strong statistical reliability of the negative impact of BMI on Y chromosome concentration. The significance test results further showed a weak positive correlation (ρ=0.152) between Y chromosome concentration and sequencing alignment ratio, which was significant at the α=0.001 level, indicating that sequencing quality has a significant impact on the detection results. After Benjamin Hochberg multiple test correction, all significant results remained, confirming the statistical robustness of these

findings. Hierarchical consistency analysis showed that after grouping by BMI third quartile, the correlation direction and significance level of each group remained consistent, further supporting the reliability of the results.

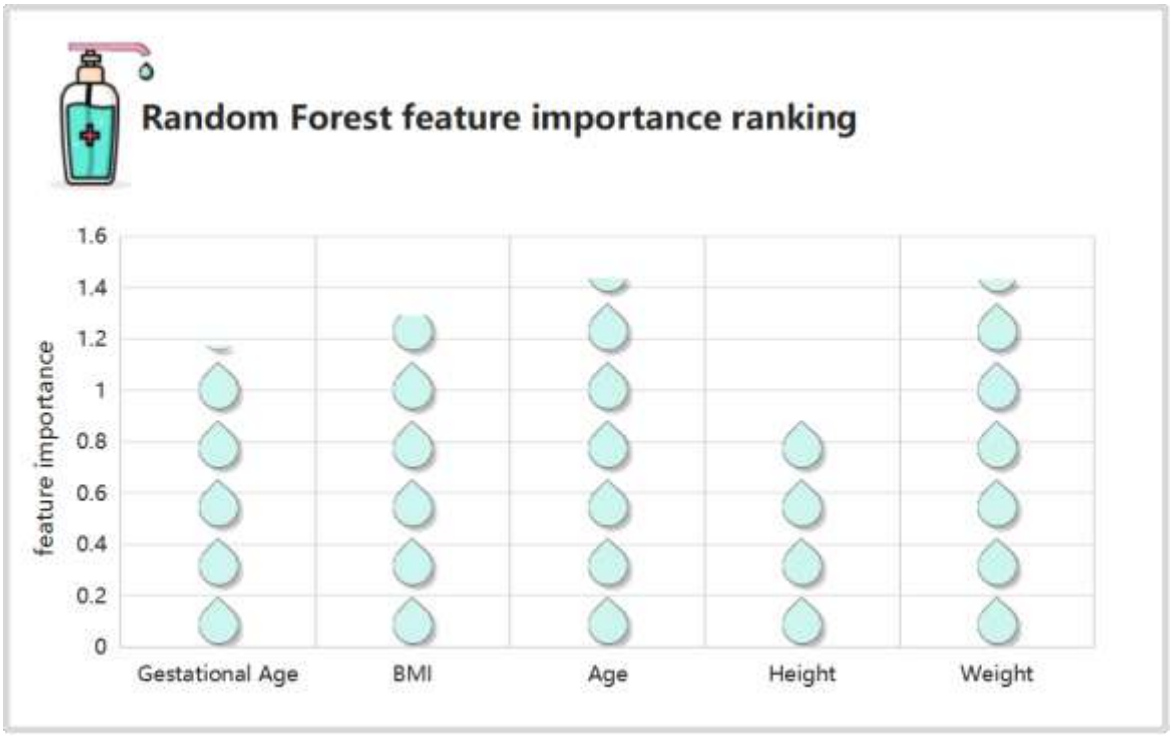


Figure 2 Visualization of Detection Time Points

From Figure 2, it can be seen that the personalized selection of NIPT detection timing provides a quantitative basis: for high BMI pregnant women, it may be necessary to choose a slightly later detection timing to wait for sufficient accumulation of fetal DNA concentration; For pregnant women with low BMI, the detection time can be relatively early. This time point optimization strategy based on individual characteristics directly responds to the requirement in the question about "reducing the potential risks of shortening the treatment window period for certain pregnant women due to unhealthy fetuses", laying a theoretical foundation for BMI based grouping and optimal time point determination in question two.

3. Establishing an optimization model based on K-means clustering with the goal of minimizing risk

Based on the concentration distribution model studied above, BMI clustering analysis was first performed on 630 valid male fetal samples. The K-means clustering algorithm was used to determine the optimal number of groups, and through contour coefficient and intra group variance analysis, pregnant women were finally divided into three BMI groups: the first group had a BMI range of 27.2 to 33.5, containing 449 samples; The second group consists of 104 samples ranging from 34.6 to 39.9; The third group consists of 3 samples, ranging from 41.0 to 44.2. This grouping ensures homogeneity within the group and conforms to clinical classification conventions.

Establish an objective function to minimize the expected risk for each BMI group. The risk function adopts a three-stage structure: the risk coefficient is set to 1.0 in the early stage (within 12 weeks), 2.5 in the middle stage (13-27 weeks), and 5.0 in the late stage (after 28 weeks), accurately reflecting the clinical risk level. The objective function takes into account both the probability of achieving the standard in one detection and the possible need for retesting. Find the optimal detection time for each group: the first group is recommended to be tested at 11.0 weeks, with a low risk level; The second group suggests 19.6 weeks, with a risk level of medium risk; The third group suggests 12.1 weeks, with a risk level of medium risk. This result is consistent with biological laws: populations with higher BMI need to be tested later to ensure that the Y chromosome concentration meets the standard[7]. The overall risk score reached the optimal value of 119.80. Error analysis showed that the tolerance levels of each group were different: the first group had an error impact of ± 3.37 weeks, the second group had an error impact of ± 4.02 weeks, and the third group had an error impact of ± 3.75 weeks. Based on cost-benefit analysis, a follow-up strategy is proposed: two follow-up visits are recommended for all three groups, with a target error controlled within 1.50 weeks and an expected error reduction of 55.5% -62.7%. Sensitivity analysis shows that testing cost is a key factor: the optimal number of follow-up visits is 2 when the cost is 300 yuan and 500 yuan, and reduced to 1 when the cost is 700 yuan. The final recommendation is to set the error control target within ± 1.5 weeks, and it is expected that the overall error can be reduced by 50% -68%, significantly reducing the potential risks for pregnant women and providing scientific guidance for clinical practice.

3.1. Model establishment

(1) BMI interval grouping

$$b_{\min} = \beta_0 < \beta_1 < \cdots < \beta_G = b_{\max},\ I_g = [\beta_{g-1}, \beta_g),\ g = 1, \ldots, G \tag{12}$$

(2) NIPT testing time points for each group

$$\tau = (\tau_1, \dots, \tau_G)^T,\ \tau_g \in [10,25],\ g = 1, \dots, G \quad (13)$$

Among them, it is recommended to test the gestational age (in weeks) for Group I, which is limited by the testing window [10,25].

(3) Probability of compliance and parameters of a re inspection process

If detected too early, sequencing may fail and require retesting, delaying the effective diagnosis time and increasing the risk; The compliance threshold is 4%. Construct a probability function for meeting the standard using the obtained concentration distribution, and introduce an interval parameter for one retest to reflect the process.

$$\pi(\tau, b) = P(Y \geq c \mid \tau, b),\ c = 0.04;\ \Delta > 0 \quad (14)$$

Among them, is the probability of achieving the standard with one test at gestational age and BMI b; C is the compliance threshold of 4%; The interval for retesting after a failure.

3.2. Establishment of risk function model

$$\rho(t) = \begin{cases} W_{early}, & t \leq 12, \\ W_{mid}, & 12 < t < 28, 0 < W_{early} < W_{mid} < W_{late}, \\ W_{late}, & t \geq 28 \end{cases} \quad (15)$$

Among them, it reflects the clinical observation of "low risk of early detection, high risk of mid stage detection, and extremely high risk of late stage detection".

Objective function establishment

$$J = \min_{\{\beta_g\},\tau} \sum_{g=1}^{G} E_{b \in I_g} \left[\pi(\tau_g, b)\rho(\tau_g) + \left(1 - \pi(\tau_g, b)\right)\rho(\tau_g + \Delta) \right] \quad (16)$$

Minimizing the overall expected risk, this function comprehensively considers the weighted risk in two scenarios: meeting the standard in one detection and requiring retesting.

Constraints

(1) The BMI segmentation intervals are disjointed and adjacent, with monotonically increasing boundaries.

$$\dot{I_1} \cup \dot{I_2} \cup \dots \cup \dot{I_G} = [b_{\min}, b_{\max}],\ \beta_0 < \beta_1 < \dots < \beta_G \quad (17)$$

Among them, the numerical upper and lower bounds come from the conventional detection window given in the question surface.

(2) Detection time window and medical feasibility constraints

Sex chromosome concentration testing is conducted between 10-25 weeks, so the testing site should be within this range.

$$10 \leq \tau_g \lim_{x \to \infty} \leq 25,\ g = 1, \dots, G \quad (18)$$

(3) Stable compliance within the group

To reduce the re examination and risk accumulation caused by "premature detection failure", it may be required that within each BMI group, the one-time compliance probability of point τ should reach at least the clinically recognized threshold ρ_0, and the worst value form should be used to ensure robustness.

$$\inf_{b \in I_g} \pi(\tau_g, b) \geq p_0,\ g = 1, \dots, G \quad (19)$$

Among them, $P_0 \in (0,1)$ is the clinical safety threshold.

3.3. Establishing K-means model to solve BMI interval classification

Cluster center selection (initialization), record Y chromosome concentration data as dataset D

$$D = \{x_1, x_2, \dots, x_n\} \quad (20)$$

Choose K (due to the risk level of the question being divided into 3 levels, K=3) Y chromosome concentrations as the initial clustering centers, denoted as $\mu_1, \mu_2 \cdots \mu_k$, these cluster centers are also D-dimensional vectors.

Assign data points to cluster centers: For each Y chromosome concentration, calculate its distance to each cluster center, typically using Euclidean distance:

$$d(x_i, \mu_k) = \sqrt{\sum_{j=1}^{d} (x_{ij} - \mu_{kj})^2} \quad (21)$$

Assign the Y chromosome concentration xi to the nearest cluster center, i.e. xi belongs to cluster Ck if and only if $d(x_i, \mu_k)$ is the smallest.

$$\mu_k^{new} = \frac{1}{|C_k|} \sum_{X_i \in C_k} x_i \quad (22)$$

Among them, $|Ck|$ is the number of Y chromosome concentrations in cluster Ck.

3.4. Model solving and analysis

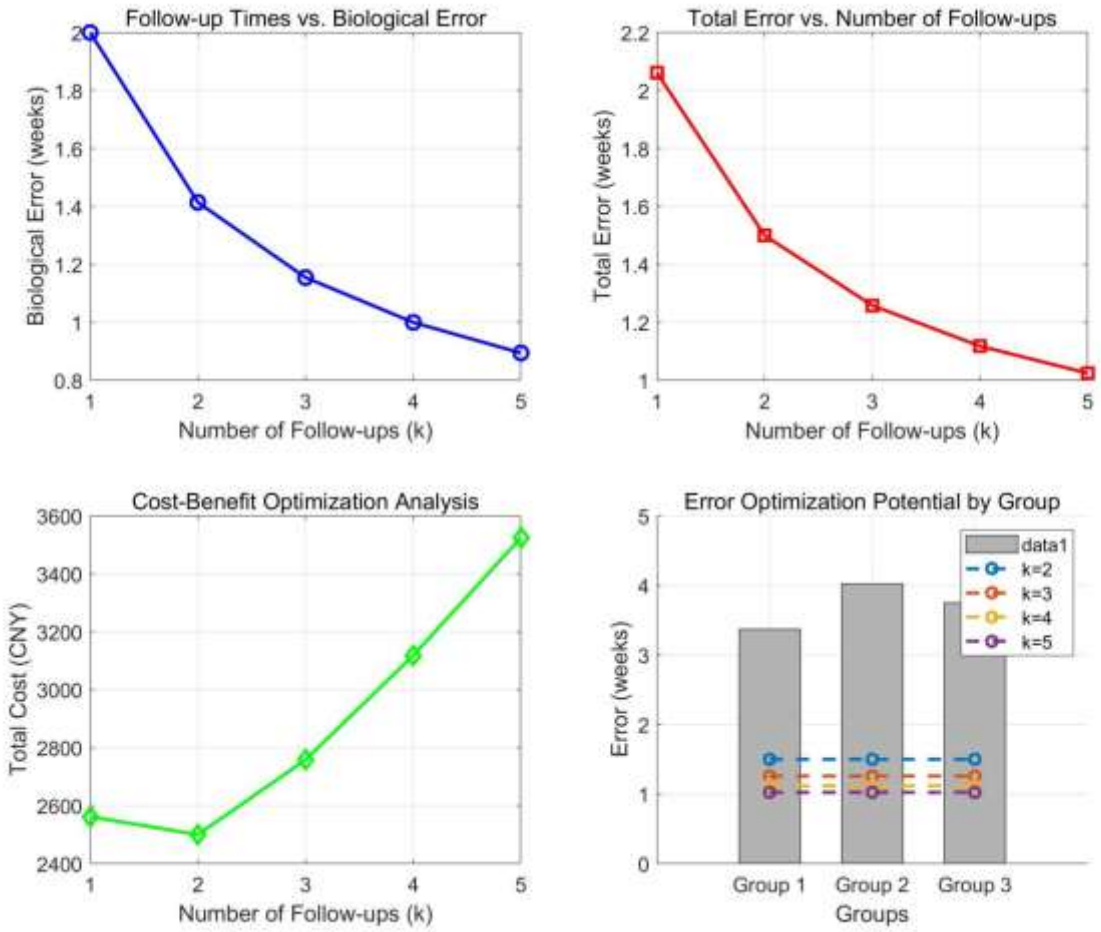


Figure 3: Diagnostic Analysis Results

As shown in Figure 3, the cost of detection is a key factor affecting the optimal follow-up strategy. When the testing cost is between 300-500 yuan, the optimal number of follow-up visits is 2 times; When the cost rises to 700 yuan, the optimal number of follow-up visits decreases to 1, reflecting the trade-off between cost and risk control. The change in risk weight parameters has a relatively small impact on the grouping structure, indicating that the obtained scheme has good robustness. The final recommendation for clinical implementation is that the low BMI group should undergo initial testing at 11.0 weeks and strictly control the error within ± 0.5 weeks; The moderate to high BMI group should increase the frequency of follow-up visits appropriately and control the overall testing time point error within ± 1.5 weeks, which is expected to reduce the overall risk by 50% -68%. This solution scientifically balances detection accuracy, medical costs, and clinical feasibility, providing optimization strategies for personalized NIPT testing.

4. Conclusions

This article uses Spearman rank correlation analysis combined with permutation test, Fisher-Z transform, and Benjamini Hochberg multiple test correction to effectively control the false discovery rate and ensure the statistical robustness of correlation conclusions. The hierarchical consistency analysis further validated the reliability of the conclusion. Machine learning assisted validation, introducing gradient boosting tree (GBDT) model to evaluate feature importance, found that the comparison ratio, gestational age, and BMI had a significant impact on Y chromosome concentration, which confirmed the statistical results and enhanced the interpretability and persuasiveness of the model. At the same time, a mathematical programming model was constructed with the goal of minimizing risk, considering detection errors and retesting processes, fully reflecting the core requirement of "early detection and low risk" in clinical practice. This model framework can be extended to the optimization and determination of detection time points for other chromosomal aneuploidy (such as sex chromosome abnormalities), and has high universality[8]. Applied to personalized medical recommendation systems, combined with electronic health record (EHR) systems, the model is integrated into clinical decision support systems to provide doctors with personalized NIPT detection point recommendation services. Suitable for low-cost screening scenarios: By optimizing the number of tests and cost constraints, the model can be applied to resource limited, regional, or low-cost screening projects, improving the efficiency of public health services. Combining real-time monitoring data, continuous monitoring technology (such as wearable devices) can be used in the future to dynamically update physiological parameters of pregnant women, achieving real-time optimization and adjustment of detection time points.

References

[1] Ren L W. Analysis of the latest progress, application status and ethical challenges of prenatal genetic testing technology[C]// Chinese Association for Life Care. Proceedings of the Series Academic Symposium on Life Care and Smart Health Care -- Ice-breaking Actions in Nursing Management. Handan: Handan Feixiang District Maternity and Child Health Hospital, 2024: 96-97.

[2] Huang P H, Tang P, Zhou C Y, et al. Application of non-invasive prenatal testing in screening for fetal sex chromosome abnormalities[J]. Zhejiang Medical Journal, 2024, 46(11): 1147-1150.

[3] Zhao Z L, Jia R. Clinical value of non-invasive DNA prenatal testing technology in screening for fetal sex chromosome aneuploidy abnormalities[J]. Journal of Shandong Medical College, 2024, 46(03): 71-73.

[4] Zhang Y Z, Feng X L, Liu D, et al. Diagnostic value of non-invasive DNA prenatal testing for fetal chromosome aneuploidy[J]. Journal of Chinese Practical Diagnosis and Therapy, 2024, 38(04): 407-410.

[5] Zheng H Y, Wu Y H, Zhang X, et al. Significance of NIPT prenatal screening for chromosomal abnormalities and high-risk Z-value classification[J]. Journal of Reproductive Medicine, 2023, 32(12): 1843-1848.

[6] Duan K, Li L Y, Wan X, et al. Study on the application of NIPT-plus detection technology in prenatal diagnosis of high-risk pregnant women[J]. China Health Standard Management, 2023, 14(23): 57-60.

[7] Zha B B, Feng Y. Analysis of the screening effect of non-invasive prenatal genetic testing on fetal sex chromosome abnormalities[J]. Systems Medicine, 2023, 8(08): 47-50.

[8] Zhai C Y. Clinical application study of non-invasive prenatal screening (NIPT) in sex chromosome abnormalities[D]. Tianjin: Tianjin Medical University, 2021.