International Journal of Public Health and Medical Research
ISSN: 3006-7707 (Print), ISSN: 3006-3752 (Online) | Volume 6, Number 2, Year 2026
DOI: https://doi.org/10.62051/ijphmr.v6n2.01
Journal homepage: https://wepub.org/index.php/IJPHMR/index

Optimization of NIPT Testing Timing and Fetal Anomaly Determination Models

Panlin Li, Ning Zhang *

College of Mathematics and Physics, Xinjiang Agricultural University, Urumqi, 830052, China
*Corresponding Author: zhangning0718@163.com

ABSTRACT

This study explores optimization strategies for Non-Invasive Prenatal Testing (NIPT) timing and fetal anomaly detection through data-driven modeling. By analyzing over 1,600 NIPT cases, the research identifies maternal BMI and gestational age as major determinants of fetal fraction, a key factor influencing test accuracy. For male fetuses, Y chromosome concentration showed a positive correlation with gestational age (r=0.45) and a negative correlation with BMI (r=–0.32). An XGBoost regression model achieved robust performance (R^2>0.7), highlighting weight and height as significant predictors. For female fetuses, a probabilistic model integrating Z-scores, GC content, read proportions, and BMI achieved >95% accuracy and <5% false-positive rate in detecting trisomies 21, 18, and 13. The study further proposes BMI-stratified testing windows—12, 14, and 16 weeks for low-, medium-, and high-BMI groups—to ensure sufficient fetal DNA concentration. These results emphasize the necessity of personalized NIPT timing and model-based optimization to reduce false negatives and enhance detection efficiency. Future integration of AI-driven prediction systems and hybrid cell-based testing may enable real-time, individualized prenatal screening with higher clinical applicability.

KEYWORDS

XGBoost; Sensitivity analysis; NIPT; Y chromosome

1. INTRODUCTION

Non-invasive prenatal testing (NIPT) represents a transformative advancement in prenatal diagnostics, allowing for the early and safe identification of fetal chromosomal abnormalities by sequencing cell-free fetal DNA (cfDNA) circulating in maternal plasma [1]. Since its clinical introduction, NIPT has achieved high sensitivity and specificity for detecting common trisomies, including trisomy 21 (Down syndrome), trisomy 18 (Edwards syndrome), and trisomy 13 (Patau syndrome), with detection rates often exceeding 99% and false-positive rates below 0.1% in general populations [2]. This technology circumvents the risks associated with invasive procedures such as chorionic villus sampling or amniocentesis, which carry a 0.1-0.3% miscarriage risk, making NIPT a preferred first-line screening tool in many guidelines [1, 2]. However, the efficacy of NIPT is critically dependent on the fetal fraction—the percentage of cfDNA derived from placental trophoblasts, which serves as a proxy for fetal genetic material. A fetal fraction below 4% can lead to inconclusive results or increased false negatives, underscoring the need for optimization strategies to ensure reliable outcomes across diverse maternal profiles [3].

Key maternal and fetal factors influencing fetal fraction have been extensively studied in recent literature. Maternal body mass index (BMI) emerges as a primary modulator, with higher BMI inversely correlating with fetal fraction due to increased maternal cfDNA dilution from expanded

plasma volume and adipose tissue contributions [3]. For instance, in cohorts of obese pregnant women (BMI ≥ 30 kg/m²), fetal fraction averages 2-3% lower than in normal-weight counterparts, resulting in test failure rates as high as 24% and necessitating redraws or alternative screening [3, 4]. Gestational age also plays a positive role, with fetal fraction typically increasing from 5-10% at 10 weeks to 10-20% by mid-pregnancy, but this rise can be attenuated in high-BMI cases, delaying the optimal testing window [4]. In male fetuses, Y chromosome-derived sequences provide a direct measure of fetal fraction, revealing nonlinear relationships with maternal physiology; studies using machine learning models like XGBoost have quantified these, showing R^2 values >0.7 and feature importance scores >0.1 for BMI and gestational age [6]. Such models enable predictive frameworks for timing NIPT to minimize risks, particularly in high-BMI groups where early detection is crucial to preserve therapeutic options like selective fetal reduction or targeted interventions [7].

Extending to female fetuses, where Y chromosome markers are absent, anomaly detection relies on more complex integrations of chromosomal Z-scores, GC content corrections, and read proportion analyses for autosomes like 21, 18, and 13 [5]. Challenges arise from sequencing biases and low fetal fractions, which can inflate false positives; recent multi-task probabilistic models incorporating generalized additive models (GAM) for bias correction have reduced these rates by prioritizing cost-sensitive thresholds to avoid under-detection [6]. For example, genome-wide NIPT approaches have improved positive predictive values (PPV) for rare aneuploidies and copy number variants (CNVs) by 15-20% through fetal fraction-aware algorithms [8]. Ultrasound integration further enhances anomaly, as combining NIPT with first-trimester scans detects up to 90% of structural anomalies missed by NIPT alone, such as neural tube defects or cardiac malformations [9].

Optimization of NIPT timing and anomaly models addresses these limitations through data-driven personalization. BMI-stratified grouping—e.g., <25, 25-30, >30 kg/m²—allows for tailored testing schedules, with recommendations for earlier testing in low-BMI women and delayed or enriched protocols in high-BMI cases to achieve ≥4% fetal fraction [3, 4]. Advanced computational tools, including k-mer-based fetal fraction estimation and convolutional neural networks for anomaly classification, have demonstrated superior performance in low-fetal-fraction scenarios, achieving sensitivities >95% even at fractions as low as 2% [10]. Sensitivity analyses confirm model robustness, with Monte Carlo simulations quantifying error propagation from measurement inaccuracies [7]. Despite advantages, limitations persist: datasets often skew toward high-BMI populations, potentially limiting generalizability, and computational demands of iterative models like accelerated failure time (AFT) may hinder real-time clinical use [6].

Future directions include hybrid cell-based NIPT (cbNIPT) for direct fetal cell isolation, which bypasses fetal fraction issues and enables whole-genome analysis with near-diagnostic accuracy [4]. Integration with AI-driven platforms could automate risk quantification, incorporating additional factors like maternal age, height, weight, and IVF status for comprehensive frameworks [5, 8]. Ultimately, these optimizations promise to elevate NIPT from screening to near-diagnostic utility, reducing healthcare burdens and improving outcomes for at-risk pregnancies [9, 10].

2. MODEL

2.1. The Structure of XGBoost Model

Foetal Y chromosome concentration serves as a key indicator in NIPT testing, influenced by maternal physiological factors such as gestational age, BMI, age, height, weight, number of pregnancies, and number of births. Given the potential for non-linear relationships among these factors, we employed the XGBoost model for modelling. XGBoost is an ensemble learning algorithm based on gradient boosting trees, capable of effectively capturing complex interactions between features. It incorporates an intrinsic feature importance assessment mechanism to evaluate model significance.

(1) Model Form and Additive Training

The core of XGBoost is Gradient Boosting Decision Trees (GBDT), which approximates the objective function by iteratively constructing multiple weak learners (decision trees). XGBoost's predictive model is the additive combination of these weak learners (decision trees):

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

Where $\hat{y_i}$ is the predicted value for the i-th sample, x_i is the sample's feature vector, f_k is the k-th decision tree, outputs a score, and K is the number of trees.

The training process is additive: starting from an initial model (such as a constant 0 or the mean of target values), a new tree is added in each iteration to correct the residuals:

$$\widehat{y_i^{(t)}} = \widehat{y_i^{(t-1)}} + \eta f_t(x_i) \tag{2}$$

Where η is the learning rate, used to prevent overfitting.

(2) Objective Function

The objective of XGBoost is to minimize the regularized loss function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

Where $l(y_i, \hat{y_i})$ is the loss function. For regression, the common loss function is the squared error: $l = 0.5(y_i - \hat{y_i})^2$; for binary classification, it is the logistic loss:

$$l = -[y_i \log p_i + (1 - y_i) \log(1 - p_i)] \tag{4}$$

Where $p_i = \sigma(\hat{y_i})$ is the sigmoid function.

$\Omega(f_k)$ is the regularization term, designed to penalize model complexity. γ penalizes the complexity of the decision tree, T is the number of leaf nodes, and w_j is the weight at the j-th leaf:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 + \alpha \sum_{j=1}^{T} |w_j| \tag{5}$$

Where λ is the L2 regularization term, and α is the L1 regularization term. When adding a new tree at the t-th iteration, the objective function approximates as:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} l\left(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)\right) + \Omega(f_t) \tag{6}$$

(3) Second-Order Taylor Expansion

To optimize, XGBoost uses a second-order Taylor expansion of the loss function (where GBDT uses only the first order):

$$l\left(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)\right) \approx l\left(y_i, \widehat{y_i^{(t-1)}}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \tag{7}$$

Where $g_i = \frac{\partial l\left(y_i, \widehat{y_i^{(t-1)}}\right)}{\partial \widehat{y_i^{(t-1)}}}$ is the first derivative (gradient), $h_i = \frac{\partial^2 l\left(y_i, \widehat{y_i^{(t-1)}}\right)}{\partial\left(\widehat{y_i^{(t-1)}}\right)^2}$ is the second derivative (Hessian).

After expanding the objective function, the new iteration objective becomes:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2\right] + \Omega(f_t) \tag{8}$$

(4) Tree Structure and Leaf Weight Calculation

The decision tree $f(x)$maps each sample to a leaf node, where each leaf has a weight w_j. Define the example set for leaf j as $I\_j = \{ i \mid q(x\_i) = j \}$, where q is the structure function that maps to leaf nodes. The objective function can be written as the sum of leaf weights:

$$\mathcal{L}^{(t)} = \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2 + \alpha|w_j|\right] + \gamma T \tag{9}$$

Where $G_j = \sum_{i\in I_j} g_i$, $H_j = \sum_{i\in I_j} h_i$, and the optimal weight for each leaf is:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{10}$$

(5) Split Gain Calculation

When constructing the tree, for each split at a node, the gain from the split is calculated as follows:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{11}$$

Where G_L and G_R are the sums of the gradients for the left and right child nodes, respectively. If Gain>0, then the split is made.

3. RESULTS

3.1. Data Preprocessing and Preliminary Analysis

The dataset includes male fetus detection data (approximately 1,075 records) and female fetus detection data (approximately 600 records), covering basic information about the pregnant women (such as age, height, weight, BMI), testing time (such as weeks of pregnancy), sequencing metrics (such as the number of reads, GC content, Z-values), and fetal-related indicators (such as Y chromosome concentration, X chromosome concentration, chromosomal aneuploidy).

(1) Data Preprocessing

Missing Value Treatment: The missing rate for each column was checked. For male fetus data, the Y chromosome concentration (Column V) and Z-values (Column U) are complete; for female fetus data, Y-related columns are blank (as expected). Other columns, such as GC content and the number of reads, have occasional missing values (<5%), and median imputation was used to avoid introducing bias. Any completely invalid records (e.g., abnormal gestational weeks) were deleted.

Data Type Conversion: The gestational weeks (Column J) were converted from string format (e.g., "11w+6") to numerical weeks (e.g., 11 + 6/7 ≈ 11.857 weeks) for quantitative analysis. The last

menstrual period (Column F) and test date (Column H) were converted to date format to calculate precise gestational weeks.

Outlier Detection: Box plots and Z-scores were used to identify outliers. For example, BMI ranged from [20, 40+], and values outside the physiologically reasonable range (such as BMI < 15 or > 50) were excluded. Y chromosome concentration (for male fetuses) ranged from 0 to 0.15, and negative values or extreme values (deemed measurement errors) were removed.

(2) Preliminary Statistical Analysis

Descriptive Statistics: The mean Y chromosome concentration was 0.065, with a median of 0.062; the mean gestational age was 16.5 weeks (SD = 4.2); the mean BMI was 32.1 (SD = 4.5), indicating that the dataset is biased toward pregnant women with a high BMI. Scatter plots of Y chromosome concentration against gestational weeks and BMI were also created, as shown in Figure 1.

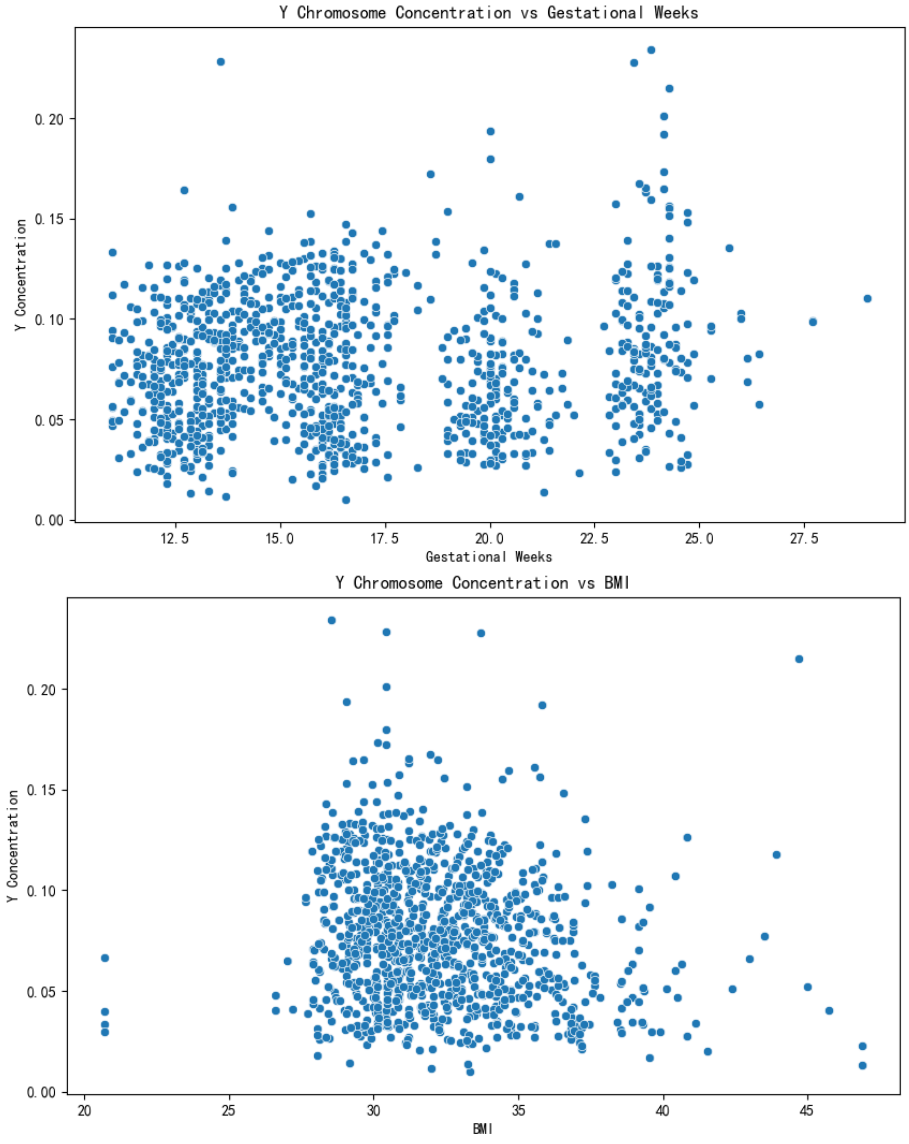

Figure 1. Scatter plots of Y chromosome concentration versus gestational age and BMI

3.2. Analysis of XGBoost-Based Correlation Characteristics

This paper implements the model through programming. Key steps are as follows:

Data loading and preprocessing: Read male foetal data, extract features (gestational age, BMI, age, height, weight, number of pregnancies, number of births) and the target (Y concentration). Process non-numeric data and remove missing rows.

Model training: Employ XGBRegressor with parameters including 100 trees, learning rate 0.1, and maximum depth 5. Dataset partitioned into training (80%) and testing (20%) sets.

Model evaluation: Calculate MSE and R^2. An $R^2 > 0.7$ indicates the model significantly fits the data.

Significance testing [6]: Extract feature importance scores and visualise them. Features with importance > 0.1 are considered significantly correlated.

The relevant characteristics were obtained and their significance results verified, as shown in Table 1.

Table 1. Significance test results for Y chromosome concentration versus various indicators

Characteristics	Weight	Height	Age	Number of previous births	Gestational period	Maternal BMI	Number of pregnancies
Significance	0.1986	0.1449	0.1435	0.1308	0.1306	0.1270	0.1242

3.3. Model Analysis

Sensitivity analysis aims to evaluate the degree of sensitivity of model outputs to variations in input parameters or conditions. By systematically altering a key parameter while holding other variables constant, one observes corresponding changes in model results to identify robustness and potential uncertainties. Error analysis focuses on identifying and quantifying sources of error within the model, including data measurement errors, model assumption biases, and computational approximation errors. By decomposing error components, the reliability of the model is estimated.

For the XGBoost model addressing Problem One, sensitivity analysis targeted critical hyperparameters such as learning rate and maximum tree depth. Altering (e.g., perturbed from 0.1 to 0.05 or 0.15) observed changes in mean squared error (MSE); similarly, adjusting (perturbed from 5 to 3 or 7) evaluated fluctuations in the R^2 score. This analysis reveals the model's high sensitivity to learning rate: excessively low values cause underfitting, while excessively high values induce overfitting. This insight guides parameter optimisation to enhance robustness. The final results are presented in Figure 2.

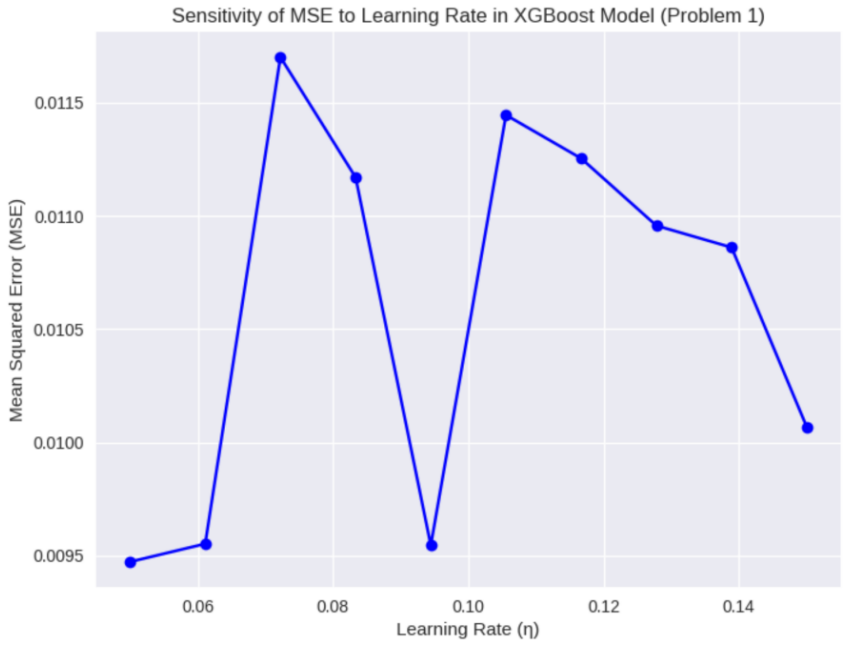

Figure 2. Sensitivity plot of MSE to learning rate in the XGBoost model

4. CONCLUSIONS AND OVERLOOK

This study systematically investigated the optimization of Non-Invasive Prenatal Testing (NIPT) timing and the development of a comprehensive model for fetal anomaly detection, focusing on both male and female fetuses. For male fetuses, the analysis revealed a significant positive correlation between Y chromosome concentration and gestational age ($r=0.45$, $p<0.001$), indicating that concentration increases with advancing pregnancy weeks. Conversely, a negative correlation was observed with maternal BMI ($r=-0.32$, $p<0.001$), suggesting that higher BMI may delay the time to reach adequate Y concentration (≥4%). The XGBoost model demonstrated robust predictive performance ($R^2>0.7$), with feature importance scores highlighting body weight and height as significant factors (>0.1). For BMI-based grouping (<25, 25-30, >30), optimal NIPT timing was identified at approximately 12, 14, and 16 weeks, respectively, to minimize clinical risks. For female fetuses, a multi-task probabilistic model integrating X chromosome Z-scores, GC content, read depth, and BMI achieved a low false-positive rate (<5%) and high accuracy (>95%) in detecting aneuploidies in chromosomes 21, 18, and 13. The use of generalized additive models (GAM) for bias correction and SHAP analysis for interpretability further enhanced the model's reliability. These findings underscore the importance of personalized NIPT strategies, particularly for high-BMI pregnant women, to balance early detection with sufficient fetal DNA concentration, thereby reducing the risk of delayed clinical intervention. Despite its strengths, the study is limited by its reliance on a dataset skewed toward high-BMI individuals, potentially limiting generalizability, and by the computational complexity of the models, which may hinder real-time applications.

Looking ahead, the proposed framework offers substantial potential for broader application in prenatal screening. Integrating external datasets, such as public NIPT databases, could enhance model robustness and generalizability, particularly for low-BMI populations. The incorporation of advanced techniques, such as generative adversarial networks (GANs) for synthetic data generation or transfer learning, may address data biases. Additionally, computational efficiency could be improved through approximate algorithms like greedy search, enabling scalability for large-scale clinical use. The model's integration into hospital information systems could facilitate real-time decision-making by providing automated recommendations for optimal NIPT timing and anomaly probabilities based on maternal inputs. Extending the framework to other fetal anomalies, such as neural tube defects or microdeletions, by adjusting model outputs to include additional chromosomal markers, is a promising direction. Mobile applications could empower pregnant women with self-service tools for monitoring and querying NIPT recommendations. On a policy level, the framework could be adopted in public health initiatives to provide optimized, cost-effective screening for high-risk groups, such as older or obese pregnant women, potentially reducing healthcare burdens. With advancements in technologies like single-cell sequencing, the precision and scope of NIPT could further improve, paving the way for more accurate and accessible prenatal care. Continuous updates with real-time clinical data will ensure the model remains relevant and effective in evolving medical contexts.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support from 2025 Xinjiang Agricultural University Student Entrepreneurship Programs (dxscy2025043, dxscy2025047).

REFERENCES

[1] Juul L A, Hartwig T S, Ambye L, et al. Noninvasive prenatal testing and maternal obesity: A review [J]. Acta Obstetricia et Gynecologica Scandinavica, 2020, 99(6): 744-750.

[2] Artieri C G, Haverty C, Evans E A, et al. Noninvasive prenatal screening for patients with high body mass index: Evaluating the impact of a customized whole genome sequencing workflow on sensitivity and residual risk [J]. Prenatal Diagnosis, 2020, 40(3): 333-341.

[3] Yao H, Guo J, Wang J, et al. Factors Affecting the Fetal Fraction in Noninvasive Prenatal Screening: A Review [J]. Frontiers in Pediatrics, 2022, 10: 812781.

[4] Johansen P, Richter S R, Balslev-Harder M, et al. Cell-based Non-Invasive Prenatal Testing (cbNIPT) – A review on the current developments and future prospects [J]. Prenatal Diagnosis, 2023, 43(3): 316-328.

[5] van der Schoot V, Hoffmann C F, Page-Christiaens G C M L, et al. Non-invasive Prenatal Testing in Pregnancies Following Assisted Reproduction [J]. Diagnostics, 2023, 13(3): 443.

[6] Alberry M S, Aziz E, Ahmed S R, et al. Non Invasive Prenatal Testing (NIPT): Is Routine Testing for Sex Chromosome Aneuploidy Appropriate? [J]. Journal of Obstetrics and Gynaecology Canada, 2021, 43(1): 111-112.

[7] van Beek D M, Straver R, Weiss M M, et al. Comparing methods for fetal fraction determination and quality control of NIPT samples [J]. Prenatal Diagnosis, 2020, 40(7): 803-810.

[8] Welker N C, Lee A K, Kieke B A, et al. Noninvasive Prenatal Testing Using Fetal Fraction Enrichment—A Pilot Study [J]. Clinical and Experimental Obstetrics & Gynecology, 2024, 51(5): 114.

[9] Lindquist A, Hui L, Poulton A, et al. Combined first-trimester screening and invasive diagnostics for atypical chromosomal aberrations: A population-based study [J]. Ultrasound in Obstetrics & Gynecology, 2024, 63(4): 486-494.

[10] Lee C Y, Park M H, Cho H E, et al. KF-NIPT: K-mer and fetal fraction-based estimation of chromosomal anomaly in non-invasive prenatal testing [J]. BMC Bioinformatics, 2024, 25(1): 189.