Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 12
ESAC 2026

Analysis of Factors Influencing Fetal Y Chromosome Concentration Based on a Linear Mixed-Effects Model

Peng Yang *, Dujing Luo, Zihong Zhang

Business School, Chengdu University of Technology, Chengdu, China, 610059

* Corresponding Author Email: 202308020223@stu.cdut.edu.cn

Abstract. The accuracy of Non-Invasive Prenatal Testing (NIPT) relies heavily on sufficient fetal Y chromosome concentration in maternal blood. Identifying key influencing factors is critical for optimizing testing timing and improving diagnostic reliability. This study employed multiple linear and polynomial regression models to explore relationships between Y concentration, gestational age, and maternal BMI. To address repeated measurements and significant individual variations, a linear mixed-effects model was introduced, incorporating fixed effects for population-level trends and random effects for individual-specific variations. Gestational age demonstrated a significant positive effect on Y chromosome concentration ($p < 0.001$), while BMI showed a weak negative correlation ($p < 0.05$). However, all models exhibited low goodness-of-fit ($R^2 \approx 0.04–0.07$), indicating substantial unexplained individual variability. Discussion and Innovation: The study highlights the necessity of personalized approaches in NIPT timing, as population-level factors alone are insufficient for accurate prediction. The application of a mixed-effects model provides a robust framework for handling hierarchical data and isolating individual differences, offering a methodological foundation for future personalized prenatal testing strategies.

Keywords: Linear Mixed-Effects Model; Y Chromosome Concentration; Gestational Age.

1. Introduction

Non-invasive prenatal testing (NIPT) has revolutionized the screening for fetal chromosomal abnormalities by analyzing cell-free fetal DNA in maternal blood [1-2]. The clinical accuracy of this technique is critically dependent on achieving a sufficient fetal DNA concentration, with a key threshold being a fetal Y chromosome fraction of ≥4% for male pregnancies [3]. However, significant inter-individual variation exists among pregnant women in factors such as gestational age, body mass index (BMI), and maternal age, which directly influence this concentration. A uniform testing protocol applied to this heterogeneous population can lead to inaccurate results or delayed diagnosis, potentially missing critical therapeutic windows and posing substantial health risks. Consequently, there is an urgent clinical need to move beyond a one-size-fits-all approach and develop personalized NIPT strategies that determine the optimal testing timepoint based on individual maternal characteristics.

The primary objective of this study is to quantitatively analyze the correlation between fetal Y chromosome concentration and two pivotal maternal factors: gestational age and BMI. We aim to construct and validate a robust mathematical model that accurately describes these relationships, thereby providing a foundational analytical framework for tailoring NIPT protocols to individual differences [4].

To achieve this, our research first employed traditional multiple linear and polynomial regression models to explore potential linear and non-linear associations. Recognizing a major limitation of these approaches—their inability to account for the repeated measurements from the same individual and the substantial inherent inter-individual variability that violates the assumption of independence—this study introduces a key methodological innovation. We developed a linear mixed-effects model to specifically handle this hierarchical data structure. This model incorporates fixed effects to estimate the population-average trends of gestational age and BMI on Y concentration, while simultaneously integrating random effects to capture subject-specific variability in baseline concentration and rate of

change [5-6]. This approach ensures more robust parameter estimation and reliable statistical inference, providing deeper insights into both population trends and individual differences, which is a crucial step towards truly personalized prenatal care.

2. Establishment and Solution of the Y-Chromosome Concentration Relationship Model

2.1. Exploratory Analysis of Linear and Nonlinear Relationships

Based on exploratory data analysis, this study first assumed a potential relationship between fetal Y-chromosome concentration, gestational age [7], and maternal BMI, and initially constructed a multiple linear regression model to quantify their associations:

$$Y = \beta_0 + \beta_1 G + \beta_2 B + \epsilon \tag{1}$$

Where Y represents the fetal Y-chromosome concentration (unit: %),G denotes the gestational age (unit: weeks), B stands for maternal Body Mass Index (BMI, unit: kg/m²), $\beta_0, \beta_1, \beta_2$ are regression coefficients to be estimated, ϵ is the random error term, assumed to follow a normal distribution $\epsilon \sim N(0, \sigma^2)$ (where σ^2 is the error variance).

To capture potential nonlinear associations (e.g., quadratic trends of gestational age or BMI on Y-chromosome concentration, or interactive effects between variables), a polynomial regression model with higher-order terms and interaction terms was further established [8-10]:

$$Y = \beta_0 + \beta_1 G + \beta_2 G^2 + \beta_3 B + \beta_4 B^2 + \beta_5 G \times B + \epsilon \tag{2}$$

The ordinary least squares (OLS) method was adopted to estimate the regression coefficients. The core principle is to minimize the sum of squared residuals between the observed values and predicted values of Y-chromosome concentration:

$$\min \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \tag{3}$$

$\widehat{Y}_i = \beta_0 + \beta_1 G_i + \beta_2 B_i$ (for the linear model) or $\widehat{Y}_i = \beta_0 + \beta_1 G_i + \beta_2 G_i^2 + \beta_3 B_i + \beta_4 B_i^2 + \beta_5 G_i B_i$ (for the polynomial model) represents the predicted Y-chromosome concentration of the i-th sample, and n is the total number of samples.

2.2. Model Testing and Significance Analysis

(1) Overall Model Significance Test (F-test)

The F-test was used to verify whether the established regression model has overall explanatory power for Y-chromosome concentration, i.e., whether at least one of the independent variables (gestational age, BMI) has a significant linear relationship with the dependent variable.

Null Hypothesis (H_0): $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (all regression coefficients of independent variables are zero, meaning the model is invalid),

Alternative Hypothesis (H_1): At least one $\beta_j \neq 0$ (the model has overall explanatory power).

The F-statistic was calculated as the ratio of the mean square regression (MSR, reflecting the variation of Y-chromosome concentration explained by the model) to the mean square error (MSE, reflecting the unexplained variation):

$$F = \frac{MSR}{MSE} = \frac{\sum \left(\hat{Y}_i - \overline{Y}\right)^2 / p}{\sum \left(Y_i - \hat{Y}_i\right)^2 / (n-p-1)} \tag{4}$$

Where: p is the number of independent variables (e.g., $p = 2$ for the linear model, $p = 5$ for the polynomial model with quadratic and interaction terms), $\overline{Y}$ is the average of the observed Y-chromosome concentration values, $n - p - 1$ is the degrees of freedom of the error term. The F-statistic follows an F-distribution with degrees of freedom $df_1 = p$ and $df_2 = n - p - 1$ under H_0. If the calculated F-value is greater than the critical F-value (or the corresponding p-value < 0.05), H_0 is rejected, indicating the model is statistically significant.

(2) Individual Coefficient Significance Test (t-test)

The t-test was used to evaluate whether each independent variable (gestational age, BMI, or their higher-order/interaction terms) has a significant individual effect on Y-chromosome concentration.

For each regression coefficient β_j (e.g., β_1 for gestational age, β_3 for BMI), the test hypotheses are: Null Hypothesis (H_0): $\beta_j = 0$ (the j-th independent variable has no significant effect on Y-chromosome concentration), Alternative Hypothesis (H_1): $\beta_j \neq 0$ (the j-th independent variable has a significant effect on Y-chromosome concentration).

The t-statistic was calculated as the ratio of the estimated coefficient to its standard error:

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \tag{5}$$

Where $SE(\hat{\beta}_j)$ is the standard error of the estimated coefficient $\hat{\beta}_j$, reflecting the sampling variability of the coefficient estimate. The t-statistic follows a t-distribution with degrees of freedom $df = n - p - 1$ under H_0. If the absolute value of the t-statistic is greater than the critical t-value (or the corresponding p-value < 0.05), H_0 is rejected, indicating the j-th independent variable has a statistically significant effect on Y-chromosome concentration.

(3) Model Goodness of Fit Evaluation

The coefficient of determination (R^2) and adjusted coefficient of determination (R^2_{adj}) were used to measure the proportion of variation in Y-chromosome concentration explained by the regression model, thereby evaluating the model's goodness of fit.

The formula for R^2 is:

$$R^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \overline{Y})^2} \tag{6}$$

Where:

1) $\sum (Y_i - \hat{Y}_i)^2$ is the sum of squared residuals (unexplained variation),

2) $\sum (Y_i - \overline{Y})^2$ is the total sum of squares (total variation of Y-chromosome concentration).

Scatter Plot Matrix of Variable Relationships is shown in figure 1. The value of R^2 ranges between 0 and 1. A larger R^2 indicates that the model explains more variation in Y-chromosome concentration, i.e., a better fit. However, R^2 tends to increase with the number of independent variables (even for irrelevant variables), so the adjusted R^2_{adj} (which corrects for the number of variables and sample size) was further used for evaluation:

$$R^2_{adj} = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2/(n-p-1)}{\sum (Y_i - \overline{Y})^2/(n-1)} \tag{7}$$

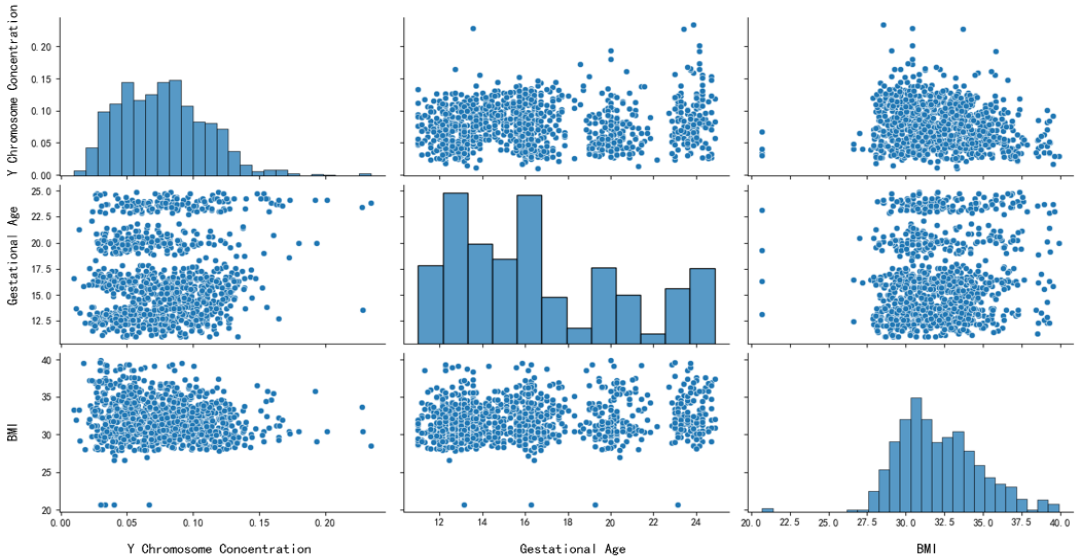

Figure 1. Scatter Plot Matrix of Variable Relationships

2.3. Mixed-Effects Model for Repeated Measurement Data

(1) Rationale for Adopting the Mixed-Effects Model

In the analysis of Y-chromosome concentration, the dataset often contains repeated measurement structures: multiple Y-chromosome concentration measurements (at different gestational age) were collected from the same individual. Traditional OLS regression assumes that all sample observations are independent, but repeated measurements from the same individual are inherently correlated (e.g., genetic background, basal metabolic level, and other time-invariant individual characteristics lead to similar Y-chromosome concentration trends for the same individual at different time points). Violating the independence assumption will cause biased estimates of standard errors of regression coefficients, thereby distorting the results of significance tests (e.g., overestimating the significance of variables) and reducing the reliability of conclusions.

The linear mixed-effects model is a specialized statistical tool for handling hierarchical or repeated measurement data. By introducing fixed effects (reflecting the average effect of population-level variables, such as the average effect of gestational age on Y-chromosome concentration across all individuals) and random effects (reflecting individual-specific deviations from the population average, such as individual differences in baseline Y-chromosome concentration or the rate of change with gestational age), the model can effectively separate and quantify population-level trends and individual variability, ensuring unbiased parameter estimates and reliable statistical inference.

(2) Structure of the Mixed-Effects Model

A two-level linear mixed-effects model was constructed, with "measurements at different gestational age" as the within-individual level (Level 1) and "individuals" as the between-individual level (Level 2).

Level 1 (Within-individual regression equation):

This level describes the relationship between Y-chromosome concentration and gestational age/BMI for a single individual:

$$Y_{ij} = \beta_{0i} + \beta_{1i} \times G_{ij} + \beta_2 \times B_i + \epsilon_{ij} \tag{8}$$

Where: Y_{ij} is the Y-chromosome concentration of the i-th individual at the j-th measurement, G_{ij} is the gestational age of the i-th individual at the j-th measurement, B_i is the BMI of the i-th individual (treated as a time-invariant variable), β_{0i} is the individual-specific intercept (baseline Y-chromosome

concentration of the i-th individual), β_{1i} is the individual-specific slope (the rate of change of Y-chromosome concentration with gestational age for the i-th individual), β_2 is the population-level fixed slope (the average effect of BMI on Y-chromosome concentration across all individuals), ϵ_{ij} is the within-individual random error, assumed to follow $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$.

Level 2 (Between-individual regression equation):

This level models the individual-specific intercept (β_{0i}) and slope (β_{1i}) as functions of population-level parameters and individual random effects:

$$\beta_{0i} = \gamma_{00} + \zeta_{0i} \tag{9}$$

$$\beta_{1i} = \gamma_{10} + \zeta_{1i} \tag{10}$$

Where γ_{00} is the population-level average intercept (average baseline Y-chromosome concentration across all individuals), γ_{10} is the population-level average slope (average rate of change of Y-chromosome concentration with gestational age across all individuals), ζ_{0i} is the individual random effect on the intercept (deviation of the i-th individual's baseline concentration from the population average), ζ_{1i} is the individual random effect on the slope (deviation of the i-th individual's concentration change rate from the population average).

The individual random effects ζ_{0i} and ζ_{1i} are assumed to follow a bivariate normal distribution:

$$\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma\right) \tag{11}$$

Where $\sum$ is the covariance matrix of random effects:

$$\Sigma = \begin{bmatrix} \sigma_{\zeta_0}^2 & \sigma_{\zeta_0\zeta_1} \\ \sigma_{\zeta_0\zeta_1} & \sigma_{\zeta_1}^2 \end{bmatrix} \tag{12}$$

In the matrix: $\sigma_{\zeta_0}^2$ is the variance of the intercept random effect (reflecting individual variability in baseline Y-chromosome concentration), $\sigma_{\zeta_1}^2$ is the variance of the slope random effect (reflecting individual variability in the rate of concentration change with gestational age), $\sigma_{\zeta_0\zeta_1}$ is the covariance between the intercept and slope random effects (reflecting whether individuals with higher baseline concentration have faster or slower concentration increases with gestational age).

Combined model equation:

Substituting the Level 2 equations into the Level 1 equation, the final form of the mixed-effects model is:

$$Y_{ij} = \underbrace{\gamma_{00} + \gamma_{10} \times G_{ij} + \beta_2 \times B_i}_{\text{Fixed Effects}} + \underbrace{\zeta_{0i} + \zeta_{1i} \times G_{ij} + \epsilon_{ij}}_{\text{Random Effects}} \tag{13}$$

2.4. Model Solution, Results, and Interpretation

(1) Model Parameter Estimation and Testing

The mixed-effects model was solved using maximum likelihood estimation (MLE), and the linear and polynomial regression models were solved using the OLS method. Detailed parameter estimates,

standard errors, t/F statistics, and p-values are provided in the supporting technical documentation (including calculation code, output reports, and result tables).

(2) Key Results

Correlation Analysis

The correlation matrix between Y-chromosome concentration, gestational age, and BMI showed:

Y-chromosome concentration was weakly positively correlated with gestational age ($r = 0.110,\ p < 0.001$)—statistically significant but with a small correlation coefficient, indicating a mild positive trend of concentration increasing with gestational age,

Y-chromosome concentration was weakly negatively correlated with BMI ($r = -0.156,\ p < 0.001$)—also statistically significant but weakly correlated, suggesting that higher BMI may be associated with slightly lower Y-chromosome concentration,

Gestational age and BMI were weakly positively correlated ($r = 0.138,\ p < 0.001$), indicating no severe multicollinearity between the two independent variables.

Linear Regression Model

Goodness of fit: $R^2 = 0.042$, adjusted $R^2 = 0.040$—the model explained only about 4% of the variation in Y-chromosome concentration, indicating limited explanatory power,

Coefficient estimates:

Gestational age: $\hat{\beta}_1 = 0.0011\ (p < 0.001)$—each additional week of gestation was associated with an average increase of 0.0011% in Y-chromosome concentration,

BMI: $\hat{\beta}_2 = -0.0022\ (p < 0.001)$—each unit increase in BMI was associated with an average decrease of 0.0022% in Y-chromosome concentration,

Overall significance: $F = 22.91,\ p = 1.82 \times 10^{-10}$—the model was statistically significant, though with low explanatory power.

Polynomial Regression Model

Compared with the linear model, the polynomial model (including quadratic terms of gestational age and BMI, and their interaction term) showed:

Goodness of fit: $R^2 = 0.068$, adjusted $R^2 = 0.063$—explanatory power improved slightly but remained low,

Coefficient estimates:

BMI first-order term: $\hat{\beta}_3 = 0.0267\ (p < 0.001)$,

BMI second-order term: $\hat{\beta}_4 = -0.000429\ (p < 0.001)$—indicating a quadratic relationship between BMI and Y-chromosome concentration (increasing first and then decreasing),

Gestational age second-order term: $\hat{\beta}_2 = 0.000131\ (p = 0.072)$—marginally significant, suggesting a potential weak quadratic trend,

Model comparison: $F = 9.65\ (p = 2.76 \times 10^{-6})$—the polynomial model was significantly superior to the linear model, but multicollinearity between higher-order terms and original variables was observed.

Mixed-Effects Model

Fixed effects:

Gestational age: $\hat{\gamma}_{10} = 0.003$ $(p < 0.001)$—the population-level average effect of gestational age on Y-chromosome concentration was positive and significant, consistent with the linear/polynomial models,

BMI: $\hat{\beta}_2 = -0.001$ $(p = 0.033)$—the population-level average effect of BMI was negative and significant, with a smaller absolute coefficient than the linear model,

Random effects:

Intercept random effect variance: $\hat{\sigma}^2_{\zeta_0} = 0.0012$ $(p < 0.001)$,

Slope random effect variance: $\hat{\sigma}^2_{\zeta_1} = 0.0008$ $(p < 0.001)$—both variances were significantly greater than zero, confirming substantial individual variability in baseline Y-chromosome concentration and its rate of change with gestational age.

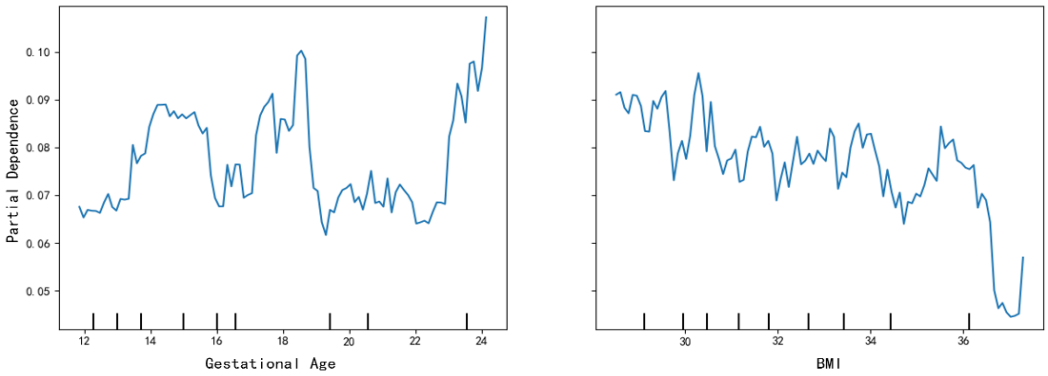

Figure 2. Partial Dependence Plot (PDP) of Gestational Age and BMI on Y Chromosome Concentration

Partial Dependence Plot (PDP) of Gestational Age and BMI on Y Chromosome Concentration is shown in figure 2.

(3) Practical Implications and Conclusions

Key role of gestational age: All three models consistently confirmed that gestational age has a significant positive effect on Y-chromosome concentration ($p < 0.001$). This is biologically plausible: as gestation progresses, the placenta expands, increasing the release of fetal cell-free DNA into the maternal bloodstream, thereby raising the proportion of fetal-derived Y-chromosome concentration.

Weak but non-negligible effect of BMI: BMI has a significant negative effect on Y-chromosome concentration ($p < 0.05$), though the effect size is small. This may be because higher maternal BMI is associated with increased plasma volume, diluting the concentration of fetal cell-free DNA, or because adipose tissue metabolism affects the clearance of fetal DNA in maternal blood.

Dominance of individual variability: The mixed-effects model revealed that individual-specific random effects (intercept and slope) account for a large proportion of the variation in Y-chromosome concentration. This explains why the linear and polynomial models have low goodness of fit—Y-chromosome concentration is not only affected by gestational age and BMI but also by individual-specific factors (e.g., genetic background, placental function, metabolic status).

Methodological value: The mixed-effects model effectively addresses the correlation of repeated measurements, providing more reliable parameter estimates than traditional OLS regression. It not only quantifies population-level trends but also captures individual differences, laying a foundation for personalized analysis of Y-chromosome concentration-related research (e.g., optimizing detection timing in non-invasive prenatal testing).

3. Conclusion

This study undertook a comprehensive statistical analysis to investigate the influence of maternal gestational age and BMI on fetal Y chromosome concentration, a critical factor for the accuracy of Non-Invasive Prenatal Testing (NIPT). The research workload encompassed extensive exploratory data analysis, the establishment and validation of multiple regression models (linear and polynomial), and the innovative application of a linear mixed-effects model to address the complex, hierarchical structure of the data involving repeated measurements from the same individuals.

Our analysis robustly confirms that gestational age exerts a significant positive fixed effect on Y chromosome concentration, a finding consistent across all employed models and supported by biological plausibility. Conversely, maternal BMI demonstrates a significant but comparatively weak negative effect. The most critical finding, however, is the identification of substantial and statistically significant inter-individual variability, as captured by the random effects in the mixed model. This variability, manifesting in both baseline concentration and the rate of change with gestational age, is the primary reason for the consistently low goodness-of-fit (R^2) in traditional regression models, indicating that these two factors alone are insufficient for precise individual prediction.

The practical application of this research is highly feasible. The mixed-effects model provides a robust methodological framework that can be integrated into clinical practice to move beyond a one-size-fits-all NIPT protocol. By acknowledging and quantifying individual differences, our findings strongly advocate for and provide the foundation for developing personalized testing strategies.

For future directions, research should focus on personalized grouping studies that incorporate a wider array of individual-specific factors (e.g., genetic background, placental function markers) to build more accurate predictive models. Furthermore, clinical implementation research is needed to translate these models into practical guidelines for determining patient-specific optimal testing timepoints based on their unique characteristics, ultimately balancing the critical needs of diagnostic accuracy and timely intervention.

References

[1] Zhang Zhenzhen, Chen Weiqing, Ding Ying. Application Value of Non-Invasive Prenatal Testing Combined with Beta-Subunit of Chorionic Gonadotropin and Serum Pregnancy-Associated Protein A in Predicting Fetal Chromosomal Abnormalities in Pregnant Women at 9–13+6 Weeks Gestation [J]. Chinese Journal of Maternal and Child Health, 2025, 40 (12): 2165-2168.

[2] Wang Qian. Value of Fetal Nuchal Translucency Measurement in Ultrasound Screening for Chromosomal Abnormalities [J]. Primary Medical Forum, 2025, 29 (15): 52-55.

[3] Wang Li, Geng Xiuxiu, Wu Tingting, et al. Application of Chromosome Microarray Analysis Technology in the Evaluation of Fetuses with Abnormal Ultrasound Soft Markers [J]. Journal of Naval Medical Science, 2025, 46 (05): 518-520.

[4] Guo Cheng, He Li, Chen Guixiu. Relationship between serum phosphatase and tensin homolog protein, fibroblast growth factor-23 concentrations with deletion of chromosome 10 in coronary heart disease patients and in-stent restenosis after interventional treatment [J]. Lingnan Journal of Cardiovascular Diseases, 2025, 31 (03): 221-226.

[5] Liu Lina, Wu Heming, Zheng Zhiyuan, et al. Chromosomal Microarray Analysis of Fetuses from Advanced Maternal Age Pregnancies with Normal and Abnormal Ultrasound Findings [J]. Guangdong Medicine, 2025, 46 (04): 538-541.

[6] Zhang Chi, He Xuelian. Application of Chromosome Karyotype Analysis Combined with Low-Depth Whole-Genome Sequencing in Detecting Fetal Chromosomal Abnormalities [J]. Chinese Journal of Eugenics and Genetics, 2025, 33 (04): 862-865.

[7] Ye Xiuqin, Bian Huanjie, Chen Yang. Diagnostic Significance of Combined RIPIS/a Spectral Parameters in Venous Catheter Blood Flow and NT Measurement for Fetal Chromosomal Abnormalities [J]. Chinese Journal of Maternal and Child Health, 2025, 40 (09): 1680-1683.

[8] [8] Zhao Jia, Song Shuo. Research Progress on Maternal Urinary Cell-Free DNA in Fetal Chromosomal Abnormality Detection [J]. Women and Children's Health Journal, 2025, 4 (05): 27-31+36.

[9] Liu Jianzhen, Chen Hongzhen, Meng Xiangrong, et al. Application of Non-Invasive Prenatal Screening in Chromosomal Aneuploidy Screening of Twin Pregnancies and Analysis of Fetal Cell-Free DNA Concentration [J]. Chinese Journal of Prenatal Diagnosis (Electronic Edition), 2023, 15 (04): 22-26.

[10] Wang Xuechun. Effects of Different Concentrations of Colchicine on Root Morphological Characteristics and Chromosomes of Fava Beans [J]. Horticulture and Seedlings, 2023, 43 (01): 25-27.