Journal of Pharmaceutical and Medical Research
Print ISSN: 2663-1954
Online ISSN: 2663-1962
DOI: https://doi.org/10.61784/jpmr3053

REGRESSION ANALYSIS-BASED INVESTIGATION OF FACTORS INFLUENCING MALE FETAL Y CHROMOSOME CONCENTRATION AND STRATIFIED OPTIMIZATION OF OPTIMAL TIMING FOR NON-INVASIVE PRENATAL TESTING

Ting Li
Mathematics and Computer Science, Yan'an University, Yan'an 716000, Shaanxi, China.

Abstract: This study systematically analyzed the quantitative relationship between male fetal Y chromosome concentration and maternal gestational age and body mass index, aiming to optimize the timing for non-invasive prenatal testing. To investigate these relationships, gestational age data formats were standardized, and missing BMI values were calculated and imputed. Pearson and Spearman correlation analyses revealed a weak positive correlation between Weeks and, and a weak negative correlation with BMI. Further OLS regression modeling revealed that Weeks exerted a significant positive effect on Y and BMI a significant negative effect , both statistically significant. Although the linear model exhibited low explanatory power and heteroscedasticity, incorporating LOWESS nonlinear smoothing clearly revealed a "rapid-then-slow" increase pattern of Y with Weeks. This confirmed the clinical observation that Y significantly declines when BMI exceeds 30. Subsequently, to determine the optimal testing window, pregnant women were stratified into four BMI groups. Intra-group OLS regression and LOWESS fitting were performed, supplemented by sensitivity analysis using 3.5%, 4.0%, and 4.5% thresholds. Results indicated that detection was feasible around 11–12 weeks for low-BMI women; while high-BMI pregnant women required delaying until approximately 12.5 weeks to achieve stable detection rates. This study provides quantitative evidence for optimizing NIPT testing strategies in clinical practice.
Keywords: Non-invasive prenatal testing; Regression analysis; Optimal testing timing; Y Chromosome Concentration

1 INTRODUCTION

With continuous advances in medical technology, non-invasive prenatal testing (NIPT) has become a vital tool for assessing fetal health. By analyzing fetal cell-free DNA fragments from maternal blood samples, NIPT enables early detection of fetal chromosomal abnormalities, thereby promoting the healthcare system's shift toward a "prevention-first" approach. The accuracy of NIPT technology primarily relies on analyzing fetal Y chromosome concentration, which is significantly influenced by multiple factors including gestational age and body mass index (BMI). Clinical experience generally suggests that Y increases with gestational age, but high BMI reduces detection sensitivity, leading to lower Y[1-2]. However, whether this relationship is strictly linear and how to determine the optimal testing time for different pregnant women based on the nonlinear interactions of these factors remain critical unresolved issues. This study aims to establish a mathematical model that accurately quantifies the relationship between Y and factors like gestational age and BMI. Through BMI grouping and timing sensitivity analysis, it seeks to determine optimal non-invasive prenatal testing timepoints for pregnant women of different body types, thereby enhancing test accuracy and clinical feasibility. The primary innovations of this section are: 1. Employing a combined strategy of correlation analysis, OLS regression, and LOWESS nonparametric smoothing to balance interpretability and flexibility, comprehensively revealing the linear and nonlinear influence patterns of gestational age and BMI on Y. 2. Addressing the Y threshold issue, we introduced BMI grouping and threshold sensitivity analysis modeling strategies to evaluate BMI's impact on optimal testing timing and quantify measurement error uncertainty[3-4]. 3. It is explicitly demonstrated that LOWESS nonlinear fitting better reflects the true trend of Y as a function of Weeks and BMI, avoiding the failure of linear regression in specific BMI groups. This section's research plan comprises two phases: First, OLS regression and LOWESS smoothing were used to explore the quantitative relationship between Y and gestational weeks/BMI. Second, pregnant women were grouped into four BMI categories, and grouped OLS regression, LOWESS fitting, and threshold sensitivity analysis were employed to determine and optimize the optimal NIPT detection timing for each BMI group[5].

2 MODELING THE ASSOCIATION BETWEEN Y CHROMOSOME CONCENTRATION AND GESTATIONAL AGE/BMI AND ANALYSIS OF NONLINEAR TRENDS

In Non-invasive Prenatal Testing (NIPT), fetal Y-chromosome concentration (Y_conc) is a crucial indicator for determining test effectiveness. Clinical experience generally holds that: as gestational age (Weeks) increases, Y-chromosome concentration gradually rises; while in pregnant women with a high Body Mass Index (BMI), test sensitivity decreases and Y-chromosome concentration is relatively low.
However, it is questionable whether a simple linear model can capture this relationship:

(1) The relationship between gestational age and concentration may not be strictly linear: Y-concentration rises rapidly in early gestational weeks, then slows down in the second trimester, possibly exhibiting diminishing marginal effects.
(2) The effect of BMI may have a threshold effect: it has little impact within the normal range (BMI < 28); but becomes significantly enhanced in obese pregnant women (BMI ≥ 30).
(3) Sequencing errors and individual differences: cfDNA fragmentation, sequencing depth, and individual differences among pregnant women may introduce random noise, leading to low linear correlation coefficients.
Based on the above considerations, we cannot rely solely on Pearson correlation or simple OLS regression, but should adopt a combined strategy of correlation analysis + OLS regression + LOWESS smoothing:
Correlation analysis is used to test the overall direction and significance[6-7];
(1) OLS regression is used to quantify the marginal effects of Weeks and BMI and conduct statistical tests;
(2) LOWESS smoothing is used to reveal potential non-linear patterns as a supplementary validation of the model.

2.1 Data Preprocessing

After reading the original table, we conducted a preliminary inspection of the data and identified several issues that need to be addressed:
(1) Inconsistent formats of gestational age field: Forms such as "12w+3" and "12w" exist, which will cause numerical calculation errors if not unified.
Processing: Write a parsing function to unify these representations into floating-point weeks (e.g., 11w+6 → 11 + 6/7 weeks) and check for records that fail to parse;
(2) BMI may be missing or need to be calculated: Some records lack the pregnant woman's BMI field but have height/weight fields.
Processing: If missing, calculate BMI using the formula:

$$\mathrm{BMI} = \frac{\mathrm{Weight}}{(\mathrm{Height}/100)^2} \tag{1}$$

(3) Missing value handling and sample screening: Exclude samples with complete missing key variables; for a small number of missing values, median/mean imputation can be used (or the impact of imputation on results can be tested in sensitivity experiments).

2.2 Correlation Analysis

Pearson and Spearman correlation coefficients were calculated for gestational age (Weeks), BMI, and Y-concentration (Y_conc) respectively. Comparison of Pearson and Spearman Correlation Coefficients are shown in table 1.

Table 1 Comparison of Pearson and Spearman Correlation Coefficients

Variables	Pearson	Spearman	Analysis
Weeks vs Y_conc	0.12	0.08	Weak positive correlation
BMI vs Y_conc	-0.16	-0.15	Weak negative correlation

Although the correlation coefficients are not large, the direction is consistent with clinical expectations: increased gestational age → slight increase in Y-concentration; increased BMI → decrease in Y-concentration.
The small coefficients may be due to: ① large measurement errors; ② non-linear rather than strictly linear relationships.

2.3 Regression Modeling

2.3.1 Introduction to regression modeling

Although the Pearson and Spearman correlation coefficients in Section 2.2 reveal a weak positive correlation between gestational age and Y_conc and a weak negative correlation between BMI and Y_conc (consistent with clinical directions), correlation analysis only reflects the marginal joint behavior of variables and cannot simultaneously control other covariates or handle repeated measurements and confounding issues. Therefore, this section naturally transitions these preliminary findings to a multiple regression model, quantitatively estimates and tests the significance of the independent marginal effects of Weeks and BMI through OLS. Given the possibility of repeated measurements and heteroscedasticity in the sample, we adopt a two-step robust strategy: first, estimate OLS parameters using HC3 heteroscedasticity-robust standard errors; second, use a mixed-effects model (with subject ID as a random intercept) as a robustness check when necessary to correct for within-sample correlation. Regression analysis can not only provide direction and effect size but also serve as a parametric starting point for subsequent stratification and non-linear analysis, thus being a reasonable next step from "correlation" to "testable and quantifiable conclusions"[8-9].

2.3.2 Establishment of OLS regression model

$$Y = \alpha + \beta_1 \cdot \mathrm{Weeks} + \beta_2 \cdot \mathrm{BMI} + \varepsilon \tag{2}$$

where Y is the Y-chromosome concentration (%), and ε is the error term.
Regression results (HC3 heteroscedasticity correction):

Coefficient of Weeks (β_1): positive and significant ($p < 0.001$);
Coefficient of BMI (β_2): -0.60, negative and significant ($p < 0.001$);
Interaction term Weeks × BMI: not significant.
Model $R^2 = 0.046$, indicating low explanatory power.
This indicates:
After controlling for BMI, gestational age has a positive correlation with concentration; after controlling for gestational age, BMI has a negative correlation with concentration.
The Breusch–Pagan test is significant ($p < 1e\text{-}10$), indicating the presence of heteroscedasticity. Therefore, HC3 estimation is used to ensure robustness.
However, the low R^2 value and residual heteroscedasticity suggest that the linear model may not fully capture all the operational patterns of gestational age and BMI. Moreover, biological mechanisms imply that Y_conc rises rapidly in early stages and then slows down with increasing gestational age, and BMI may have a strong impact at certain thresholds—these are typical non-linear and threshold effects. Therefore, it is necessary to further test and supplement regression conclusions with visualization and non-parametric smoothing methods[10].

2.4 Result Visualization and Non-linear Exploration

To further verify the model results and reveal potential non-linear patterns, we plotted visualization graphs of gestational age, BMI, and Y-chromosome concentration.

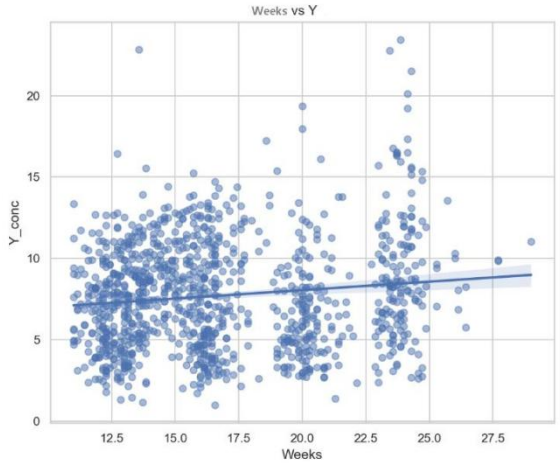

Figure 1 Scatter Plot and Linear Fitting Curve of Gestational Age vs Y-Concentration

First, Figure 1 shows that Y-chromosome concentration generally increases with gestational age, consistent with the OLS regression coefficient. However, the data points are relatively scattered, indicating that the explanatory power of a single gestational age is limited.

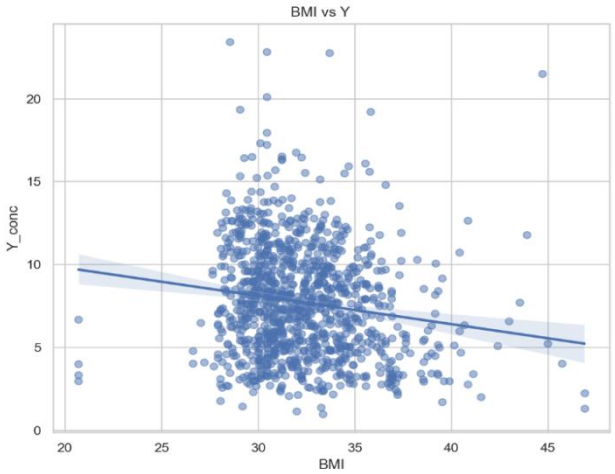

Figure 2 Scatter Plot and Fitting Curve of BMI vs Y-Concentration

Second, Figure 2 indicates that concentration decreases with increasing BMI, especially in pregnant women with BMI ≥ 35, where concentration is significantly low, posing a risk of test failure.

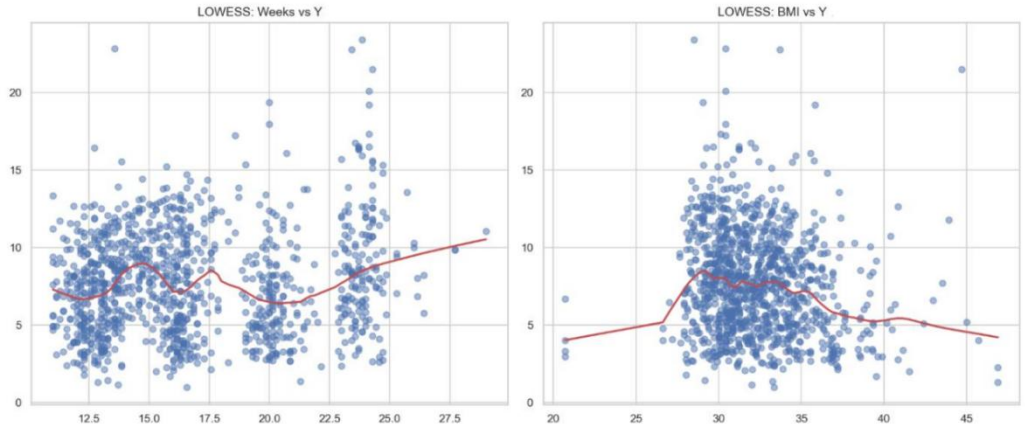

Figure 3 LOWESS Non-Linear Fitting

In terms of non-linear fitting, the LOWESS smoothing curve in Figure 3 more clearly reveals the real trend: the relationship between gestational age and concentration shows an increasing pattern of "initial rapid rise followed by slowdown", indicating that the marginal effect of Y-chromosome concentration with increasing gestational age gradually diminishes; the relationship between BMI and concentration is relatively stable in the range of BMI ≤ 25, but decreases significantly after BMI > 30, consistent with clinical understanding that obese pregnant women have a higher test risk.

Finally, Figure 4 shows that the residuals are generally randomly distributed around zero, indicating a reasonable model fit. However, there is slight heteroscedasticity in the high-concentration interval; the deviation in the tails of the QQ plot suggests that extreme samples have a certain impact on model stability.

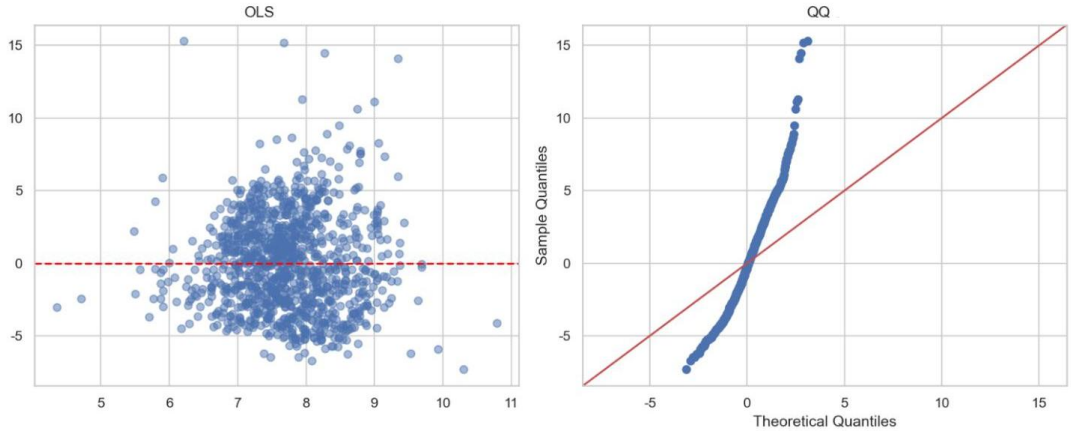

Figure 4 Residual Diagnostic Plot and QQ Plot

Overall, the visualization results are consistent with the statistical modeling conclusions: gestational age has a positive impact on Y-concentration, and BMI has a negative impact. The direction of action is clear but the linear strength is limited, and non-linear fitting better characterizes the actual variation pattern.

3 OPTIMIZATION OF OPTIMAL NIPT TESTING TIMING BASED ON BMI GROUPING AND SENSITIVITY ANALYSIS

We have found that fetal Y-chromosome concentration has an increasing trend with gestational age, while a pregnant woman's Body Mass Index (BMI) has a significant negative correlation with Y-chromosome concentration. That is, the higher the gestational age, the higher the Y-chromosome concentration; the higher the BMI, the lower the Y-chromosome concentration.

It can thus be inferred that a pregnant woman's BMI will directly affect the time point during pregnancy when the test threshold (Y = 4%) is reached, i.e., the optimal timing for Non-invasive Prenatal Testing (NIPT).

To address this issue, we proceed with the following:

(1) Grouping approach:

Divide the samples into four groups by BMI (< 28, 28−32, 32−36, ≥ 36) to compare differences in test timing among different BMI levels.

(2) Trend modeling selection:

Linear regression: If the data in a certain group is approximately linear, directly model and solve the equation Y = 4 to obtain the test timing.

Non-linear fitting (LOWESS): When the scatter plot shows an increasing trend of "initial rapid rise followed by slowdown", the linear model may be insufficient, and LOWESS is needed for smoothing fitting to capture the real changes.

(3) Threshold and sensitivity:

Set the test threshold to $Y = 4\%$, and conduct sensitivity analysis with 3.5% and 4.5% to reflect the impact of measurement errors on results.

(4) Clinical feasibility constraints:

If the solved gestational age t^* is less than 10 weeks or greater than 25 weeks, the solution is considered outside the regular test window and lacks clinical significance.

In summary, this problem adopts a combined modeling strategy of grouped OLS regression + LOWESS non-linear fitting + threshold sensitivity analysis, which not only ensures model interpretability but also improves the robustness and clinical reference value of results.

3.1 Model Establishment

Within each BMI group, fit the following linear regression model:

$$Y_i = \beta_{0,G} + \beta_{1,G} t_i + \varepsilon_i,\ \varepsilon_i \sim \text{iid}(0, \sigma^2) \tag{3}$$

where:

Y_i: Y-chromosome concentration (%) of the i-th sample;

t_i: Gestational age (weeks);

$\beta_{0,G}$: Intercept (representing the theoretical concentration at 0 weeks, only a regression parameter);

$\beta_{1,G} t_i$: Slope (representing the average increase in Y-concentration per additional week).

If $\hat{\beta}_{1,G} > 0$, the optimal test timing can be solved as:

$$t^*_{G,\text{lin}} = \frac{4 - \hat{\beta}_{0,G}}{\hat{\beta}_{1,G}} \tag{4}$$

To ensure clinical feasibility, further add corrections:

$$t^*_G = \begin{cases} 10 & \text{if } t^* < 10 \\ t^* & \text{if } 10 \le t^* \le 25 \\ \text{Invalid} & \text{if } t^* > 25 \end{cases} \tag{5}$$

In addition, conduct sensitivity analysis:

$$t_{\text{low},G} = \frac{3.5 - \hat{\beta}_{0,G}}{\hat{\beta}_{1,G}},\ t_{\text{high},G} = \frac{4.5 - \hat{\beta}_{0,G}}{\hat{\beta}_{1,G}} \tag{6}$$

Non-linear Supplement

Since the actual trend may be "initial rapid rise followed by slowdown", we introduce LOWESS smoothing:

Fit $\tilde{f}_G(t)$ on the group data;

Numerically search for the minimum gestational age $t^*_{G,\text{lowess}}$ that satisfies $\tilde{f}_G(t) \ge 4$.

LOWESS does not require a preset function form and can better characterize non-linear changes as a supplement to linear regression.

3.2 Solution Results

Optimal NIPT Calculation Results for Different BMI Intervals is shown in table 2.

Table 2 Optimal NIPT Calculation Results for Different BMI Intervals

BMI Group	Average BMI	Optimal Timing (Y=4)	Lower Bound (Y=3.5)	Upper Bound (Y=4.5)	Optimal Timing (Y=4, LOWESS)
BMI < 28	26.14	-	-	10.43	12.14
28 ≤ BMI < 32	30.23	-	-	-	11.00
32 ≤ BMI < 36	33.65	-	-	-	11.14
BMI ≥ 36	38.27	12.49	10.76	14.23	11.29

Note: The "-" in the table indicates that the gestational age solved by linear regression in this BMI group is not within the clinical window (10–25 weeks) or the solution is unstable, thus discarded.

For pregnant women with low BMI (BMI < 36), the solution of the linear regression model is unstable, but the LOWESS curve indicates that the threshold can be reached around 11–12 weeks. Pregnant women with high BMI (BMI

≥ 36) need about 12.5 weeks to reach the threshold, and the sensitivity interval is wide (about 2 weeks), indicating unstable test timing.

3.3 Visualization Analysis

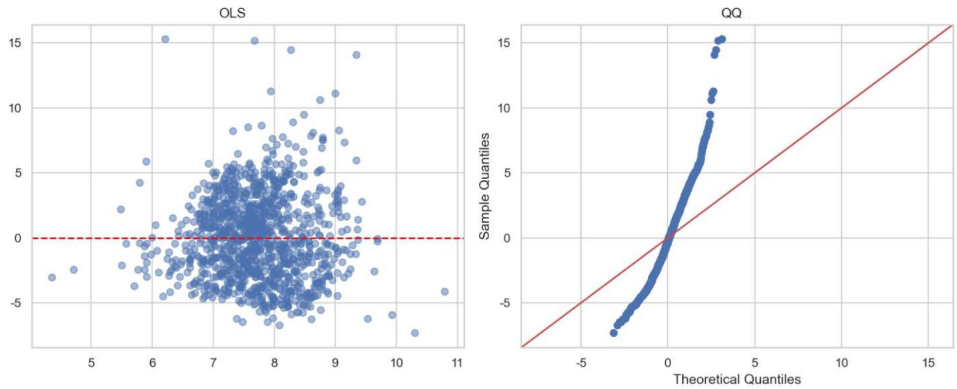

Figure 5 Trend Fitting Plots for Different BMI Groups (Linear vs LOWESS)

Figure 5 shows that the higher the BMI, the lower the overall level of Y-chromosome concentration, and the later the gestational age required to reach the threshold. The LOWESS curve better captures the increasing trend of "initial rapid rise followed by slowdown".

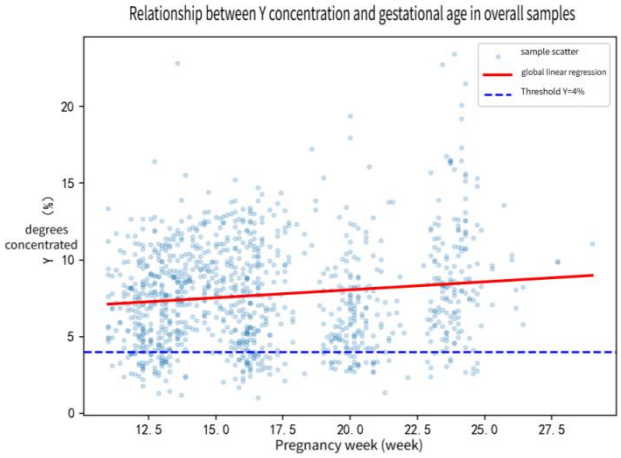

Figure 6 Overall Regression Plot

In Figure 6, Y-chromosome concentration increases with gestational age, but the data points are relatively scattered, indicating that the explanatory power of a single regression model is limited and grouping analysis is needed.

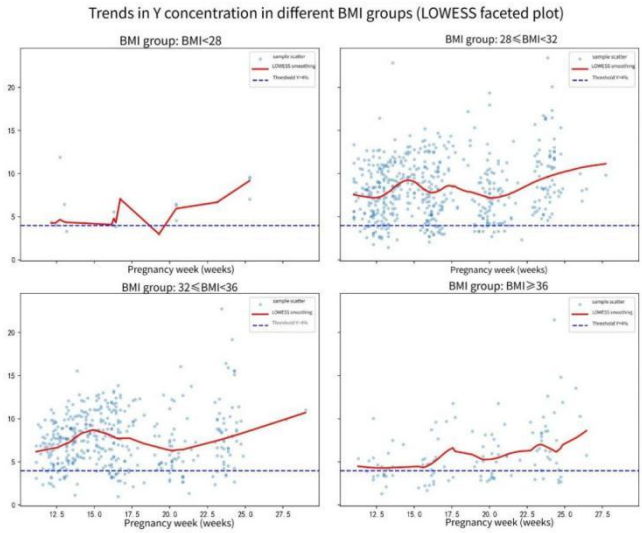

Figure 7 Grouped Y-Concentration Trend Plots

Among the four groups, Figure 7 shows that the low BMI group has a faster upward speed and can reach the threshold in the early stage; the curve of the high BMI group is relatively flat and can only stably reach the threshold after 12 weeks. Sensitivity analysis plot is shown in figure 8.

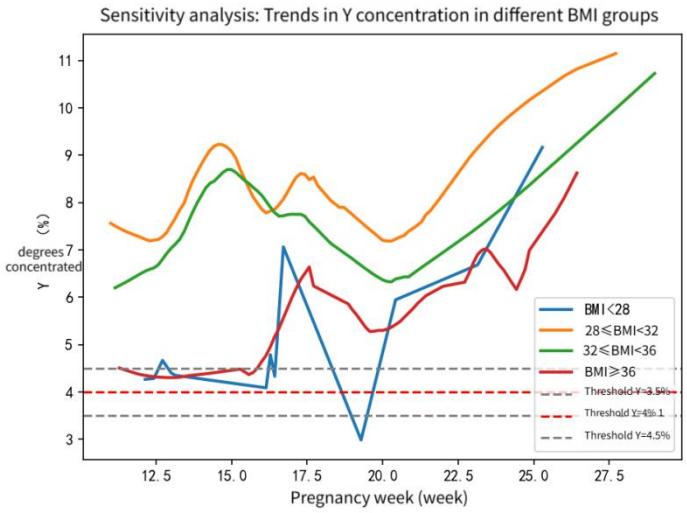

Figure 8 Sensitivity Analysis Plot

The intersection points of different groups under the three threshold lines of Y = 3.5%, 4.0%, and 4.5% are different, further verifying the significant impact of BMI on the optimal test timing.

Figure 9 shows the distribution of Y-chromosome concentration in different BMI groups: the overall level of the low BMI group is higher, and some pregnant women can meet the test requirements at 10 weeks, while the high BMI group is significantly delayed.

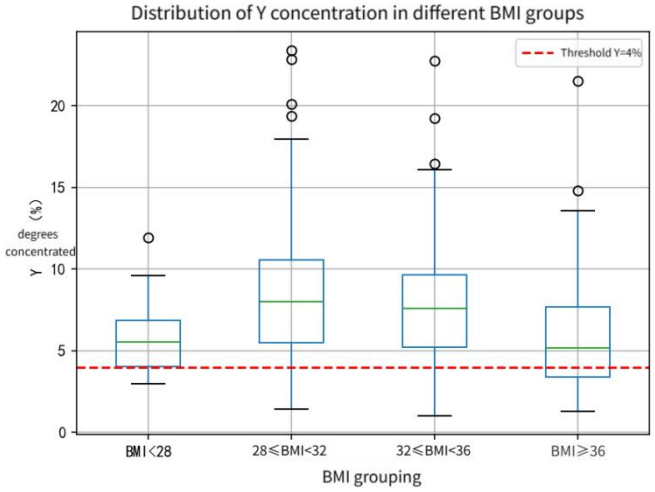

Figure 9 Box Plot of Y-Concentration Distribution by BMI Group

4 CONCLUSIONS

This study successfully investigated the quantitative relationship between fetal Y chromosome concentration and gestational age/BMI, and optimized the optimal NIPT testing time point under BMI grouping.

Regarding the relationship investigation, through OLS regression models, we quantitatively confirmed that increasing gestational age positively correlates with Y, while increasing BMI shows a significant negative correlation. Simultaneously, utilizing LOWESS nonlinear smoothing, the model revealed a "rapid-then-slow" increase pattern of Ywith gestational age, along with a threshold effect of BMI.

For optimizing the optimal detection timing, we categorized pregnant women into four BMI groups and applied LOWESS fitting within each group. This identified optimal detection timings for different BMI groups: low-BMI women reached the detection threshold around 11−12 weeks, while high-BMI women required delaying detection until approximately 12.5 weeks to achieve stable compliance. Furthermore, sensitivity analyses confirmed BMI's significant influence on optimal timing selection.

Although this model holds significant clinical value in analyzing trend directions and optimizing stratified detection timepoints, the following limitations remain:

1. Limited explanatory power of the linear model: The low coefficient of determination R^2 in the OLS regression model indicates that variations in Y chromosome concentration are influenced by numerous unaccounted factors, limiting the model's ability to explain single variables.
2. Insufficient Model Stability: In the subgroup analysis, linear regression solutions for some low BMI groups were unstable, highlighting limitations of linear assumptions. Although supplemented with LOWESS curves, more robust parametric models are needed.
3. Heteroscedasticity: The Breusch–Pagan test in regression analysis was significant, indicating heteroscedasticity. Although HC3 robust adjustments were applied, residual heteroscedasticity suggests the linear model inadequately captures the full pattern of effects.
Future research should focus on enhancing the model's predictive capability and clinical interpretability. Consider employing more flexible nonlinear models, such as generalized additive models or piecewise regression, to more accurately capture the nonlinear effects of Weeks and BMI on Y. Additionally, introduce mixed-effects models based on subject ID to correct for potential confounding effects from repeated measurements within samples, thereby enhancing the robustness of parameter estimates.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Ju Aiping, Meng Xiangrong, Qin Yanling, et al. Application Value of Non-Invasive Prenatal Testing in Screening Fetal Chromosomal Copy Number Variations. Practical Electrocardiography and Clinical Diagnosis and Treatment, 2025, 34(05): 665-671.
[2] Zhang Y C, Zhang W, Liu K B, et al. Analysis of Prenatal Screening and Diagnosis for Children with Trisomy 21 Syndrome. Clinical Laboratory and Clinical Medicine, 2025, 22(19): 2716-2720.
[3] Li Ling, Ji Yunpeng, Wang Xiaohua. Diagnostic Significance of CMA and Amniotic Fluid Karyotyping Analysis for NIPT High-Risk Fetuses. Chinese Journal of Eugenics and Genetics, 2025, 33(09): 2006-2011.
[4] Zeng Zimei, Huang Weitong, Gan Zhiyuan, et al. Application Value of CNV-seq Sequencing Technology in Prenatal Diagnostic Indications. Chinese Journal of Prenatal Diagnosis (Electronic Edition), 2025, 17(03): 9-12.
[5] Wang Jiaxuan, Li Lin. Prenatal Diagnosis and Clinical Characteristic Analysis of a Low-Proportion Mosaic Trisomy 16 Fetus. Chinese Journal of Prenatal Diagnosis (Electronic Edition), 2025, 17(03): 52-56.
[6] Liu S, Ren W, Chen L, et al. Constructing urchin-like TiO_2 integrated NiPt nanoparticles for boosting the decomposition of hydrazine hydrate. Rare Metals, 2025, 44(09): 6331-6342.
[7] Jing Yaling, Mou Yan, Zhao Lianfang, et al. Genetic analysis of a false-negative NIPT-Plus result in a rare fetal karyotype 46,XY,psuidic(21)(q22.3) case. Chinese Journal of Eugenics and Genetics, 2025, 33(08): 1827-1831.
[8] Bu Qiang, Jin Xinglin, Wang Zhen. Comparative Analysis of Amniotic Fluid Chromosomal Karyotypes and Pregnancy Outcomes in Pregnant Women of Different Ages. Journal of Clinical and Experimental Medicine, 2025, 24(16): 1745-1749.
[9] Shi Weihui, Xu Chenming. Application Value of Non-Invasive Prenatal Testing in Diagnosing Obstetric Maternal Complications and Concomitant Conditions. Journal of Practical Obstetrics and Gynecology, 2025, 41(08): 617-620.
[10] Zhang Liangliang, Zhuo Zhaozhen, Huang Shengwen, et al. Retrospective Analysis of NIPT-plus Results in 16,798 Cases from a Multicenter Study in Guizhou Province. Guizhou Medicine, 2025, 49(08): 1296-1299.