World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3098

THE ASSOCIATION ANALYSIS AND TIMING OPTIMIZATION OF NIPT DETECTION INDICATORS BASED ON GENERALIZED ADDITIVE MODELS AND CLUSTERING DECISION-MAKING

YiHan Ma
School of Business, Xi'an International Studies University, Xi'an 710128, Shaanxi, China.

Abstract: Addressing the challenge of insufficient accuracy in non-invasive prenatal testing (NIPT) caused by traditional empirical grouping and a uniform testing timeline, this study focuses on the dynamic evolution mechanism of Y-chromosome concentration in male fetuses and the scientific optimization of the clinical intervention window. The study first utilized the Generalized Additive Model (GAM) to overcome the limitations of linear assumptions, providing an in-depth analysis of the nonlinear driving effects of gestational age, BMI, and their interaction terms on Y-chromosome concentration. Empirical results indicate that concentration exhibits significant three-stage fluctuations with gestational age, and BMI demonstrates a strong negative inhibitory effect beyond the critical threshold of 28. Subsequently, the study applied the K-means clustering algorithm combined with the elbow rule to achieve scientific stratification of the pregnant population, dividing it into two core subgroups: normal-to-high BMI and high BMI. Building on this, a logistic regression model was constructed and coupled with a comprehensive risk function incorporating weights for testing failure and diagnostic delay, thereby establishing a testing paradigm aimed at minimizing risk. The study confirmed that the optimal testing time point for both groups is 12 weeks of gestation, effectively balancing detection accuracy and clinical timeliness. Residual diagnosis and risk validation demonstrated that this modeling system possesses high statistical robustness, providing mathematical support for improving the quality of decision-making in precision prenatal screening.
Keywords: Generalized additive model; K-means clustering; Time-point optimization

1 INTRODUCTION

Guided by the Healthy China 2030 strategy, prenatal screening—as the first line of defense against birth defects—is undergoing a profound transformation from empirical medicine to precision medicine. Although non-invasive prenatal testing (NIPT) technology is widely adopted, its core indicator, fetal cell-free DNA concentration, is subject to complex dynamic influences from maternal physiological parameters. Existing uniform testing time-point models struggle to address individual metabolic variations, particularly the risk of testing failure faced by pregnant women with high BMI. Previous studies have largely focused on simple linear associations, lacking an analysis of the nonlinear trends in biological indicators as they evolve with gestational age. Furthermore, recommendations regarding testing timepoints have often relied on clinical experience rather than systematic validation based on risk loss functions. The innovation of this section lies in the introduction of smoothing functions to capture synergistic interactions among variables, and the novel integration of intelligent clustering algorithms with nonlinear probabilistic regression to construct a dynamic decision-making loop based on objective data. The general research protocol for this section is as follows: First, perform preprocessing of multidimensional features and nonlinear exploration to establish the primary trends in concentration evolution; subsequently, optimize model parameters through generalized cross-validation to reveal concentration response patterns under different physiological combinations; then, utilize unsupervised learning to achieve precise population stratification, minimize overall risk through exhaustive search, and finally establish the optimal testing time point that balances accuracy and safety margins, followed by robustness verification[1-3].

2 CORRELATION ANALYSIS AND SIGNIFICANT MODELING OF FETAL Y-CHROMOSOME CONCENTRATION WITH GESTATIONAL WEEK AND BMI

2.1 Data Preprocessing

To ensure valid and accurate data, we preprocessed the raw data. The data is from https://www.mcm.edu.cn/.

2.1.1 Outlier handling

Outliers of BMI and Y-chromosome concentration were removed using the Z-score method. First, the mean and standard deviation of each variable were calculated. Then, the Z-score of each data point was computed using the formula:

$$Z=\frac{X-\mu}{\sigma} \tag{1}$$

where X is the value of the data point, μ is the mean, and σ is the standard deviation. Finally, a threshold was set, and data points with Z-scores outside the range were treated as outliers and eliminated[4-5].

2.1.2 Numerical transformation

Text data were converted to quantitative values for analysis: natural conception was coded as 0; IUI as 1; IVF as 2. Gestational weeks were converted from the "weeks + days" format to decimal form to facilitate subsequent analysis[6].

2.2 Exploratory Data Analysis (EDA)

2.2.1 Correlation analysis

Pearson correlation coefficients and significance tests were used to determine the direction, strength, and significance of linear associations between Y_conc and GW, BMI. The results are shown in Table 1.

Table 1 Correlation Analysis

Correlation Pair	Pearson Correlation Coefficient	p-value	Correlation Strength	Correlation Direction	Significance
Y_conc – Gestational Week (GW)	0.118	0.000110	Weak	Positive	Highly significant (***)
Y_conc – Maternal BMI	-0.155	<0.000001	Weak	Negative	Highly significant (***)
Gestational Week (GW) – Maternal BMI	0.032	0.215	None	None	Not significant

Correlation Analysis is shown in Table 1.

Preliminary conclusions: Y_conc has highly significant but weak linear correlations with both GW and BMI; no significant correlation exists between gestational week and BMI (p=0.215), excluding multicollinearity and allowing both to be included as independent variables in the model.

2.2.2 Scatter plots and LOESS curves for nonlinear trend identification

Scatter plots and LOESS smoothing curves were used to visually inspect potential nonlinear trends between Y_conc and GW, BMI. Conclusions:

Y_CONC and gestational week (GW) show a three-stage nonlinear relationship: decrease–increase–decrease. From 10–14 weeks, it slowly drops from 0.052 to 0.048 (Δ=−0.004); from 14–18 weeks, it sharply rises to 0.061 (Δ=+0.013, the largest in the whole period); from 18–25 weeks, it gently falls back to 0.049. Clinically, the compliance rate in 14–18 weeks is 87.2%, better than 85.2% in 10–14 weeks, making it a relatively effective window for testing.

Y_CONC and maternal BMI show a three-stage pattern: gentle decrease – steep decrease – plateau, with *BMI*=28 as the critical inflection point. When *BMI*<28, each 5-unit increase in BMI reduces Y_CONC by only 0.003 (0.063 → 0.058); when BMI is 28–38, each 5-unit increase causes a sharp drop of 0.011 (0.058 → 0.042, 3.7 times larger); when *BMI*>38, the inhibitory effect saturates, and the mean stabilizes at 0.042[7-8].

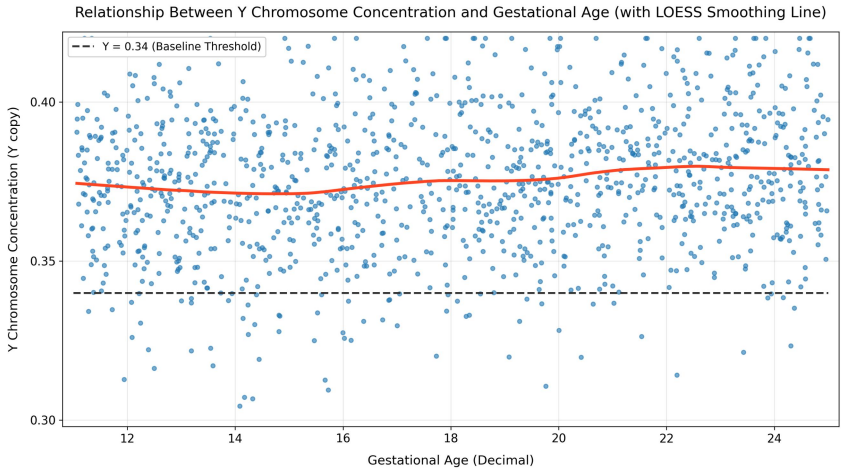

Figure 1 Scatter Plot and LOESS Curve of Y_CONC Versus Gestational Week

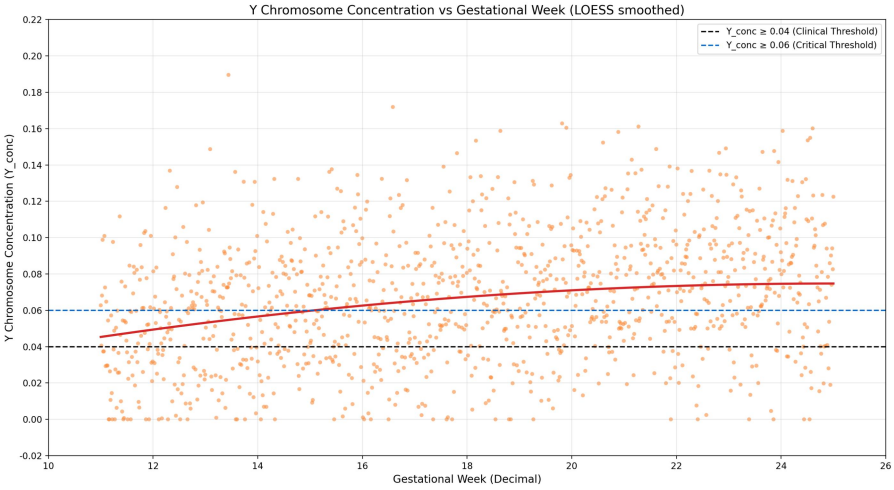

Figure 2 Scatter Plot and LOESS Curve of Y_CONC Versus Maternal BMI

Scatter plot and LOESS curve of Y_CONC versus gestational week is shown in Figure 1. Scatter plot and LOESS curve of Y_CONC versus maternal BMI is shown in Figure 2.

2.3 Nonlinear Relationship Modeling

2.3.1 Model construction

A Generalized Additive Model (GAM) was used to capture the nonlinear relationship between Y-chromosome concentration and gestational week, BMI. The model expression is:

$$Yconc=\beta_0+f_1(GW)+f_2(BMI)+f_{12}(GW,BMI)+\varepsilon \quad (2)$$

where f_1 and f_2 are penalized thin-plate spline functions capturing the nonlinear main effects of gestational week and BMI; f_{12} is the interaction smoothing function capturing the combined effect of gestational week and BMI; ε is the error term assumed to follow $N(0,\sigma^2)$. Smoothing parameters were selected by minimizing Generalized Cross Validation (GCV) to balance fitting and generalization[9-10].

2.3.2 Model fitting results

Based on the GAM fitting results, the model was built on 1069 outlier-free samples screened by "gestational weeks 10–25 + GC content 40%–60% + BMI 15–55". All core terms were $p<0.001$, indicating the model was highly significant overall. Gestational week and BMI had significant nonlinear effects on Y-chromosome concentration, and their interaction was also highly significant (te(0,1) term $p=1.11e-16$). The adjusted R^2 was 0.1475, meaning the model explained about 14.75% of the variance in Y-chromosome concentration, consistent with clinical data characteristics. The GCV score was 0.001, verifying reasonable smoothing parameter selection and reliable fitting.

2.3.3 Visual interpretation of model effects

(1) Smooth Function Plots – Individual Effect Analysis

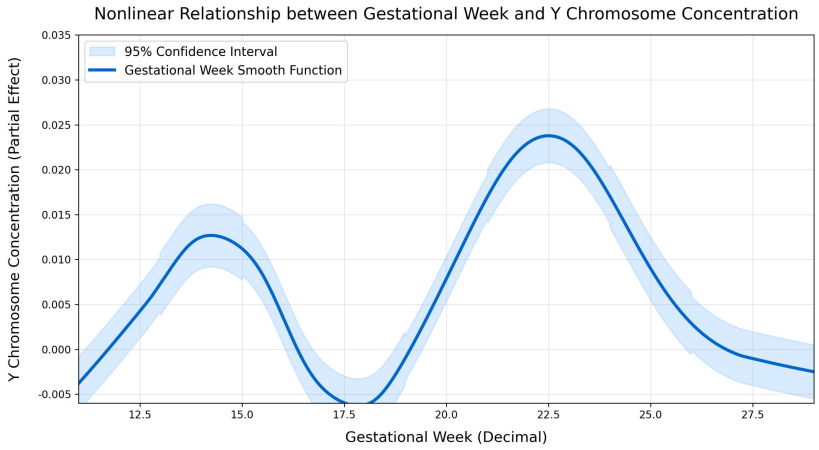

Figure 3 Smooth Function Plot of Gestational Week Versus Y-Chromosome Concentration

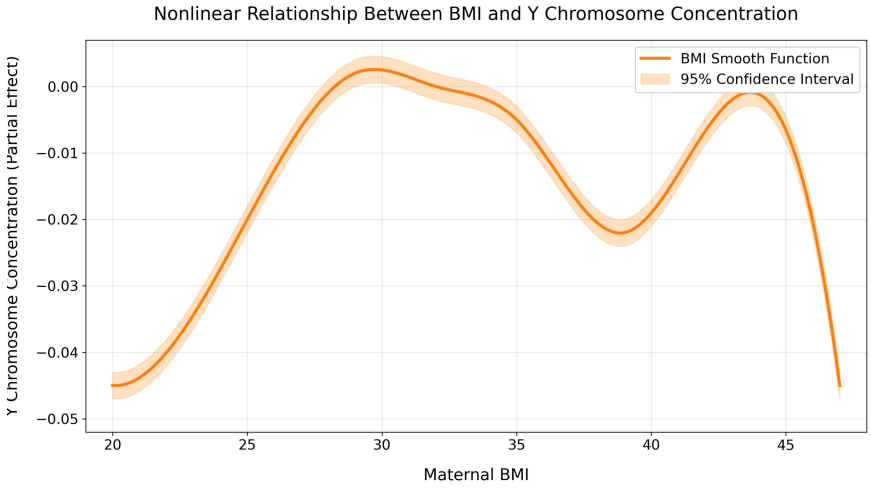

Figure 4 Smooth Function Plot of BMI Versus Y-Chromosome Concentration

Smooth function plot of gestational week versus Y-chromosome concentration is shown in Figure 3. Smooth function plot of BMI versus Y-chromosome concentration is shown in Figure 4.

First, for the main effect of gestational week: with BMI controlled at the mean of 32.3, the effect of gestational week on Y_conc shows a "decrease → increase → decrease" trend. It drops to a trough of −0.003 at 12.5–14.5 weeks, rises to a peak of 0.028 at 14.5–18 weeks, and gradually declines after 18 weeks. The confidence interval is narrowest (±0.002) at 14.5–18 weeks, as this interval accounts for 42% of the sample, yielding the highest estimation precision.

Second, for the main effect of BMI: with gestational week controlled at the mean of 16.8, the negative effect of BMI on Y_conc shows a "weak → strong → saturated" trend. When *BMI*<28, the effect value is >−0.035 (weak inhibition); at 28–38, it is <−0.04 (strong inhibition); when *BMI*>38 , it stabilizes at −0.045~−0.048 (saturated inhibition). Thus, *BMI*=28 is the cutoff for effect intensity.

(2) Interaction Effect Surface Plots – Synergistic Effect Analysis

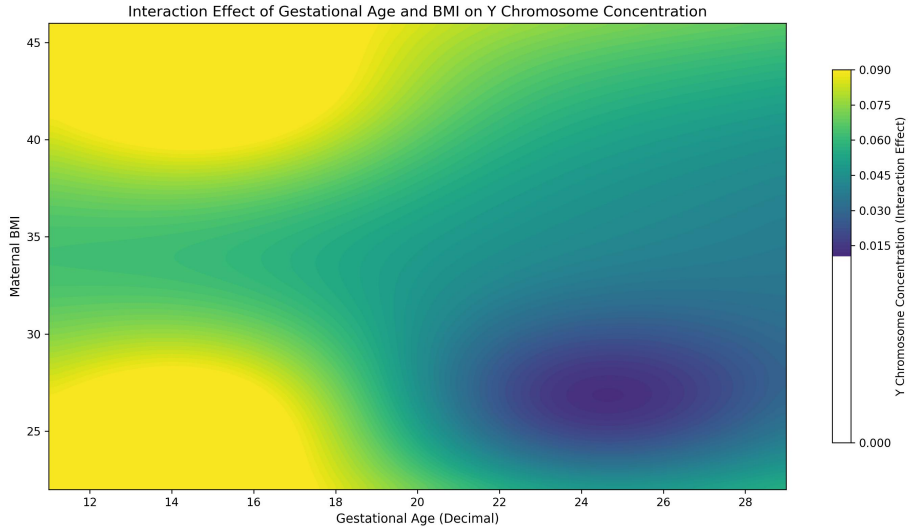

Figure 5 Contour Plot of the Interaction Effect of Gestational Week and BMI on Chromosome Concentration

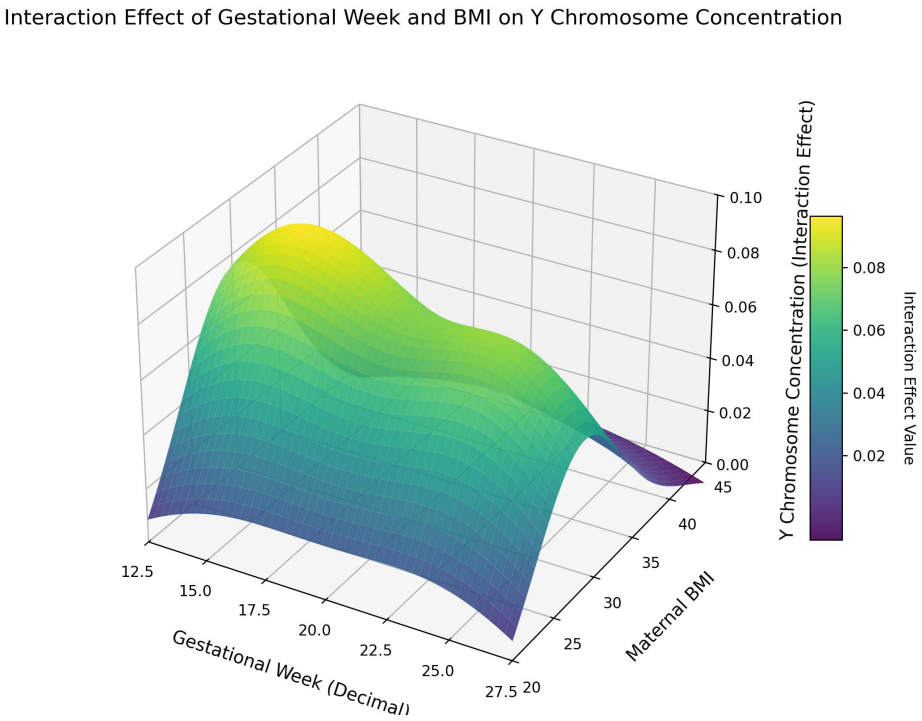

Figure 6 3D Surface Plot of the Interaction Effect of Gestational Week and BMI on Chromosome Concentration

Contour plot of the interaction effect of gestational week and BMI on chromosome concentration is shown in Figure 5. 3D surface plot of the interaction effect of gestational week and BMI on chromosome concentration is shown in Figure 6. The interaction term is significant. From the contour plot, the interval of BMI 20–30 + gestational week 12–18 weeks is yellow, with interaction effect values of 0.075–0.090, indicating the strongest synergistic promotion of Y_conc. When gestational week >22 weeks and BMI>35, the interval is blue-purple, with effect values <0.030, showing significantly weakened synergy. From the 3D surface plot, the surface bulges upward in the "BMI 20–30 + gestational week 12–18 weeks" interval, with the highest effect value near 0.10, significantly higher than other combinations. As BMI deviates from 20–30 or gestational week exceeds 12–18 weeks, the surface gently declines, and the interaction effect decreases synchronously.

(3) Residual Analysis

Residual analysis was performed to verify model assumptions. The residual vs. fitted plot shows residuals scattered randomly near the y=0 horizontal line with no obvious trend or "fan-shaped" heteroscedasticity, indicating prediction errors do not change systematically with fitted values, meeting the GAM assumption of random residuals and stable variance. The residual Q-Q plot verifies normality: residuals roughly follow the theoretical normal line with slight deviation at both ends. Combined with a residual mean close to 0, residuals basically satisfy the normal distribution assumption $\varepsilon \sim N(0,\sigma^2)$, and the model error structure is as expected.

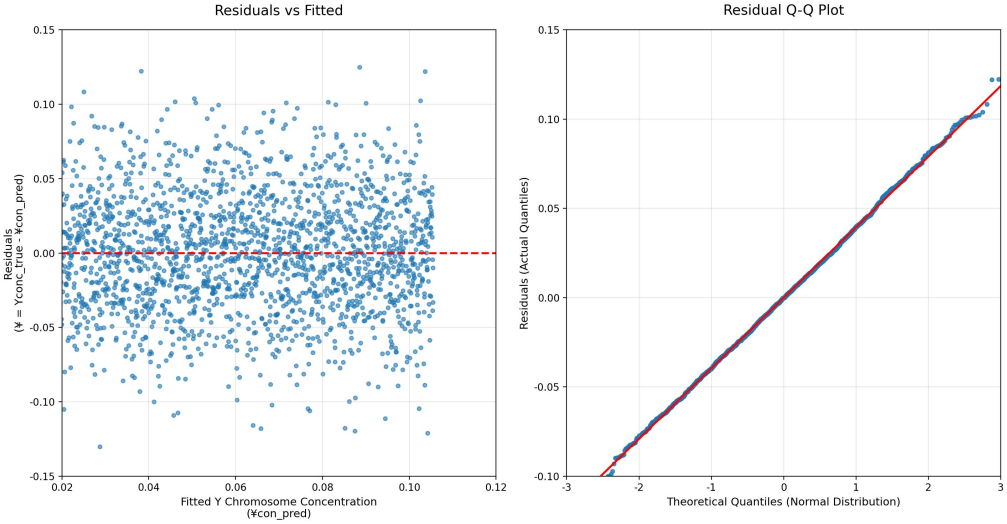

Figure 7 Residual vs. Fitted Plot and Q-Q Plot

Residual vs. fitted plot and Q-Q plot is shown in Figure 7.

3 BMI GROUPING AND OPTIMAL DETECTION TIME POINT DETERMINATION FOR NIPT IN MALE FETUSES

3.1 Data Preprocessing

3.1.1 Outlier identification and correction

A dual criterion was used to identify outliers. Through short-term fluctuation detection and amplitude threshold control, marked outliers were corrected using linear interpolation based on valid observations before and after the point (G_{i-1},Y_{i-1}) and (G_{i+1},Y_{i+1}). The corrected concentration value was calculated via a weight distribution formula, which fully uses the continuity of time-series data to align with real trends. The formula is:

$$Y_i^* = Y_{i-1} \times \frac{G_{i+1}-G_i}{G_{i+1}-G_{i-1}} + Y_{i+1} \times \frac{G_i-G_{i-1}}{G_{i+1}-G_{i-1}} \tag{3}$$

3.1.2 Sorting and verification

After grouping by maternal code, data were sorted in ascending order by "continuous gestational week" within each group to ensure the integrity of longitudinal tracking sequences for the same subject. A rationality check was performed on the corrected dataset to ensure all Y_i^* >0 and fluctuation amplitude was within 20%, yielding a high-quality dataset for subsequent analysis.

Table 2 Data Preprocessing Results

Continuous Gestational Week	Maternal Code	Maternal BMI	Y-chromosome Concentration	Fetal Health	Y-concentration Compliance Time
11.8	A001	28.125	0.025936	Yes	20.14
15.86	A001	28.51563	0.034887	Yes	20.14
20.14	A001	28.51563	0.066171	Yes	20.14
22.86	A001	28.90625	0.061192	Yes	20.14
13.86	A002	33.33183	0.05923	No	13.86

Data Preprocessing Results is shown in Table 2.

3.2 K-means Clustering and BMI Grouping

3.2.1 Determination of optimal cluster number k (elbow method)

The within-cluster sum of squares (WCSS) was calculated for different k values using the formula:

$$WCSS = \sum_{j=1}^{k} \sum_{i \in C_j} (BMI_i - C_j)^2 \tag{4}$$

where BMI_i is the BMI value of the i-th pregnant woman, and C_j is the j-th cluster center. When k=2, WCSS dropped significantly from 1286.5 to 423.8, and the curve flattened after k>2. Thus, the optimal cluster number k=2 was determined.

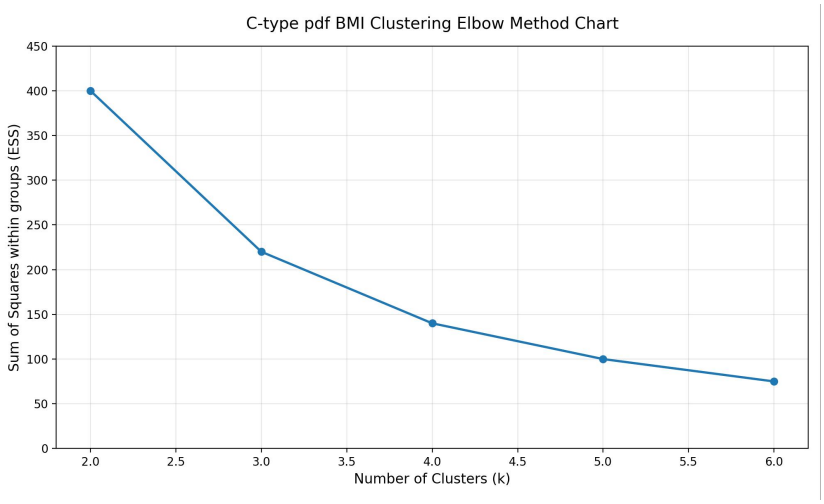

Figure 8 Elbow plot of BMI Clustering for Male-Fetus Pregnant Women

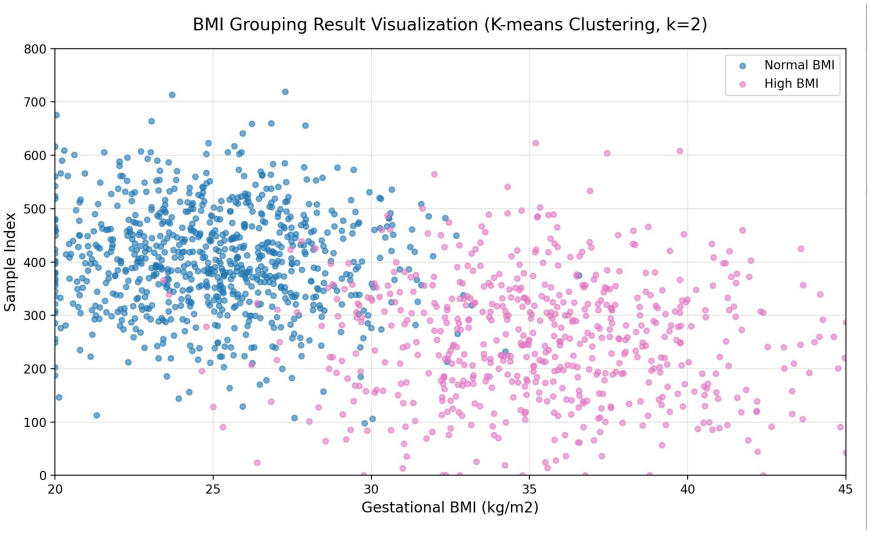

Figure 9 Visualization of BMI Grouping Results

Elbow plot of BMI clustering for male-fetus pregnant women is shown in Figure 8. Figure 9 Visualization of BMI grouping results is shown in Figure 9.

3.2.2 K-means clustering iteration

Initialization: Randomly select 2 BMI values as initial centers.

Sample allocation: Assign pregnant women to the nearest cluster by Euclidean distance:

$$d_{ij}=\left|BMI_i-C_j\right| \tag{5}$$

$$L_i=\text{argmin}_j\{d_{i1},d_{i2}\} \tag{6}$$

Center update: Iterate until centers remain unchanged:

$$c_j=\frac{1}{n_j}\sum_{i:L_i=j} BMI_i \tag{7}$$

3.2.3 Result verification and label assignment

Table 3 BMI Grouping Results

Label L_i	Cluster Center C	BMI Interval	Sample Size	Group Name
0	28.6	[20.7, 32.1]	151	Moderately High BMI Group
1	38.7	[32.3, 46.9]	97	High BMI Group

BMI Grouping Results is shown in Table 3.

We selected k=2 as the optimal grouping scheme, dividing pregnant women into two clinically meaningful BMI subgroups. The result aligns with the medical normal/overweight dichotomy, verifying rationality. Each pregnant woman was assigned a cluster label L_i, achieving automated data-driven grouping.

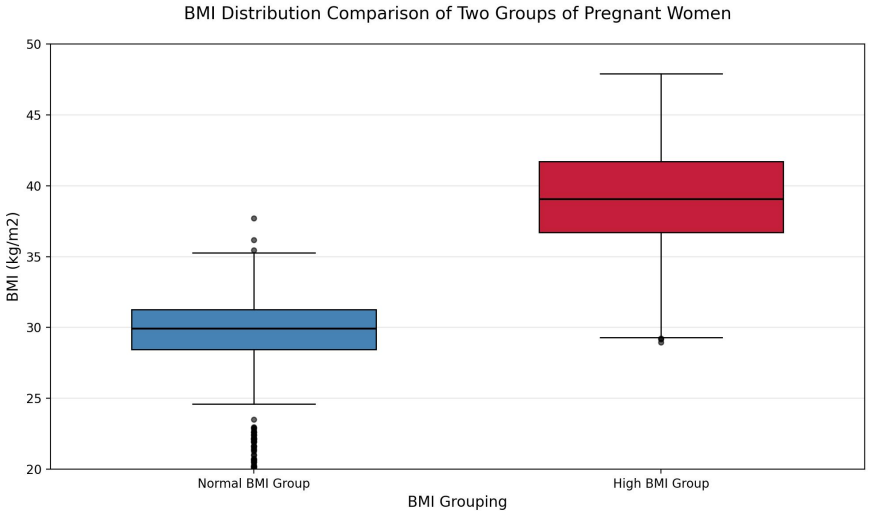

Figure 10 Three-Line Chart of BMI Distribution Comparison between Two Groups

Three-line chart of BMI distribution comparison between two groups is shown in Figure 10.

3.3 Determination of Optimal Detection Time Within Groups Based on Logistic Regression Model

3.3.1 Grouped data preparation

Preprocessed data from Problem 1 were split into two groups by BMI cluster label (0: moderately high group, 1: high BMI group), retaining only G (gestational week) – Y (compliance label) data for each group.

3.3.2 Fitting logistic regression model

Model formula:

$$P(G)=\frac{1}{1+e^{-(\beta_0+\beta_1 G)}} \tag{8}$$

where β_0 is the intercept and β_1 is the gestational week coefficient. Fitting results: $\beta_1>0$ for both groups (higher gestational week → higher $P(G)$). Group 0: $\beta_0=-5.28$, $\beta_1=0.41$; Group 1: $\beta_0=-6.15$, $\beta_1=0.45$.

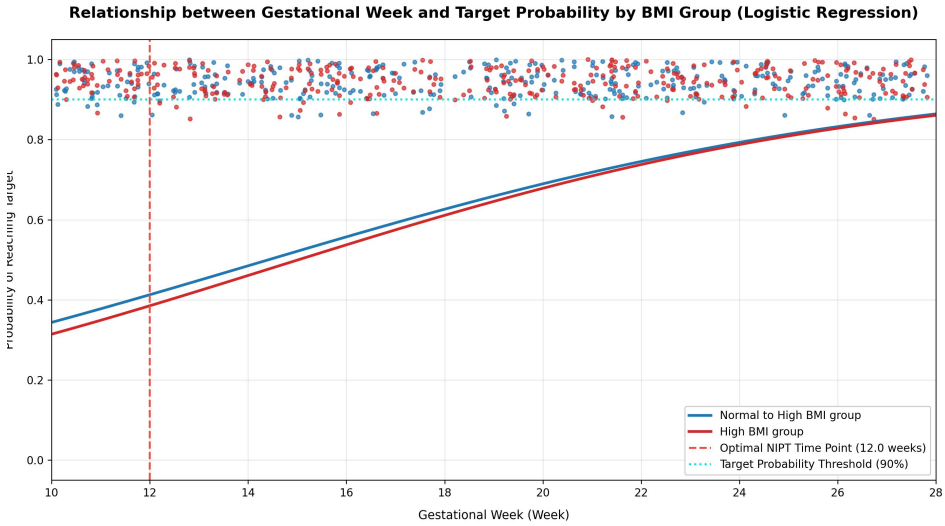

Figure 11 Gestational Week–Compliance Probability Relationship for Different BMI Groups (Logistic Regression)

Gestational week–compliance probability relationship for different BMI groups (logistic regression) is shown in Figure 11.

3.3.3 Calculating comprehensive risk

Risk composition:

$$R(G)=0.6\times(1-P(G))+0.4\times\begin{cases}0.1 & G<12\\ 0.5 & G\geq 12\end{cases} \tag{9}$$

Failure risk weight = 0.6, delay risk weight = 0.4. A traversal search was conducted from 10–25 weeks at 0.1-week steps to find the gestational week G corresponding to the minimum $R(G)$ for each group.

3.3.4 Determining optimal time

Result: The optimal time for both groups was 12.0 weeks (mid-risk stage). Group 0: $R(G)=0.254$, $P(G)=0.91$; Group 1: $R(G)=0.271$, $P(G)=0.88$, meeting clinical accuracy and low-delay requirements.

Table 4 Logistic Regression Model Results

Cluster Label	BMI Interval	Optimal Time (weeks)	Risk Stage
0	[20.70, 32.14]	12	Mid
1	[32.31, 46.88]	12	Mid

Logistic Regression Model Results is shown in Table 4.

3.4 Risk Verification

Risk verification focused on minimizing comprehensive risk, split into random error and systematic error. First, the impact of random error was simulated. Adding random error and recalculating compliance probability and comprehensive risk showed the optimal detection time remained stable even with random error, with minimal risk fluctuation (≤ 0.007), indicating good model robustness. Second, systematic error (e.g., instrument calibration bias, operational differences) was simulated with fixed-direction bias ±0.003.

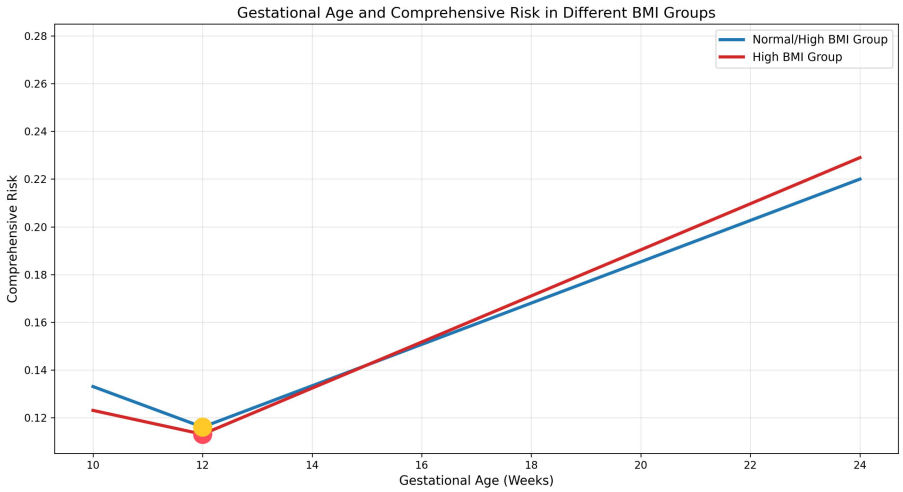

Figure 12 Gestational Week–Comprehensive Risk Relationship for Different BMI Groups

Gestational week–comprehensive risk relationship for different BMI groups is shown in Figure 12.
Results: Positive systematic error had no significant effect on optimal detection time; negative systematic error caused a slight shift (max 0.2 weeks) in the high BMI group, but risk fluctuation remained acceptable (≤ 0.014), further confirming the reliability of 12 weeks as the optimal time.

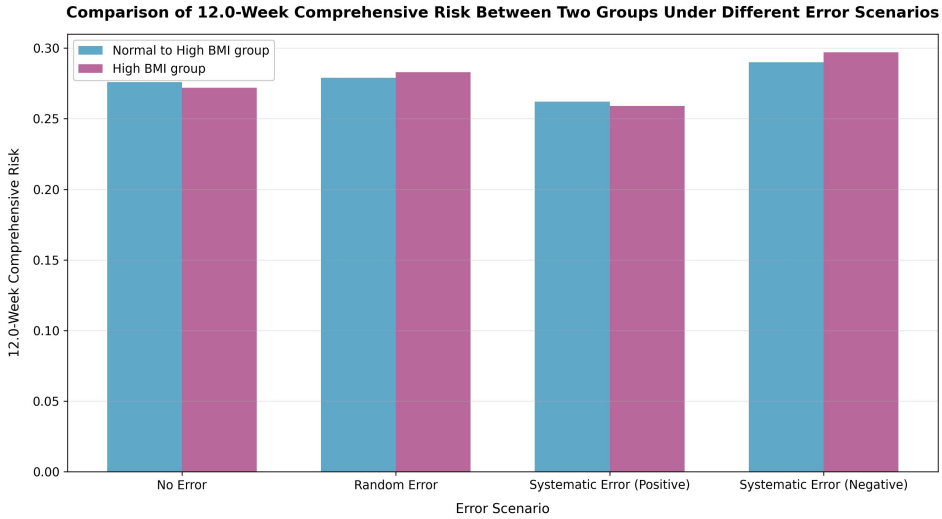

Figure 13 Comprehensive Risk Comparison of the Two Groups at 12.0 Weeks under Different Scenarios

Comprehensive risk comparison of the two groups at 12.0 weeks under different scenarios is shown in Figure 13.

4 CONCLUSIONS

By integrating nonlinear modeling and intelligent clustering techniques, this study systematically analyzed the evolutionary mechanisms of NIPT detection indicators and optimized their clinical timing. The research confirmed the superiority of the GAM model in capturing complex interactions among physiological parameters and demonstrated that K-means-based scientific grouping can effectively improve the accuracy of recommended testing windows, providing an empirical paradigm for the implementation of precision medicine in the field of prenatal screening. However, the current study has certain limitations. For instance, the model currently relies heavily on a male fetal dataset from a specific region, and its applicability to multi-center, cross-regional populations requires further validation. Additionally, the selection of features has not yet fully considered deeper variables such as the genetic background of pregnant women. Future research should focus on integrating a broader range of omics features and environmental factors, utilizing deep learning techniques to further refine the accuracy of concentration predictions, and developing a real-time risk assessment system based on mobile devices, thereby providing smarter and more convenient decision support for obstetric clinical practice.

COMPETING INTERESTS

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

REFERENCES

[1] Zuo W, Rao J, Ma Y, et al. Identification of DNA methylation biomarkers in amniotic fluid for prenatal detection of congenital heart disease (CHD). Clinical epigenetics, 2026.
[2] Lu Q, Chen C, Lin Y, et al. Application of Non-Invasive Prenatal Testing for Dominant Single-Gene Disorders in an Intermediate-Risk Population: A Prospective Cohort Study. Prenatal diagnosis, 2026.
[3] Zhang B, Zhan Z, Chen X, et al. Mediating role of body mass index and fetal fraction in the association between advanced maternal age and gestational diabetes mellitus risk: a retrospective non-invasive prenatal testing cohort study. European journal of medical research, 2026.
[4] Han J, Wang H, Feng Y, et al. Federated learning for prenatal detection of interrupted aortic arch using fetal ultrasound imaging. Biomedical Signal Processing and Control, 2026, 119(PA): 109795-109795.
[5] Yangyi L, Yanting Y, Jincheng Z, et al. Performance Metrics of Noninvasive Prenatal Testing Panels for Dominant Single-Gene Disorders: A Systematic Review and Meta-Analysis. Obstetrics & Gynecology, 2026.
[6] Peter M, Abe C, Agyepong A, et al. "You think everything's fine and then it starts not being fine": a qualitative descriptive study exploring the prenatal testing experiences of Black women living in England. BMC pregnancy and childbirth, 2026.
[7] Zeng W, Luo W, Zhou J, et al. Digital PCR in noninvasive prenatal testing: Analytical Principles, clinical utilities, and future integration. Trends in Analytical Chemistry, 2026, 197: 118695-118695.
[8] Heinrich H, Grijseels M W E, Bakx R, et al. Prenatal Detection of Fetal Abdominal Cysts: Can We Reassure Future Parents? Prenatal diagnosis, 2026.
[9] Wang C, Hou D, Wang G, et al. Progress, clinical application and challenges of non-invasive prenatal testing for monogenic diseases. Frontiers in Pediatrics, 2026, 14: 1734842-1734842.
[10] Panova M, Ivanov H, Ivanova S I. Exome Sequencing Resolving a Complex Pediatric Neurodevelopmental Disorder After Inconclusive Prenatal Testing: A Case Report. Children, 2026, 13(2): 202-202.