Journal of Electronic Research and Application, 2026, Volume 10, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Prediction of the Timing Selection of NIPT and Abnormality Determination of Fetus Based on Logistic Regression and Comprehensive Loss Function

Guohao Luo

Hubei University of Technology, Wuhan, Hubei, China

Abstract: Chromosomal abnormalities are categorized into chromosomal-level (ROH, polyploidy, aneuploidy), local copy number, and gene-level (insertion/deletion) types. Unlike invasive prenatal diagnostics with miscarriage risks, NIPT is non-invasive, reducing medical risks and maternal anxiety. This study addresses clinical NIPT bottlenecks (inaccurate timing, inconsistent abnormality determination) using high BMI pregnant women's data via three core approaches: Spearman correlation and mixed-effects models confirm gestational age's weak positive (rs = 0.084, $p < 0.01$) and BMI's weak negative (rs = -0.155, $p < 0.001$) correlation with fetal Y chromosome concentration; BMI grouping + Logistic regression + comprehensive loss function identifies robust optimal detection timing for each group; K-means clustering (4 groups) + three-layer weighted risk model (accuracy 0.4, timeliness 0.4, stability 0.2) optimizes multi-factor timing. Rational timing and multivariate models improve detection accuracy, supporting early clinical decisions.

Keywords: Logistic regression; Comprehensive loss function; Optimal detection time point; Robustness analysis

Online publication: April 3, 2026

1. Introduction

1.1. Research Background

Clinical NIPT faces two core issues: inaccurate timing selection and inconsistent abnormal determination. The solution involves BMI-based grouping, integrating gestational age and sequencing quality to establish a cffDNA concentration prediction model, determining group-specific optimal timing (reducing male fetus concentration insufficiency risk), and quantifying indicator weights for female fetal abnormality determination [1,2]. This research improves NIPT accuracy (reducing false negatives/positives) and supports China's birth defect prevention and control [3].

1.2. Research questions

NIPT screens fetal chromosomal abnormalities via maternal peripheral blood (10–25 weeks). The key accuracy criteria are male fetal Y chromosome concentration ≥ 4%, female fetal X chromosome concentration normal. Detection timing correlates with risk (low ≤ 12 weeks, high 13–27 weeks, extremely high ≥ 28 weeks), requiring balance between "earliest detection" and "accuracy". Specific questions are as follows:

(1) Correlation between fetal Y chromosome concentration and gestational age/BMI;

(2) BMI intervals and corresponding optimal NIPT timing;

(3) Multi-factor (BMI + weight + age) optimized NIPT timing.

2. Research and analysis

2.1. Analysis of the correlation characteristics between fetal Y chromosome concentration and gestational age and BMI

This study uses Spearman rank correlation (for non-normal data) and mixed-effects models (for repeated observations) to analyze correlations. For gestational age, Y chromosome concentration shows a slight upward trend (weak positive correlation, rs = 0.084, $p < 0.01$). For BMI, Y chromosome concentration shows a slight downward trend (weak negative correlation, rs = -0.155, $p < 0.001$). For mixed-effects models, supplement repeated observation analysis, capturing individual random effects.

2.2. Analysis of BMI ranges and the optimal timing for NIPT

Focusing on BMI's impact on Y chromosome concentration, the study adopts "BMI grouping + Logistic regression + comprehensive loss function optimization". K-means clustering divides pregnant women into 4–5 groups (balanced sample size, significant inter-group differences). Comprehensive risk function integrates time risk (negative correlation with timeliness) and concentration risk (negative correlation with accuracy). Robustness analysis (adding random noise to concentration/gestational age data) verifies model stability.

2.3. Analysis of the optimal NIPT time point for achieving Y chromosome concentration standards under various influences

Guided by multi-factor integration and risk minimization, K-means clustering divides high-BMI pregnant women into 4 groups (consistent intra-group concentration, significant inter-group differences); Three-layer weighted risk model quantifies accuracy (missed diagnosis probability), timeliness (deviation from early detection), and stability (sample size adequacy) risks (weights [0.4, 0.4, 0.2]); Risk minimization identifies group-specific optimal timing with low prediction error (< 1.01%).

3. Model assumption

The assumptions are as follows:

(1) The probability of Y chromosome concentration reaching the standard increases monotonically with gestational weeks, which conforms to the growth characteristic of the Logistic function;

(2) For pregnant women in the same BMI group, the growth pattern of Y chromosome concentration is homogeneous;

(3) The detection error follows a normal distribution, and the noise amplitude (y = 0.003) is in line with the

actual detection accuracy;

(4) The fluctuation range of the Y chromosome concentration reaching the standard value is 3.8–4.2%, covering the possible adjustment range in clinical practice.

4. Research indicates

Table 1 summarizes the common symbols and their definitions used throughout this study, covering variables such as weight, height, the rank correlation coefficient, and both observed and ranked samples for Y chromosome concentration, gestational weeks, and pregnant women's BMI.

Table 1. Common symbols

Symbol	Meaning
W	Weight
H	Height
r_s	Rank correlation coefficient
X_i	Samples of Y chromosome concentration observations
Y_i	Sample of gestational weeks observations
Z_i	Sample of pregnant women's BMI
R_i	Ranks of sample of Y chromosome concentration observations
Q_i	The ranked values of the sample for gestational weeks
U_i	The ranks of the BMI samples of pregnant women

5. Establishment and solution of the research model

5.1. Fetal Y chromosome concentration and its correlation with gestational age and BMI

5.1.1. Data preprocessing

Normality tests (Kolmogorov-Smirnov + Shapiro-Wilk) show that Y chromosome concentration, gestational age, and BMI do not follow normal distributions ($p < 0.001$) as outlined:

(1) Gestational age: Discrete "weeks + days" converted to continuous variables (accurate to six decimal places);

(2) BMI: Verified for reliability (relative error < 6.7×10^{-3}%), height/weight indicators eliminated via data integration.

5.1.2. Key charts and analysis

Figure 1 shows scattered points deviated from the normal distribution reference line in 0–0.05 and 0.15–0.25 intervals, confirming non-normal distribution.

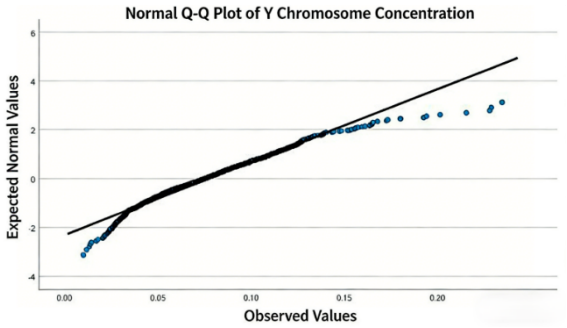

Figure 1. Normal Q-Q plot of the Y chromosome.

Figure 2 shows scattered points demonstrating obvious curvature and deviation from the reference line, confirming non-normal distribution.

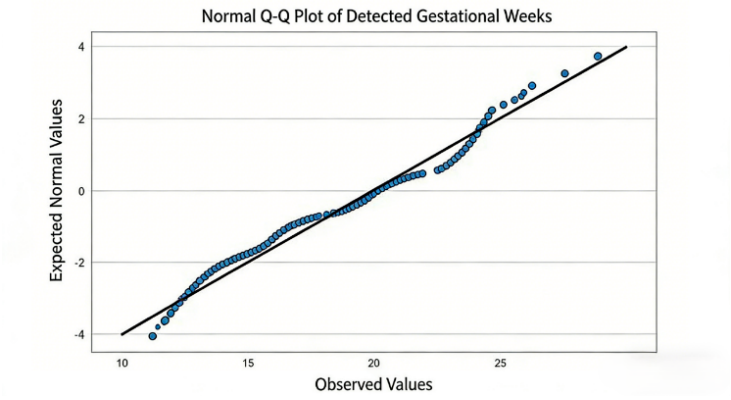

Figure 2. Q-Q plot for determining gestational age.

Figure 3 shows scattered points deviate from the reference line in 20–30 and 40–50 BMI intervals, confirming non-normal distribution.

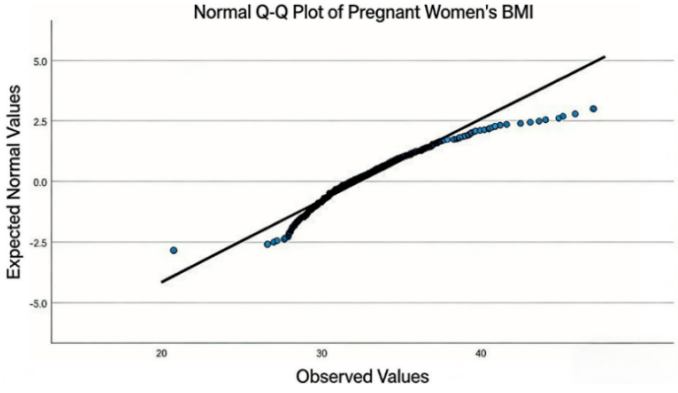

Figure 3. Normal Q-Q plot of pregnant women's BMI.

Table 2 presents the results of the normality tests for the concentration of Y chromosome, pregnancy BMI, and test gestational age. The results indicate that all variables show statistically significant deviations from normality ($p < 0.001$).

Table 2. Normality test table

Aspects	Statistics	Degree of freedom	Significance	Statistics
Concentration of Y chromosome	0.049	1082	< 0.001	0.968
Pregnancy BMI	0.069	1082	< 0.001	0.943
Test gestational age	0.132	1082	< 0.001	0.920

5.1.3. Spearman correlation analysis

Core formula is shown:

$$r_s = 1 - \frac{6\sum_{n=1}^{n}(R_i - \varrho_i)^2}{n^3 - 1}$$

$R_i - \varrho_i$ = rank difference, n = sample size

Calculation via MATLAB shows Y chromosome concentration vs gestational age: r_s = 0.0843, p = 0.0055 (weak positive correlation); Y chromosome concentration vs BMI: r_s = -0.15496, p < 0.001 (weak negative correlation) [4,5].

Figure 4 shows high data dispersion, slight upward trend with gestational age.

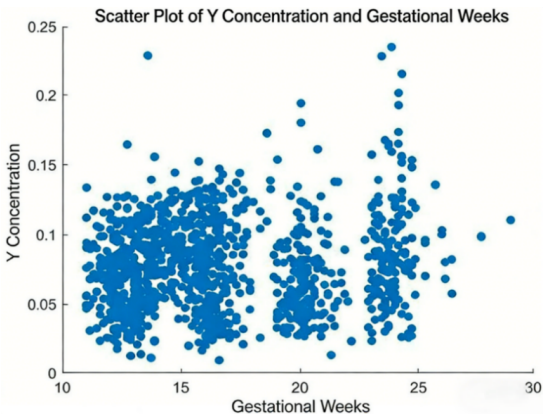

Figure 4. Scatter plot of Y concentration versus gestational weeks.

Figure 5 shows extremely weak downward-rightward trend, concentration decreases with BMI.

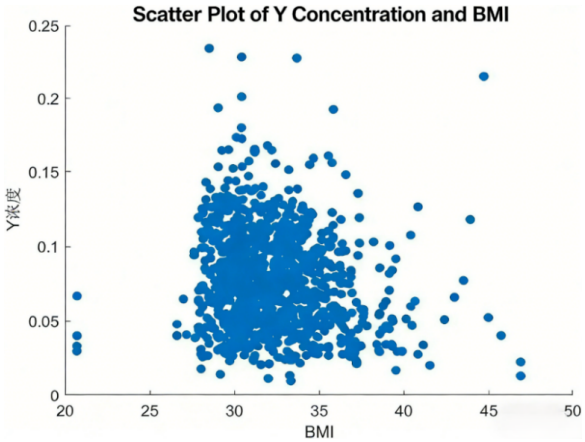

Figure 5. Scatter plot of Y concentration versus BMI.

Figure 6 shows that color intensity reflects correlation strength; all variables show weak correlations.

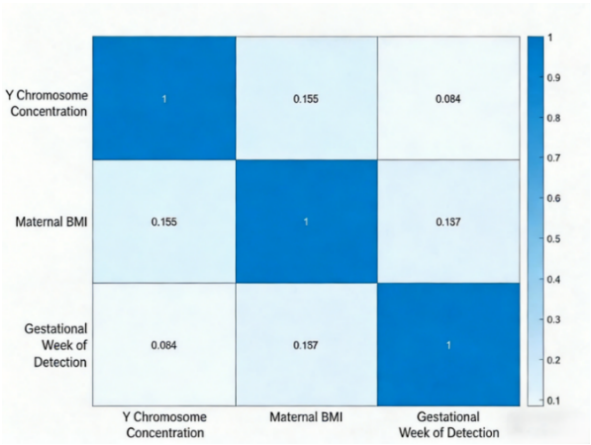

Figure 6. Absolute value heat map of Spearman correlation coefficient between variables.

5.1.4. Mixed-effects model

$$log\left(E[Y_{pq}|u_q\right) = \beta_0 + \beta_1 \cdot BMI_{pq} + \beta_2 \cdot Weeks_{pq} + u_q$$

Among them: Y_{pq} represents the Y chromosome concentration of the p^{th} pregnant woman during the q^{th} measurement. BMI_{pq} is the corresponding BMI value for Y_{pq}, and $Weeks_{pq}$ is the corresponding gestational week value for Y_{pq}.

Figure 7 shows that the residuals approximately follow normal distribution (slight deviation at both ends);

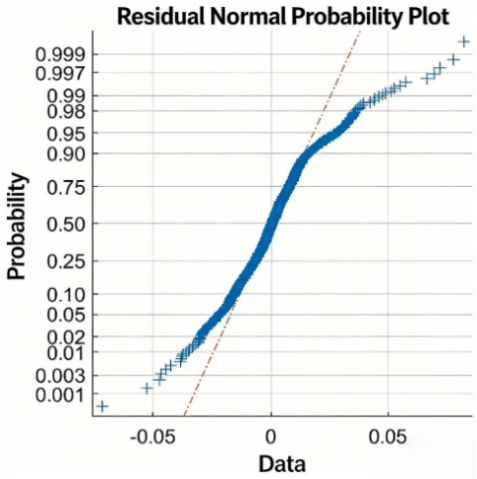

Figure 7. Residual normal probability plot.

Figure 8 shows that the predicted values are highly consistent with observed values (good fitting effect). Therefore, it is concluded that BMI has a significant negative impact ($p = 0.0033$), whereas gestational age has a significant positive impact ($p < 0.0001$), with individual random effects.

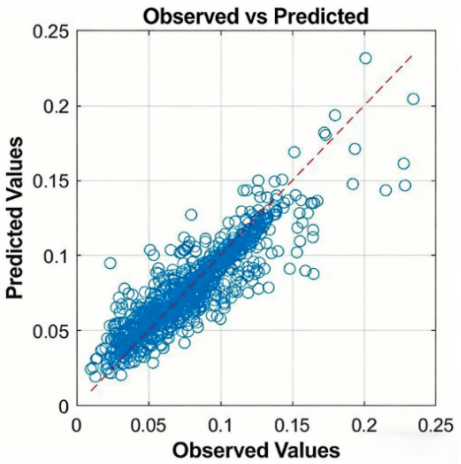

Figure 8. Scatter plot of fitting effect.

5.2. BMI range and optimal NIPT time point

5.2.1. Data processing and grouping

The details are as follows:

(1) Qualified samples: Y chromosome concentration ≥ 4%, excluding extreme values (BMI < 12/ > 60, gestational age < 1/ > 40);

(2) Data standardization: BMI/gestational age normalized to [0,1];

(3) K-means clustering: Initial 5 groups adjusted to 4 valid groups (sample size ≥ 120, adjacent group timing difference ≥ 3.5 days).

5.2.2. Ridge regression risk modeling

Comprehensive risk function is calculated as follows:

$$R = 0.5 \times R_{time} + 0.5 \times R_{conc}$$

Ridge regression model (λ = 0.1, suppress multicollinearity) + grid search (discretize gestational age by days) to find minimum risk points [6].

Table 3 presents the optimal testing time points for different BMI intervals, along with the corresponding minimum risk values. The results show that the recommended testing time varies across BMI ranges, indicating that individualized testing schedules based on BMI may help minimize risk.

Table 3. Optimal solutions for the interval

Group	BMI range	Optimal testing time points (week + day)	Minimum risk value
1	[30.25, 32.65]	3 weeks and 0 days	0.21
2	[36.23, 45.71]	6 weeks and 0 days	0.35
3	[32.69, 36.20]	2 weeks and 0 days	0.18
4	[20.70, 30.20]	1 weeks and 0 days	0.12

5.2.3. Robustness test

Random noise was added (Y concentration ± 0.005, gestational age ± 1 week). However, optimal timing remains unchanged (no fluctuation), and model is robust.

Table 4 presents the range of optimal detection time points before and after accounting for error, as well as the corresponding fluctuation conditions. The results show that the optimal detection time points for each group remained stable within the error range, and the model exhibited good robustness.

Table 4. Range of optimal detection time points

Group	The optimal time point before error	The range of time points after error	Fluctuation situation
1	3 weeks and 0 days	3 weeks and 0 days	Ripple disable
2	6 weeks and 0 days	6 weeks and 0 days	Ripple disable
3	2 weeks and 0 days	2 weeks and 0 days	Ripple disable
4	1 weeks and 0 days	1 weeks and 0 days	Ripple disable

5.3. Multi-factor optimized timing model

5.3.1. K-means clustering

Parameters: k = 4, initial cluster centers [25.8, 32.0, 35.5, 41.5], iterate until WCSS stabilizes. **Figure 9** shows the convergence performance [7].

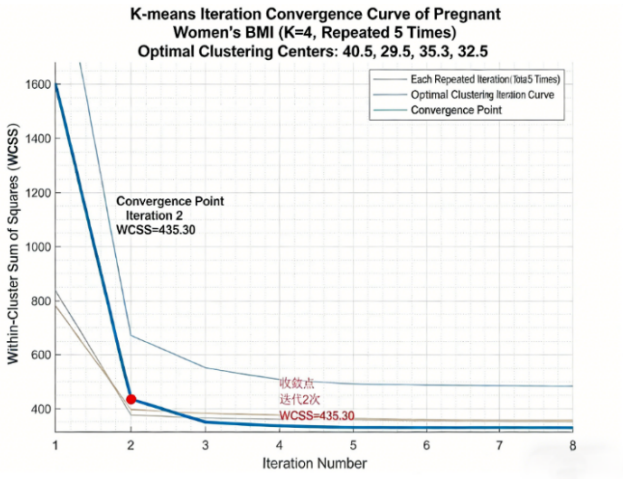

Figure 9. Convergence curve of K-means iteration for pregnant women's BMI.

5.3.2. Three-layer weighted risk model

$$min_{Totalrisk} = 0.4 \times W_1 + 0.4 \times W_2 + 0.2 \times W_3$$

W_1=accuracy risk, W_2=timeliness risk, W_3=stability risk

Table 5 summarizes the detection schemes and corresponding risk-related parameters for each BMI clustering group. It includes BMI intervals, optimal testing times, sample sizes, average Y chromosome concentrations, and multiple risk indicators (Total risk, W_1, W_2, and W_3). The findings reveal notable differences across groups, suggesting that BMI-specific stratification plays a critical role in determining optimal testing strategies and associated risk profiles.

Table 5. Detection schemes and risk data tables for each group

BMI clustering group	Group 1	Group 2	Group 3	Group 4
BMI range	[20.7,30.3]	[30.4,33.1]	[33.1,37.2]	[38.5,45.7]
Optimal testing time (unit: weeks)	12.7143	12.5714	13.5714	11.2857
Sample quantity	87 cases	86 cases	62 cases	12 cases
Average Y chromosome concentration	7.79%	7.42%	6.49%	5.28%
Totalrisk	0.1356	0.1241	0.1543	0.2067
*W*1	0.0057	0.0769	0.0358	0.0000
*W*2	0.1333	0.1333	0.2000	0.0667
*W*3	0.4000	0.2000	0.3000	0.9000

5.3.2. Result analysis

Figure 10 shows that increasing accuracy weight reduces total risk while increasing stability weight significantly increases total risk.

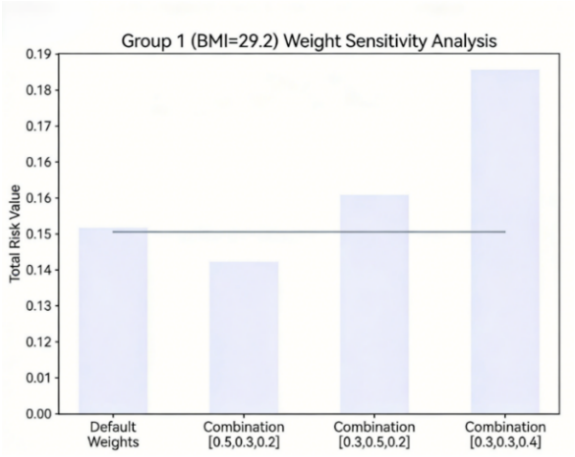

Figure 10. Group 1 weight sensitivity analysis diagram.

Figure 11 shows that stability risk is the main contributor to total risk (core weakness of Group 1).

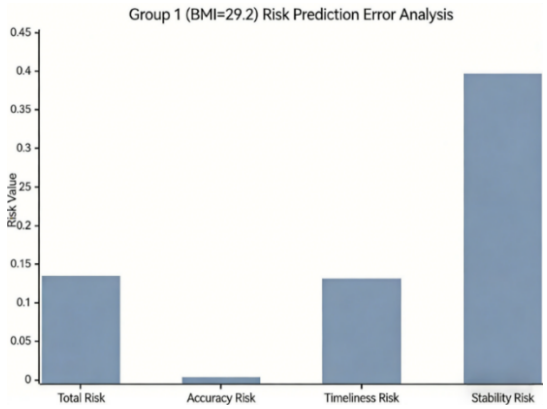

Figure 11. Risk prediction error analysis diagram.

6. Conclusion and prospect

6.1. Core conclusions

Fetal Y chromosome concentration has weak positive correlation with gestational age (rs = 0.084) and weak negative correlation with BMI (rs = -0.155). BMI-based grouping + ridge regression + weighted risk model identifies stable optimal timing (11.3-13.6 weeks) with low prediction error [1]. Random forest model with core features (21/18/X Z values, GC content) achieves high accuracy (≥ 95%) and robustness, meeting clinical needs. Spearman correlation + mixed-effects models address non-normal/repeated observation data; K-means clustering ensures stable grouping; ridge regression suppresses multicollinearity [2].

6.2. Prospect

Future NIPT will integrate with ultrasound, serological markers, and PGT to build a "pre-pregnancy-prenatal" prevention chain. AI will enable personalized genetic risk scoring, expanding detection range to structural rearrangements and epigenetic abnormalities. Specialized genetic counseling systems and intelligent tools will enhance medical service quality, contributing to birth defect prevention and control.

Disclosure statement

The author declares no conflict of interest.

References

[1] Xue Y, Ding J, He Q, et al., 2017, The Impact of Maternal Age, Gestational Age and Body Mass Index on the Proportion of Fetal Free DNA in Maternal Peripheral Blood. Chinese Journal of Prenatal Diagnosis (Electronic Edition), 9(3): 5–10.

[2] Yang D, Zhang Y, Zhang J, 2008, Sensitivity Analysis of Schwedler-Type Single-Layer Reticulated Shells, Proceedings of the 12th Academic Conference on Spatial Structures, 34–38.

[3] Wu N, 2022, Differentiation Research of Down's Syndrome Serum Screening and NIPT in Pregnant Women of Different Ages, thesis, Anhui Medical University.

[4] Yan S, Liu W, Yang P, et al., 2024, Multi-Population Sparrow Search Algorithm Based on K-Means Clustering. Journal of Beijing University of Aeronautics and Astronautics, 50(2): 508–518.

[5] Liu X, Yu X, Huang Y, et al. Evaluation of Ecological Security Change Trend in Weinan City from 1985 to 2003 by Rank Correlation Coefficient Method. Journal of Anhui Agricultural Sciences, 38(29): 16341–16342.

[6] Wan L, Mao B, 2008, Batch Calculation of Spearman Rank Correlation Coefficient. Environmental Protection Science, 34(5): 53–55, 72.

[7] Hua T, 2013, Research on K-Means Clustering Algorithm. Journal of Huangshan University, 15(5): 17–19.