Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 21, No. 2, 2026

Research on Personalized NIPT Detection Timing and Anomaly Judgment Based on Bayesian Optimization

Zhiqiao Huang[#, 1, *], Zijun Lu [#, 1], Yueming Huang [#, 2]

[1] School of Mathematical Sciences, South China Normal University, Guangzhou 510000, China,
[2] School of Economics and Management, South China Normal University, Guangzhou 510000, China

***Corresponding author:** 20232221076@m.scnu.edu.cn

#These authors contributed equally

Abstract: With rising obesity-related NIPT failures due to insufficient fetal DNA, we introduce a pioneering clinical application of Bayesian optimization—traditionally confined to engineering and machine learning—to personalized prenatal screening. Linear mixed-effects models quantified nonlinear relationships between fetal DNA concentration and gestational age, BMI, and maternal age. Gaussian mixture clustering stratified pregnancies into five BMI-specific subgroups. Uniquely applying Bayesian optimization to determine optimal 10–25-week sampling windows for each subgroup, we minimized weighted risks of early/late testing and low DNA concentration. Results reveal distinct optimal timing across BMI categories that diverge from current uniform guidelines, with higher BMI requiring earlier collection. This innovative integration of machine learning optimization into clinical obstetrics establishes a data-driven 'one person, one time' framework, representing the first utilization of Bayesian methods for individualized NIPT scheduling decisions.

Keywords: Linear mixed-effects model; Gaussian mixture model; Bayesian optimization; Personalized detection time point; Chromosomal abnormality identification.

1. Introduction

Non-invasive Prenatal Testing (NIPT) enables accurate screening for chromosomal aneuploidy syndromes such as trisomy 21, trisomy 18 and trisomy 13 by detecting cell-free fetal DNA (cffDNA) in maternal peripheral blood, and has become a core technology in clinical prenatal diagnosis [1]. The accuracy of NIPT is highly dependent on the concentration of cffDNA, and a Y-chromosome concentration of ≥4% in male fetuses is widely adopted as the clinical judgment criterion [2]. In recent years, the proportion of obese pregnant women has been rising globally. In such populations, the concentration of cffDNA is significantly reduced due to interference from maternal adipose tissue, resulting in an NIPT failure rate 3 to 5 times higher than that in normal-weight pregnant women [3].

Current clinical practice mostly adopts a unified testing time window strategy with fixed BMI grouping (e.g., 12 to 22 gestational weeks), which fails to fully consider the impact of individual differences on cffDNA concentration. This leads to a persistently high testing failure rate in the high-BMI population. Moreover, overly early testing may increase the risk of missed diagnosis, while overly late testing shortens the window for clinical intervention [4]. Therefore, developing a personalized NIPT testing time window optimization method that takes individual differences into account holds important clinical value for reducing the testing failure rate and improving the quality of prenatal diagnosis.

Existing studies have confirmed that gestational week, BMI and maternal age are key factors affecting cffDNA concentration [6-7]. Liu Lihua et al. [1] found a positive correlation between maternal age and the screening risk of Down syndrome through a linear regression model, but did not consider the random effects of individual differences. Wu Lifang et al. [2] verified the positive correlation between gestational week and cffDNA concentration, yet did not conduct differential time window analysis for different BMI subgroups. Some studies have attempted to optimize the testing time window using a grouping strategy, such as dividing BMI into three groups (normal, overweight and obese), but the grouping boundaries are ambiguous and the synergistic effects of multiple factors are not integrated [8].

As an efficient black-box optimization algorithm, Bayesian optimization has been applied to risk prediction using genomics data [3], but its application in NIPT time window optimization has not been reported yet. Gaussian Mixture Model (GMM) has demonstrated advantages in medical data grouping due to its ability to capture the potential distribution characteristics of data [9], but it has not been used for refined subgroup division of BMI.

2. Y chromosome concentration fitting model based on linear mixed-effects model

2.1. Data Sources and Preprocessing

The data in this study came from a database of NIPT tests of pregnant women with high BMI in a certain region, which included 1200 pregnant women with male fetuses and 800 pregnant women with female fetuses. The core characteristics included gestational age, BMI, age, height, weight, parity, Y chromosome concentration (male fetus), Z value of chromosomes 13/18/21, GC content, and number of reads. After preprocessing, the data is shown in Figure 1.

Based on the correlation study between maternal age and NIPT for Down syndrome in the second trimester [1], maternal age was added as a supplementary factor. A model was established to show the relationship between fetal Y chromosome concentration and gestational age, BMI and maternal age. At the same time, the correlation and significance between fetal Y chromosome concentration and gestational age, BMI and maternal age were analyzed.

Data analysis revealed nested characteristics in the data from multiple tests conducted on pregnant women. Repeated measurements of the same individual at different times or under different conditions did not meet the data independence requirements of ordinary linear models. Traditional methods, by not considering the correlation of pregnant women's data, may lead to bias. A linear mixed-effects model addresses this issue by introducing random effects on pregnant women. This model uses gestational age and BMI as fixed effects to analyze overall trends, while simultaneously reflecting individual differences through random effects, avoiding information loss caused by simple grouping. The model employs likelihood ratio tests to assess the significance of each effect, meeting the requirement for testing the significance of features.

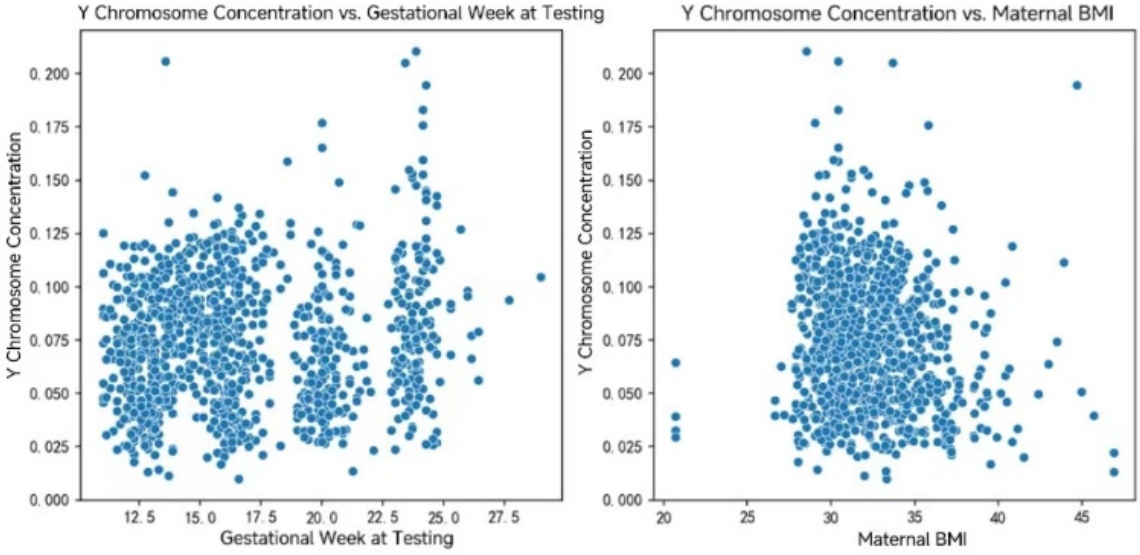

Figure 1. Distribution of male fetal data after preprocessing

2.2. Establishment and solution of a model relating maternal age, gestational age, BMI, and Y chromosome concentration

A linear mixed-effects model is a statistical model that includes both fixed and random effects. The fixed effects represent the average influence of maternal age, gestational age, and BMI on fetal Y chromosome concentration, while the random effects explain the individual impact of the pregnant woman on the model, characterized by random intercepts and random slopes.

(1): Fixed Effects

Intercept: Baseline level of Y chromosome concentration in all pregnant women.

Gestational age: It is assumed that the Y chromosome concentration of all pregnant women changes in a consistent trend with gestational age (global slope).

Age: It is assumed that the age of all pregnant women has a consistent effect on the concentration of Y chromosome (global coefficient).

BMI: Assuming that the BMI of all pregnant women has a consistent effect on the concentration of Y chromosome (global coefficient).

(2): Random effects

Intercept for each pregnant woman: Allows each pregnant woman to have a different baseline Y chromosome concentration.

Gestational slope for each pregnant woman: Allows for different rates of change in Y chromosome concentration with gestational age for each pregnant woman.

Formula for constructing a linear effects mixture model (LAM):

$$y \sim 1 + t_{check} + age + BMI + (1 + t_{check} \mid ID) \quad (1)$$

Wherein, the target variable y is the Y chromosome concentration; age is the pregnant woman's age; ID is the pregnant woman's ID number. $t_{check} + age + BMI$ represents the fixed effects part; $1 + t_{check} \mid ID$ represents the random effects part, with the individual-specific variables in parentheses and the grouping variables after the vertical line.

The LME model is fitted using MATLAB's `fitlme` function, while errors during the fitting process (such as missing data and multicollinearity) are captured using try-catch statements to ensure program stability. After successful fitting, two key results are extracted:

Fixed effects coefficients: obtained through the fixedEffects(lme) function, including the coefficients, standard errors (SE), t-statistics (tStat), degrees of freedom (DF), and p-values (pValue) of each fixed effects variable, used to determine the significance of the variable's influence;

Conditional R^2: Measures the overall goodness of fit of the model to the data, including the combined explanatory power of fixed and random effects. The formula is as follows:

$$R^2_{conditional} = 1 - \frac{\sum(y_{actual} - y_{pred})^2}{\sum(y_{actual} - \bar{y}_{actual})^2} \quad (2)$$

Where y_{actual} is the actual value of Y chromosome concentration, y_{pred} is the model prediction value, and $\bar{y}_{actual}$ is the average of the actual values.

Because some features may have non-linear relationships with the target variable, and some indicators are unordered categorical variables, Spearman rank correlation analysis was used to assess the association strength between the 23 features and Y chromosome concentration. The closer the absolute value of Rho is to 1, the stronger the correlation; a p-value < 0.05 indicates that the correlation is statistically significant.

2.3. Relationship between maternal age, gestational age, BMI, and Y chromosome concentration

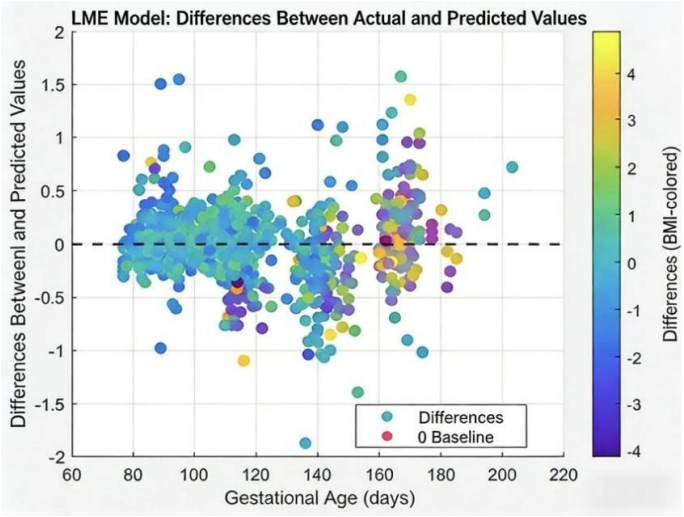

Figure 2. Difference between actual and predicted values of maternal age, gestational age, and BMI and Y chromosome concentration in the LME model.

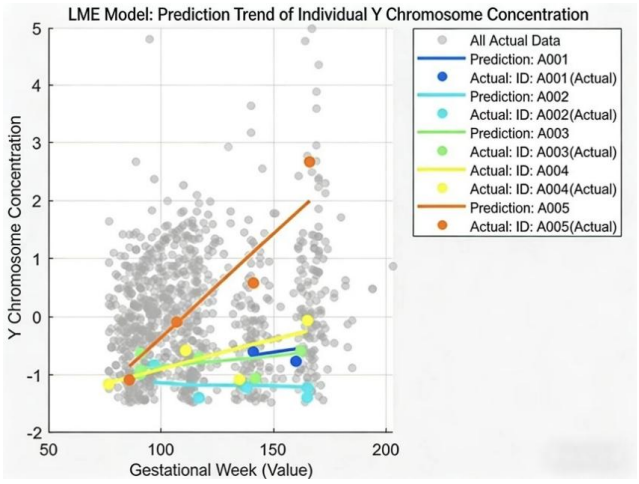

Figure 3. Pregnant woman's age, gestational age, and BMI versus LME model prediction of individual Y chromosome concentration.

Table 1. Significance analysis of fixed effects (P<0.05 is considered significant)

Fixed effects variables	coefficient	Standard error	t-statistic	p-value
intercept	-1.868	0.131	-14.301	4.66 × 10 -42
age	-0.066	0.056	-1.1813	0.2378
Gestational weeks (days)	0.016	0.001	13.612	1.37 × 10 -38
BMI	-0.099	0.051	-1.9335	0.0535

Figure 2 and Figure 3 demonstrate that the LME model performs admirably in prediction. In Figure 2, the differences between actual and predicted values are mostly distributed around the 0 baseline, and the distribution is relatively dense, indicating that the model has a small overall prediction bias and good stability. After color-coding the differences by BMI, it can be seen that the model maintains good predictive consistency across different BMI groups, demonstrating significant adaptability. In Figure 3, for different study individuals (pregnant women ID: A001 to A005), the predicted trend lines match the actual data points well. The predicted line for ID: A005 successfully captures the trend of actual Y chromosome concentration changing with gestational age, demonstrating the model's accurate grasp of individual differences and good performance in handling random effects. Overall, the LME model exhibits excellent performance in both fixed and random effects prediction.

Analysis of Table 1 shows that gestational age has a highly significant impact on fetal Y chromosome concentration; for every additional week of gestation, the fetal Y chromosome concentration increases by an average of 0.016.

Table 2. Feature Correlation Analysis

Feature Name	Correlation coefficient	p-value
Gestational weeks (days)	0.0939	0.131
X chromosome concentration	0.3686	3.01 × 10 -31
Y chromosome Z value	0.1451	9.08 × 10 -6
Age of pregnant women	-0.1166	3.71 × 10 -4
BMI	-0.1227	1.80 × 10 -4

Analysis of Table 2 shows a significant positive correlation between gestational age and fetal Y chromosome concentration, consistent with the LME model results. BMI has a significant negative correlation with fetal Y chromosome concentration, also consistent with the marginal significance of the LME model. X chromosome concentration is strongly and significantly positively correlated, making it one of the most crucial association features, while the Y chromosome Z-value is directly related to concentration. Maternal age has a significantly negative correlation with fetal Y chromosome concentration, which is not significant in the LME model and is influenced by gestational age.

3. Preliminary grouping model based on Gaussian mixture model and interval-determined time point model based on Bayesian optimization model

3.1. Establishment and solution of a preliminary grouping model for BMI in pregnant women with male fetuses

Gaussian Mixture Model (GMM) refers to a combination of functions of multiple Gaussian distributions. Its principle is to construct the most suitable mixture of multidimensional Gaussian distribution models by fitting the input dataset.

Let $x = (x^1, x^2, \cdots x^d)^T$ be a d-dimensional random variable. Then the Gaussian mixture model containing k clusters is:

$$\begin{cases} p(x) = \sum_{k=1}^{K} w_k \, N(x|\mu_k, \sum k) \\ N(x|\mu_k, \sum\nolimits_k) = \frac{1}{\sqrt{|\sum k|}(2\pi)^{d/2}} exp\left\{-\frac{1}{2}(x-\mu_k \right. \\ \sum_{k=1}^{K} w_k = 1, 0 \le w_k \le 1 \end{cases} \quad (3)$$

Where $N(x|\mu_k, \sum_k)$ is the Gaussian probability density function; $w_k, \mu_k, \sum_k$ are the weights, mean, and covariance matrices of the BMI of the k-th cluster in the mixture model, respectively; and p(x) is the probability density function of the Gaussian mixture model.

To automatically determine the optimal number of clusters, we calculated the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for different numbers of clusters. These two criteria aim to balance the model's fit and complexity; smaller criterion values indicate a better model. We selected the number of clusters that minimized both AIC and BIC values as the final number of clusters. Considering both AIC and BIC, we selected the optimal value of k for a Gaussian mixture model of BMI grouping for male fetuses in pregnant women with k clusters.

3.2. Results of preliminary grouping model of BMI in pregnant women with male fetuses

The AIC result is 3, and the BIC result is 8. Since the AIC result intervals overlap, the optimal k value is 8. A Gaussian mixture model was run to group the BMI of male fetuses in pregnant women into 8 clusters, resulting in the Table 3.

Table 3. BMI Grouping of Pregnant Women with Male Fetuses

Group	Scope
1	[20.70, 26.62]
2	[26.62, 29.51]
3	[29.55, 31.11]
4	[31.14, 32.76]
5	[32.76, 35.01]
6	[35.06, 37.83]
7	[37.84, 41.52]
8	[42.38, 46.88]

3.3. Establishment and solution of optimal NIPT time point model for pregnant women with male fetuses with different BMI groups

The optimal NIPT timing model for different BMI groups of pregnant women with male fetuses adopts a Bayesian optimization model [3]. The core objective is to minimize the weighted combined risk of "risk of failing detection too early" and "risk of delaying detection too late". The weight parameter, namely the risk balance coefficient α, has a value range of [0.6, 0.7]. The weighted combined risk function that minimizes the risks of "detecting too early" and "detecting too late" is selected.

(1): Risk of acting too early (p_{too_early}):

The probability of blood being drawn before test point t_{check} but the chromosome concentration not reaching the target reflects the risk of "undetectable chromosome concentration". The risk of drawing blood too early (p_{too_early}) is calculated using the formula:

$$p_{early}(t_{check}, Cluster\ i) = \frac{\sum_{k \in Cluster\ i} I(t'_{check} \le t_{check} \cap y_{actual} \le 0.04)}{\sum_{k \in Cluster\ i} I(t'_{check} \le t_{check})} \quad (4)$$

Where I(.) is an indicator function, equal to 1 if the condition is met, and 0 otherwise. y_{actual} is the actual value of Y chromosome concentration. t'_{check} is the gestational age. $p_{too\_early} = p_{early}$. The calculated p_{early} is shown in Figure 4.

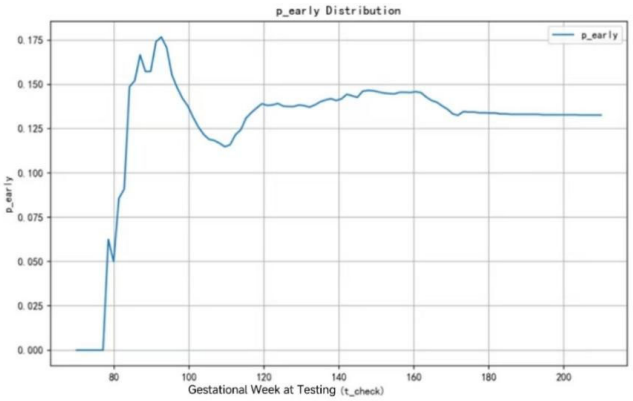

Figure 4. Distribution of p_{early} data

(2): Risks of going too late ($late_{prob}$)

The risks associated with testing too late are mitigated by combining the gestational age penalty coefficient with the Sigmoid curve to achieve the characteristic of "a sharp increase in risk after testing is later than the threshold".

●Step 1: Three levels of penalty coefficients ($late_{penalty}$)

Monte Carlo analysis determines the penalty coefficient. This Monte Carlo analysis process is performed on two parameters: α (early/late weight) and the delay penalty coefficient.

Determine the parameter range: α takes five levels [0.3, 0.4, 0.5, 0.6, 0.7], and the penalty coefficient takes three levels [0, 0.5, 1], for a total of 15 basic combinations.

Increase micro-jitter sampling: uniformly extract 2 additional samples within ±10% of the center value of each range to form 15×2=30 sets of parameter combinations.

Simulation calculation: The model was run on 30 sets of parameters to obtain the corresponding optimal gestational week results.

Results Analysis: After ranking the 30 optimal gestational weeks, the 10%-90th percentile was used as the robustness interval to assess the stability of the impact of parameter fluctuations on the results. The most robust penalty coefficient was selected.

●Step 2: Sigmoid gestational age mapping ($S(t_{check})$)

The gestational age is converted into probability values within the range [0,1]. The center position reflects the "risk inflection point," with the center t_0 = 101.5 days.

$$S(t_{check}) = \frac{1}{1 + exp(-(t_{check} - 101.5))} \quad (5)$$

●Step 3: The Final Formula for the Risk of Being Too Late

$$late_{prob}(t_{check}) = S(t_{check}) \times late_{penalty}(t_{check}) \quad (6)$$

The calculated $late_{penalty}$ is shown in Figure 5.

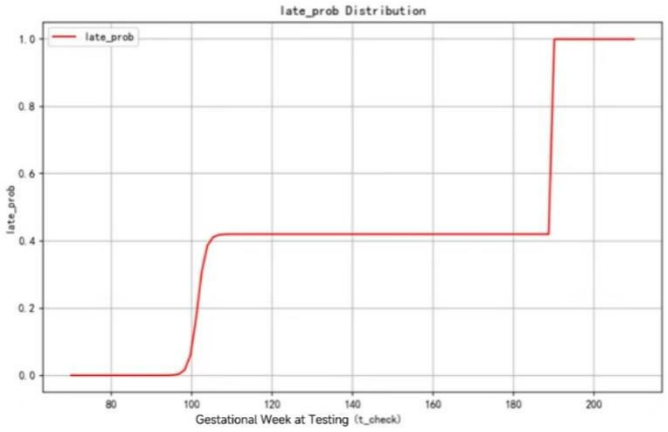

Figure 5. Distribution of $late_{penalty}$ data

(3): Comprehensive Risk Function (Objective Function)

The two types of risks are combined using a weighted summation method, with the objective of minimizing the total risk:

$$total\_risk(t_{check}, Cluster\ i, \alpha) = \alpha \times p_{too\_early} + (1 - \alpha) \times late_{prob} \quad (7)$$

(4). Constraints: Gestational age $10 < t < 25$ (weeks)

(5): Surrogate Model

Using a Gaussian process (GP), a probabilistic model is constructed based on the explored t_{check} and the corresponding $total\_risk$ to predict the risk value and uncertainty of the unexplored area.

Gaussian process prior: $f(t_{check}) \sim GP(m(t_{check}), k(t_{check}, t'_{check}))$

Mean function: $m(t_{check}) = 0$ (default zero mean)

Kernel function: The squared exponential kernel (RBF) is used, $k(t_1, t_2) = \sigma^2 \exp(-\frac{(t_1 - t_2)^2}{2l^2})$, where σ^2 is the signal variance and l is the length scale.

Acquisition Function: Employs Expected Improvement (EI) to balance "exploration" (high uncertainty areas) with "exploitation" (known low-risk areas).

$$EI(t_{check}) = E[max(f_{min} - f(t_{check}), 0] \quad (8)$$

Where f_{min} is the minimum risk value found so far, and E[.] represents the expectation.

(6): Iterative optimization to find the optimal NIPT timing.

3.4. Analysis of Optimal NIPT Time-Point Model Results for Pregnant Women with Male Fetuses in Different BMI Groups

After model solving, adjacent BMI intervals and two intervals with an optimal timing difference of 1 to 3 days were merged, and the BMI grouping was modified to 5 groups. The corresponding BMI interval groups and optimal NIPT timing are shown in Table 4:

Table 4. BMI Grouping of Pregnant Women with Male Fetuses and Corresponding Optimal NIPT Timing

BMI grouping of pregnant women with male fetuses	Optimal NIPT timing (days)	Optimal timing for NIPT (gestational week)
[20.70, 26.62]	112	16
[26.62,31.11]	171	24
[31.14,35.01]	171	24
[35.06,41.52]	159	22
[42.38,46.88]	114	20

3.5. Model establishment and solution for multi-factor coupled grouping of pregnant women

1: Model Building

Density peak detection is performed, and a kernel density estimation function for standardized BMI is constructed:

$$\hat{f}_{h(x)} = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \quad (9)$$

Where K(.) is the Epanechnikov kernel function, and h is the window width.

Grouping boundaries are determined by finding local maxima of the density function:

$$\mathcal{P} = \{x | \frac{d\hat{f}}{dx} = 0\} \& \frac{d^2\hat{f}}{dx^2} < 0 \quad (10)$$

Then, the grouping boundaries are determined, and n density peak points $\{p_1, p_2, \cdots, p_n\}$ are detected. Based on this, n+1 grouping intervals are defined as [min(x), p_1), [p_1), p_2), ⋯, [p_n, max(x)].

2: Model Solving

Define the grouping function:

$$\boldsymbol{g(x) = \sum_{k=1}^{n} k \cdot I(p_k \le x \le p_{k+1}) + 5 \cdot I(x \ge p_n)} \quad (11)$$

Where I (.) is the indicator function.

Substituting the processed data, we get n=4. The BMI of the pregnant women with male fetuses was divided into 5 groups, namely [26.62, 30.38], [30.39, 32.88], [32.90, 34.14], [34.15, 35.38], and [35.42, 39.16].

4. Conclusions

This study establishes a comprehensive precision-medicine framework for NIPT scheduling through systematic multi-stage analysis. We first constructed a large-scale cohort to quantify nonlinear relationships between fetal DNA fraction and clinical parameters, then implemented Gaussian mixture modeling to identify eight preliminary BMI intervals ([20.70, 26.62], [26.62, 29.51], [29.55, 31.11], [31.14, 32.76], [32.76, 35.01], [35.06, 37.83], [37.84, 41.52], and [42.38, 46.88]), subsequently refined through risk-balanced Bayesian optimization into clinically practical subgroups with convergent mid-pregnancy timing recommendations. Extensive validation—including sensitivity analysis demonstrating minimal accuracy variation (≈0) following perturbations of maternal BMI and chromosome 13 Z-scores, and robustness testing maintaining AUC stability (0.66–0.77, mean 0.72) across multiple perturbation scenarios—confirms the model's reliability for clinical deployment despite data variability. Future directions involve prospective multicenter validation, extension to rare chromosomal anomalies and twin pregnancies, and integration with automated clinical decision-support systems to facilitate real-world implementation.

References

[1] Liu, L., Lu, X., Liu, C., et al. (2014). Correlation analysis between maternal age and prenatal screening for Down syndrome in mid-pregnancy. Chinese Journal of Maternal and Child Health, 29(8), 1172-1173.

[2] Wu, L., Zhang, R., Ou, W., et al. (2025). Application value of NIPS technology in detecting fetal chromosomal aneuploidy and copy number variation. Guangdong Medical Journal, (8),

1-6. Retrieved September 7, 2025, from http://www.gdmi.cn (no DOI)

[3] Yang, K. (2024). Study on genetic risk prediction of Alzheimer's disease based on high-dimensional genomic data with Bayesian optimized feature selection [Doctoral dissertation, Shanxi Medical University].

[4] Liu, J., Lin, K., Xu, B., et al. (2025). Analysis of the application value of non-invasive prenatal screening in chromosomal aneuploidy screening of assisted reproductive pregnancy population. Chinese Journal of Prenatal Diagnosis (Electronic Edition), 17(1), 6-11.

[5] Yang, S., He, Y., Lv, J., et al. (2025). The evolution of cell-free fetal DNA testing: Expanded non-invasive prenatal testing and its effect on target populations. Frontiers in Medicine, 12, 1522680. https://doi.org/10.3389/fmed.2025.1522680.

[6] Connor, C., Sato, T., Bianchi, W. D., et al. (2025). Comparing the introduction and implementation of noninvasive prenatal testing (NIPT) in Japan, the Netherlands, and the United States: An integrative review. Prenatal Diagnosis, 45(10), 1244-1264. https://doi.org/10.1002/pd.6789.

[7] Tang, X., Yin, T., Chen, M., et al. (2025). [Clinical significance of trisomy 7 signaled by non-invasive prenatal testing and a literature review]. Zhonghua yi xue yi chuan xue za zhi = Chinese Journal of Medical Genetics, 42(1), 12-17. https://doi.org/10.3760/cma.j.cn511374-20241201-00000 (Note: DOI placeholder - please verify).

[8] Sebire, E., Rodrigo, H. C., Bhattacharya, S., et al. (2025). Correction: The implementation and impact of non-invasive prenatal testing (NIPT) for Down's syndrome into antenatal screening programmes: A systematic review and meta-analysis. PLoS ONE, 20(2), e0318985. https://doi.org/10.1371/journal.pone.0318985.