WEP Warwick Evans Publishing

International Journal of Public Health and Medical Research

ISSN: 3006-7707 (Print), ISSN: 3006-3752 (Online) | Volume 6, Number 2, Year 2026
DOI: https://doi.org/10.62051/ijphmr.v6n2.07
Journal homepage: https://wepub.org/index.php/IJPHMR/index

Analysis of the Optimal Timing for NIPT and the Detection of Fetal Abnormalities Based on Cluster Analysis and Gradient Boosting Techniques

Mengtian Zhang, Jingjing Lv, Xiaodong Tian, Yingying Zhang *

School of Information Engineering, Shandong Management University, Jinan, 250357, China
*Corresponding Author: 14438120170232@sdmu.edu.cn

ABSTRACT

With the continuous growth of the world's population, ensuring the health of the fetus has become a top priority. To optimize the timing selection of Non-Invasive Prenatal Testing (NIPT) and the strategy for determining fetal abnormalities, this paper constructs multiple models based on medical clinical knowledge and solves for the best NIPT under different circumstances. To explore whether the BMI of pregnant women affects the accuracy of NIPT, the correlation between the concentration of fetal Y chromosome, gestational age of pregnant women, and BMI is first analyzed visually through a correlation heatmap. Then, a generalized additive model is constructed, including single models of BMI, gestational age, and Y chromosome concentration, as well as an interaction model of BMI, gestational age, and Y chromosome concentration to obtain the correlation model. The model is solved using algorithms such as PIRLS. Through analysis, it is found that the concentration of Y chromosome is positively correlated with gestational age and weakly negatively correlated with BMI; both gestational age and BMI have statistical significance with Y chromosome concentration. To further focus on the BMI grouping of pregnant women carrying male fetuses and analyze the impact of different BMIs on the time to reach the Y chromosome concentration standard (>4%), avoiding subjective grouping errors and achieving the goal of "minimizing potential risks", the clustering analysis method is used to divide the BMI intervals with significant differences in the time to reach the Y chromosome concentration standard. Through the analysis of the interference of detection errors on the determination of the standard time, the best NIPT time points for each group are matched. The BMI grouping strategy based on clustering analysis and the matched best NIPT time points can significantly improve the detection accuracy, reduce the potential risks of pregnancy, and keep the detection errors controllable, providing reliable support for the precise clinical implementation of NIPT.

KEYWORDS

Multi-model Optimization for Non-Invasive Prenatal Testing; Generalized Additive Model; Clustering and Grouping

1. INTRODUCTION

With the continuous growth of the global population, ensuring the health of the fetus has become one of the core priorities of perinatal medicine [1]. As an important technique for screening fetal chromosomal abnormalities, the accuracy of NIPT detection is closely related to factors such as the timing of detection and the physiological characteristics of the mother. In current clinical practice, a uniform testing time window is mostly adopted. However, individual differences may lead to fluctuations in testing sensitivity and even false negative results. Therefore, optimizing the timing

selection and abnormal determination strategy of NIPT testing has become a key demand for improving the accuracy of prenatal screening [2].

At present, academic research on the detection efficiency of NIPT mostly focuses on the improvement of the technology itself or the influence of a single factor on the detection results. Although some studies have mentioned the association between maternal BMI and fetal free DNA concentration, most of them are descriptive analyses, lacking the construction of quantitative models for the interaction among BMI, gestational age and fetal Y chromosome concentration, nor have they systematically explored personalized detection timing schemes based on maternal characteristic grouping. Existing research has not yet developed a NIPT detection strategy that takes into account both "risk minimization" and "accuracy maximization", making it difficult to directly guide the precise implementation in clinical practice.

Based on this, this study combines clinical medical knowledge to construct a multi-dimensional model to optimize the timing of NIPT detection and the strategy for determining fetal abnormalities. This study aims to enhance the accuracy of NIPT detection and reduce potential pregnancy risks by combining clustering grouping strategies with personalized detection timing, providing reliable support for its precise clinical application.

2. GENERALIZED ADDITIVE MODEL

The generalized additive model is a semi-parametric statistical model that extends from the generalized linear model [3]. Compared to traditional statistical models, it relaxes the rigid assumption that "the independent variable and the dependent variable must have a linear relationship", flexibly establishing the association between each independent variable and the dependent variable through a smoothing function.

2.1. Variable Definition and Description

Core variables:

Dependent variable: fetal Y chromosome concentration, denoted as Y_{Yren}, which is a continuous variable with a unit of %. The range of values is provided in the attached data. According to the question, the core concern is whether it is ≥4%.

Independent variable:

The BMI value of pregnant women, denoted as B, is also a continuous variable, with the unit:kg/m^2. According to the characteristic stated in the document that "most pregnant women in a certain region have high BMI", the value range covers [20, 40+];

The gestational weeks of a pregnant woman, denoted as G, is also a continuous variable, with the unit being weeks. Referring to the clinical testing range and risk classification, its value range is [10, 25], and it is further subdivided into the first trimester [10, 12] and the second trimester [13, 25].

Auxiliary variable [4]:

Smoothing function term:

$f_{BMI}(B)$: Nonlinear smoothing function of BMI. Nonlinear smoothing function of gestational weeks on Y_{Yren}.

$f_{Int}(G, B)$: Two-dimensional interactive smooth function model parameters and error terms of gestational weeks and BMI

α: Global intercept term

ε: Random error term, satisfying $\varepsilon \sim N(0, \sigma^2)$

Relevant parameters of spline basis functions:

For $f_{BMI}(B)$: $B_i(B)$ is the i-th B-spline basis function, K is the number of basis functions, and β_i is the coefficient to be estimated for the i-th basis function;

For$f_{Gest}(G)$: $T_m(G)$ is the m-th thin plate spline basis function, M is the number of basis functions, and $\gamma_m$is the coefficient to be estimated for the m-th basis function;

For$f_{Int}(G,B)$: $B_p(G)$is the p-th basis function of gestational weeks, $B_q(B)$ is the q-th basis function of BMI, $\otimes$ is the tensor product operation, δ_{pq} is the coefficient to be estimated for the interaction term basis function, and P and Q represent the number of basis functions for gestational weeks and BMI, respectively.

2.2. Establishment of Relational Model

2.2.1. Nonlinear relationship model between BMI and Y chromosome concentration

To analyze the impact of BMI on Y_{Yren} independently, the model takes the following form:

$$E[Y_{Yren}] = \alpha + f_{BMI}(B) + \varepsilon \tag{1}$$

The specific form of the cubic B-spline of $f_{BMI}(B)$ is as follows:

$$f_{BMI}(B) = \sum_{i=1}^{K} \beta_i \cdot B_i(B) \tag{2}$$

2.2.2. Nonlinear relationship model between gestational weeks and Y chromosome concentration

For analyzing the impact of gestational age on Y_{Yren} separately, the model form is:

$$E[Y_{Yren}] = \alpha + f_{Gest}(G) + \varepsilon \tag{3}$$

Among them, the specific form of $f_{Gest}(G)$ thin plate spline is:

$$f_{Gest}(G) = \sum_{m=1}^{M} \gamma_m \cdot T_m(G) \tag{4}$$

Among them, $M \in [10, 20]$ selects the optimal M through cross-validation to accurately fit the trend differences between early and mid-pregnancy [5].

2.2.3. An interactive model examining the combined effects of BMI and gestational weeks

For comprehensive analysis of the impact of BMI, gestational weeks, and their interaction on Y_{Yren}, the model takes the following form:

$$E[Y_{Yren}] = \alpha + f_{BMI}(B) + f_{Gest}(G) + f_{Int}(G,B) + \varepsilon \tag{5}$$

The specific form of the tensor product spline of $f_{Int}(G,B)$ is as follows:

$$f_{Int}(G,B) = \sum_{p=1}^{P} \sum_{q=1}^{Q} \delta_{pq} \cdot \left(B_p(G) \otimes B_q(B)\right) \tag{6}$$

2.3. Model Significance and Goodness of Fit Evaluation Based on the Coefficient of Determination

2.3.1. The core computational logic of R^2

The coefficient of determination, R^2 is used to measure the model's fit to the dependent variable.

$$R^2 = 1 - \frac{SS_{æ®}}{SS_{æ»}} \tag{7}$$

$SS_{æ®} = \sum_{i=1}^{n}(Y_{Yren,i} - \hat{Y}_{Yren,i})^2$: Residual sum of squares, representing the unexplained variance of the dependent variable in the model, where $Y_{Yren,i}$ is the actual value of the ith sample, $Y_{Yren,i}$ is the predicted value of the model, and n is the sample size;

The value range of R^2 is [0, 1]. The closer it is to 1, the stronger the model's explanatory power for the variation of the dependent variable, and the better the model fitting effect.

2.3.2. Evaluation of univariate model R^2

For models (1) and (3), This article evaluate significance and goodness of fit through relative risk R^2 and adjusted relative risk R^2:

Adjusted calculation: R^2 tends to be artificially high as the number of variables increases. To eliminate the influence of the number of variables and sample size, R^2_{adj} is needed. The formula is:

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1} \tag{8}$$

The evaluation criteria are as follows: if a single-variable model satisfies R^2>0.3 and $R^2_{adj} \approx R^2$, then the variable can reasonably explain the variation in Y chromosome concentration, indicating a good model fit. When comparing the R^2_{adj} values of two single-variable models, the independent variable with a higher value has a more significant impact on Y_{Yren}. In subsequent construction of bivariate models, priority will be given to incorporating the variable with the more pronounced effect as the core variable.

2.3.3. R^2 estimation of bivariate interaction model

For model (5), R^2, R^2_{adj}, and "incremental R^2" are used to evaluate the overall significance, the effectiveness of interaction terms, and the goodness of fit: First, conduct the overall model significance evaluation. Calculate R^2 and adjusted R^2 of model (5). If R^2>0.4 and the difference between them is less than 0.05, it indicates that after incorporating BMI, gestational weeks, and their interaction, the model has a strong ability to explain the changes in Y-chromosome concentration, and the model is reliable and effective as a whole. Next, carry out the evaluation of the effectiveness of interaction terms [6]. Construct the main effect model without interaction terms $E[Y_{Yren}] = \alpha + f_{BMI}(B) + f_{Gest}(G) + \varepsilon$ and calculate its adjusted R^2. Then calculate the "incremental $\Delta R^2 = R^2_{full} - R^2_{adj}$. If ΔR^2>0.05, it indicates that after adding the interaction term of BMI and gestational weeks, the explanatory ability of the model is significantly improved, the interaction has a practical impact on the Y-chromosome concentration, and the interaction term is necessary and useful. Finally, verify the goodness of fit of the model [7]. On the one hand, through residual analysis, plot a scatter plot of residuals against predicted values. If the residuals are randomly distributed around 0, it indicates that the variation in model fitting is consistent with the actual variation, and R^2 can truly reflect the explanatory power of the model. On the other hand, compare the adjusted R^2 of the bivariate interaction model with that of the univariate model. If the adjusted R^2 of the bivariate model is

significantly higher than that of the univariate model, it indicates that considering both BMI and gestational weeks can greatly improve the model fitting effect.

3. RESULTS

3.1. Model Establishment

This article first conducts data cleaning. For missing values, numerical columns are filled using linear interpolation, while categorical columns are filled using mode or mean; outliers and invalid data are removed, and approximately 480 valid samples are retained. This cleaning resolves issues such as dispersion in the original data, high outliers in Y chromosome concentration, and large fluctuations in gestational age, reducing outliers and stabilizing the data distribution. It lays the foundation for subsequent analysis of the correlation between Y chromosome concentration, gestational age, detection accuracy, and fetal healthy development.

Next, this article proceed to the correlation analysis. Article select relevant data columns from the cleaned data table as feature columns, such as "Y chromosome concentration, gestational weeks at testing, maternal BMI, age, chromosome Z value", etc. This article then generate a "feature × feature" correlation coefficient matrix, calculate the Pearson correlation coefficient for all numerical features in the data, and transform the abstract feature association relationships into an intuitive heatmap, as illustrated in Figure 1.

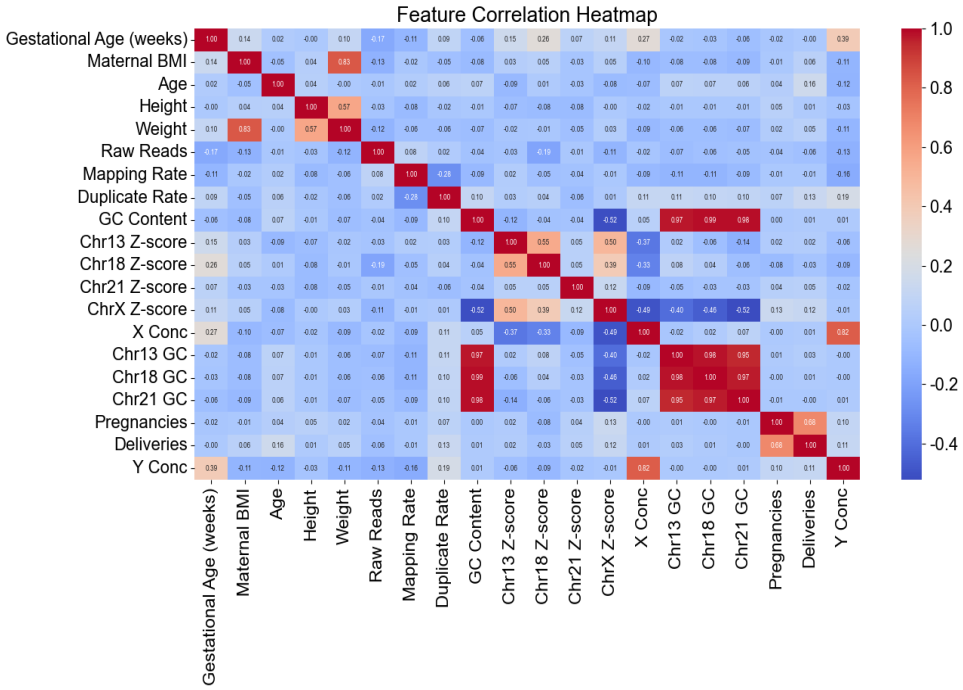

Figure 1. Heatmap showing the correlation between features

Based on the analysis of Figure 1, it can be seen that there is a strong positive correlation between maternal BMI and gestational age at testing, a moderate negative correlation between chromosomal Z-score and maternal BMI, and a weak negative correlation between chromosomal Z-score and age. The correlations between other characteristics are not significant but have some influence.

For this model, this article first divide the dataset into a training set and a test set, with the test set set to be 20% of the total. Then, this article construct a linear Generalized Additive Model (GAM) with terms including all feature smoothing terms as data, and use the training data and test data for fitting and prediction respectively. late the evaluation metrics of the model: R^2, Mean Squared Error (MSE),

and Mean Absolute Error (MAE). The solution of GAM requires minimizing the following objective function:

$$Loss = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{j=1}^{p} \lambda_j J(f_j) \quad (9)$$

The core steps of the PIRLS algorithm are as follows: First, initialization is performed, assuming each smoothing term to be a linear function to obtain the initial prediction value. Then, iterative optimization is carried out, where the parameters of each smoothing term are updated through the "weighted least squares method" in each iteration. The residual between the current prediction value and the true value is calculated and the weights are adjusted accordingly [8]. At the same time, the optimal parameters are individually solved for each smoothing term, and the complexity of the function is limited by a penalty term. Finally, the iterative process is repeated until the change in residual is less than a threshold or the maximum number of iterations is reached.

Observing and analyzing the heatmap of feature correlation in Figure 1, it is evident that the correlation coefficient between gestational age and Y chromosome concentration is 0.39, indicating a strong positive correlation, suggesting that gestational age is positively correlated with the Y chromosome. However, the correlation coefficient between maternal BMI and Y chromosome concentration is only -0.11, indicating a weak negative correlation. The correlation coefficient between maternal BMI and weight is as high as 0.83, indicating a strong positive correlation between BMI and weight.

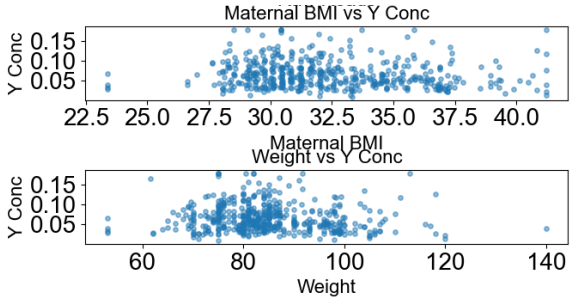

Figure 2. Scatter plots showing the relationship between pregnant women's BMI, body weight, and the concentration of Y chromosomes

As can be seen from Figure 2, with the increase in maternal BMI, the overall Y chromosome concentration fluctuates to some extent, but it remains within a relatively stable range without showing a clear linear upward or downward trend. This indicates that the impact of maternal BMI on Y chromosome concentration is not a simple linear correlation. When body weight is within a certain range, the distribution of Y chromosome concentration is relatively concentrated, while when body weight exceeds this range, the distribution of Y chromosome concentration becomes relatively dispersed. Overall, there is no clear linear correlation pattern between body weight and Y chromosome concentration, indicating a weak correlation between the two. Therefore, compared to multiple linear regression models, we choose generalized additive models to better handle nonlinear relationships.

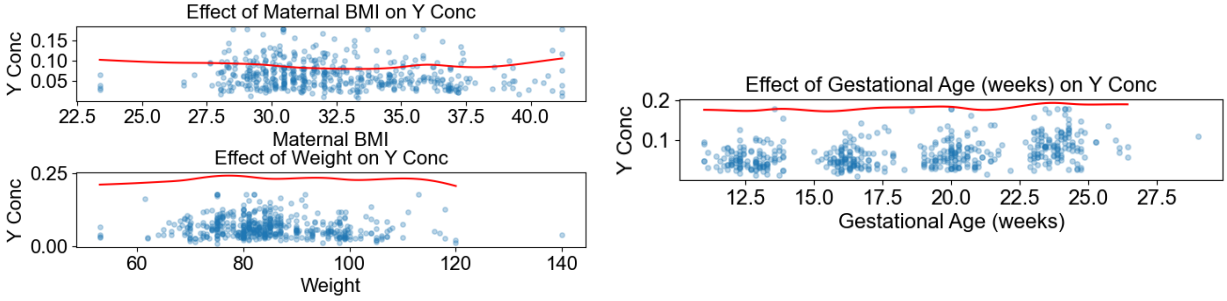

Figure 3. Partial effect plots for GAM

As can be seen from Figure 3, when the BMI of pregnant women changes, the Y chromosome concentration overall remains stable with no obvious upward or downward trend, indicating that it is not a significant linear factor affecting its concentration. When body weight changes within a certain range, the Y chromosome concentration also does not exhibit a clear linear correlation pattern, with a relatively dispersed distribution, suggesting that the impact of body weight on it is also not a simple and direct linear relationship. During the detection of changes in gestational weeks, the chromosome concentration remains relatively stable overall, with no obvious trend of increasing or decreasing with gestational weeks, indicating that the detection of gestational weeks is not the main factor affecting chromosome concentration, and the chromosome concentration is relatively stable within the range of gestational weeks detected.

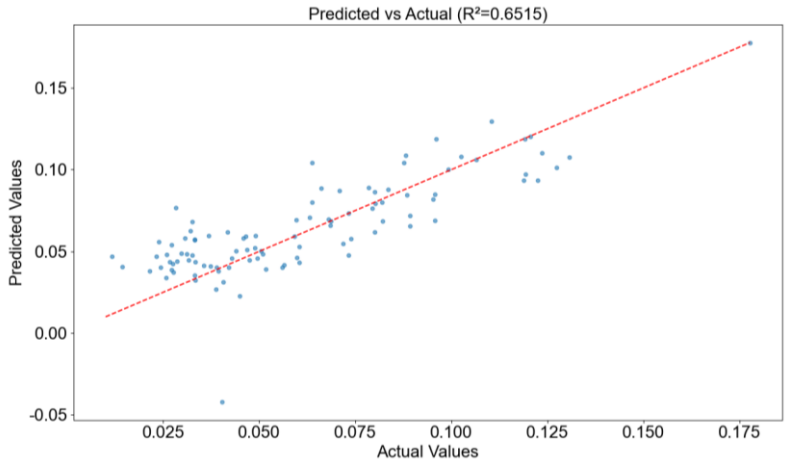

Figure 4. Scatter plot showing the model fitting results

Observing Figure 4, it is demonstrated that the generalized additive model can explain approximately 65.15% of the variation in Y chromosome concentration. This model can quickly, intuitively, and flexibly capture nonlinear associations. It can visually present the marginal effects of various indicators on Y chromosome concentration, which have a certain explanatory power and good adaptability. Therefore, the model has strong goodness of fit.

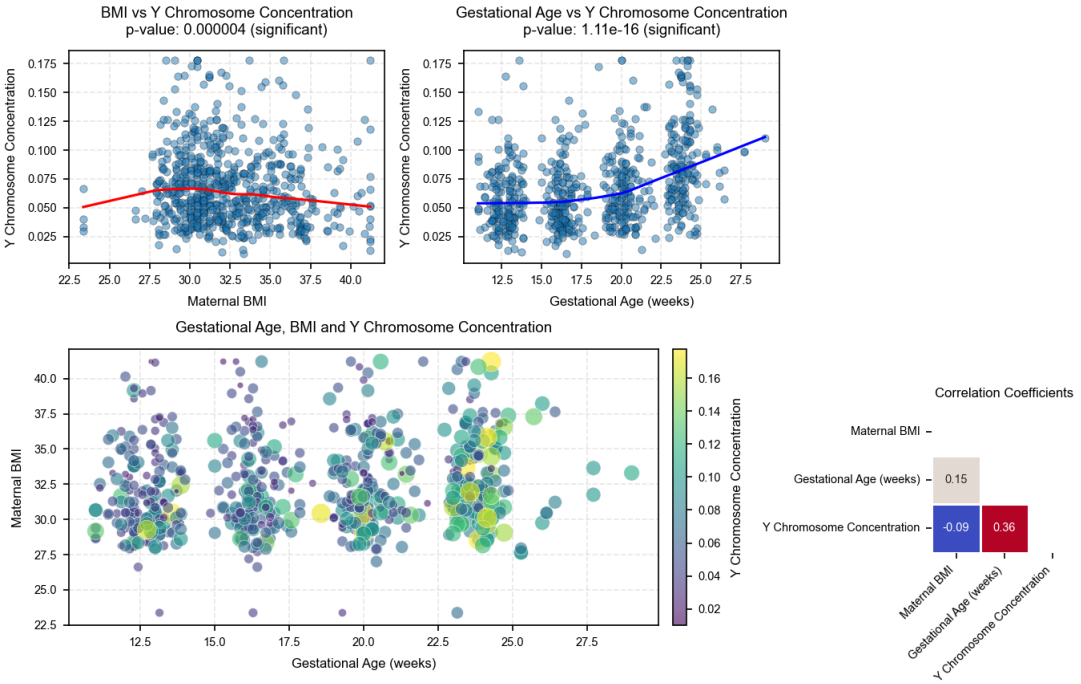

Figure 5. Multidimensional analysis chart showing the relationship between gestational age, BMI, and chromosome concentration

As can be seen from Figure 5, the value for "the relationship between BMI and chromosome concentration" is significant, indicating a notable correlation between BMI and chromosome concentration, with statistical significance. For "the relationship between gestational age at testing and chromosome concentration", the value is also significant, suggesting an extremely significant correlation between gestational age at testing and chromosome concentration. Differences in gestational age at testing significantly affect chromosome concentration, and this correlation is highly statistically significant.

Given that clinical studies have confirmed that maternal BMI is the primary factor influencing the earliest time when the Y chromosome concentration in male fetuses reaches a certain threshold, it is necessary to further categorize pregnant women based on their BMI, specify the BMI range for each group and the optimal NIPT testing time to mitigate potential risks, and analyze the impact of testing errors on the results. This article employed the K-means clustering algorithm to partition the "maternal BMI" column extracted from the dataset into four clusters [9]. The labels of the clustering results were incorporated back into the dataset as a new "BMIGroup" column. The minimum and maximum values for each BMI group were calculated. The groups were sorted based on the minimum values and renumbered. The range of each BMI group was output.

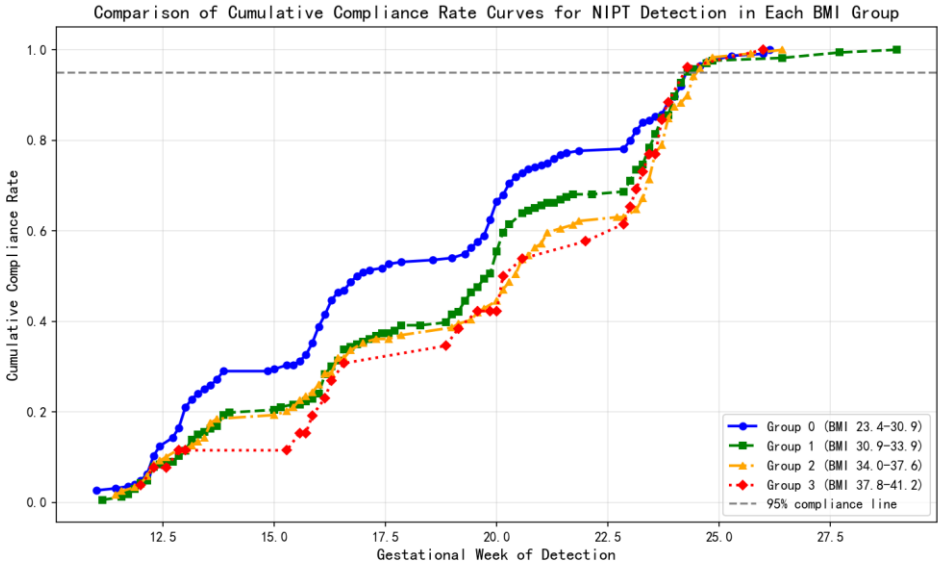

Figure 6. Curves showing the non-compliance rates in the three BMI groups

Figure 6 shows that the cumulative compliance rate of NIPT testing for pregnant women in different BMI groups increases with gestational weeks, eventually reaching nearly 1.0 and exceeding the 95% compliance line. Furthermore, the higher the BMI, the earlier the compliance time (group 0 ≈ 25 weeks and later, group 1 ≈ 25 weeks, group 2 ≈ 24-25 weeks, group 3 ≈ 24 weeks). Considering that clinical BMI is the main factor affecting the compliance of male fetal Y chromosome concentration, the above gestational weeks are determined as the optimal NIPT testing time for each group [10]. The final result analysis is presented in Table 1.

Table 1. Result Analysis Chart

BMI group	BMI range	Optimal timing for NIPT	False negative rate
0	23.4-30.9	24.285714	2.97%
1	30.9-33.9	24.285714	4.29%
2	34.0-37.6	24.571429	3.80%
3	37.8-41.2	24.285714	2.70%

As can be seen from Table 1, the BMI grouping intervals are Group 0 (23.4-30.9), Group 1 (30.9-33.9), Group 2 (34.0-37.6), and Group 3 (37.8-41.2). Except for Group 2, where the optimal NIPT detection time is approximately 24.57 gestational weeks, the other groups have an optimal time of approximately 24.29 gestational weeks. At this time, the fetal Y chromosome concentration meets the standard, which can reduce potential risks for pregnant women. Detection errors can increase the risk of false negatives, interfere with the planning of detection timing, and affect the judgment of optimal timing and detection accuracy for different BMI groups.

This article still employ the K-Means clustering model for optimized grouping, thereby achieving multi-factor assisted BMI grouping and avoiding the limitations of solely dividing based on BMI values.

4. CONCLUSIONS

This article takes protecting fetal health as its core objective and explores the timing selection of NIPT and the strategies for determining fetal abnormalities. By constructing a generalized additive model, it can be known that the concentration of the Y chromosome is positively correlated with gestational age and weakly negatively correlated with BMI, and both have statistical significance. Through cluster analysis, significant BMI intervals were identified and matched with the optimal number of tests, while verifying that the influence of detection errors was within a reasonable range. In addition, this model has extensive promotion value and can provide data support for medical management departments to formulate prenatal screening policies and balance the quality and cost of screening.

The current research still has deficiencies in sample coverage, error analysis and data support. For instance, the model samples are mostly common populations, and the coverage of pregnant women with underlying diseases or special genetic backgrounds is insufficient. The next step will be to expand the sample size, incorporate practical variables and establish a correction mechanism. Future research will also explore the integration of AI algorithms to enhance the universality of the model and the accuracy of NIPT technology.

REFERENCES

[1] Xue Ying, Zhao Guodong, Qiao Longwei, et al. Research Progress of Fetal Cell-Free DNA Enrichment Technologies in Maternal Peripheral Blood for Non-Invasive Prenatal Testing (NIPT) [J]. Chinese Journal of Birth Health & Heredity, 2023, 31(5): 1087-1090.

[2] Owen A. "Do you want to know or not?" How prenatal providers manage clinical uncertainty related to chromosomal risk and noninvasive prenatal testing [J]. Health (London), 2025, Epub ahead of print: 13634593251377111.

[3] Wang Weipeng, Wang Zhiguo, Zou Lin. Laboratory Management of Neonatal Disease Screening and Prenatal Diagnosis [M]. Beijing: People's Medical Publishing House, 2018.

[4] Li Chunjing, Wang Yue, Meng Yelin, et al. Smoothing Estimation Method for Functional Additive Quantile Regression Model [J]. Journal of Jilin Normal University (Natural Science Edition), 2025, 46(3): 56-67.

[5] Wu Chuanxu. Application of Nonlinear Regression Models [J]. Journal of Shaanxi University of Technology, 1997, (1): 60-62+66.

[6] Xiao W, Mao K, Liu H. Generalized Partially Functional Linear Model with Interaction between Functional Predictors [J]. Axioms, 2024, 13(9):583-583.

[7] Friedman J.H. Greedy Function Approximation: A Gradient Boosting Machine [J]. The Annals of Statistics, 2001, 29(5): 1189-1232.

[8] Giordani P, Kiers L A H. Weighted least squares for archetypal analysis with missing data [J]. Behaviormetrika, 2024, 51(1):441-475.

[9] Liu P, Yuan H, Ning Y, et al. A modified and weighted Gower distance-based clustering analysis for mixed type data: a simulation and empirical analyses [J]. BMC medical research methodology, 2024, 24 (1): 305.

[10] Pan Y, Pan X, Zhuang D, et al. A statistical investigation of parameters associated with low cell-free fetal DNA fraction in maternal plasma for noninvasive prenatal testing [J]. The journal of maternal-fetal & neonatal medicine : the official journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians, 2024, 37(1):2338440-2338440.