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.v6n2.03
Journal homepage: https://wepub.org/index.php/IJPHMR/index

Optimal BMI Grouping and Timing Selection Based on NIPT Test Data

Shuhan Wang *, Wenqi Sun, Chunyu Lin

School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, 100044, China

# These authors contributed equally.

*Corresponding Author: wangsh20232023@163.com

ABSTRACT

To further enhance the accuracy of non-invasive prenatal testing (NIPT) technology, this study analysed the correlation between NIPT results and factors such as gestational age and maternal body mass index (BMI), based on real-world data from 1,082 male fetuses and 605 female fetuses in China. First, Spearman's correlation coefficient analysis was conducted, revealing strong correlations between Y chromosome concentration and gestational age, BMI, and X chromosome concentration. Subsequently, K-means clustering revealed a distribution relationship between the optimal NIPT testing time point and different BMI intervals among pregnant women carrying male fetuses. Furthermore, a comprehensive BMI grouping model was developed by integrating multivariate clustering with optimization of the mean Y chromosome concentration. Constrained by a post-clustering Y-chromosome concentration mean exceeding 0.65% within each group and inter-group mean differences below 0.1%, the comprehensive BMI grouping results were: (20.70, 30.21], (30.21, 31.81], (31.81, 33.93], and (33.93, 46.88]. The corresponding optimal NIPT timepoints were 12.14, 13.28, 14.35, and 14.60 weeks respectively.

KEYWORDS

NIPT; K-Means Multivariate Clustering; Optimization Model; L-BFGS-B Algorithm

1. INTRODUCTION

With advancements in living standards and medical technology, an increasing number of expectant families seek screening for chromosomal disorders to fulfill their desire for healthy offspring. Non-invasive Prenatal Testing (NIPT), based on next-generation sequencing technology, analyzes cell-free DNA in maternal peripheral blood during pregnancy (including fetal DNA) to obtain genetic information about fetal chromosomes 21, 18, and 13 [1-2]. This enables screening for common chromosomal abnormalities such as Down syndrome, Edwards syndrome, and Patau syndrome. According to American College of Obstetricians and Gynecologists (ACOG), NIPT is now widely used in over 90 countries and regions globally, becoming one of the most extensively applied prenatal screening technologies. It benefits millions of families annually, including approximately one million pregnant women in China.

According to Jayashankar SS et al. [3], Non-Invasive Prenatal Testing (NIPT) was first discovered in 1988, it was primarily thought to be able to detect common aneuploidies, such as Patau syndrome (T13), Edward Syndrome (T18), and Down syndrome (T21). NIPT has shown promise as a simple and low-risk screening test, leading various governments and private organizations worldwide to support research studies aimed at standardizing its implementation. Zhou Gongyi et al. [4] evaluated

the clinical efficacy of NIPT by comparing its accuracy and safety with invasive prenatal diagnostic results. They concluded that NIPT is a non-invasive, safe, and highly effective screening method suitable for fetal chromosomal abnormality screening. Dai P [5] report a multiplex droplet digital PCR NIPT (dPCR-NIPT) assay that can detect trisomies 21, 18 and 13 in a single tube reaction with a better sensitivity and specificity and a much cheaper price than the NGS-NIPT.

In recent years, an increasing number of scholars have analyzed the accuracy, timeliness, and platform-specific characteristics of NIPT testing. Sebire E [6] aims to assess the extent of NIPT introduction into national screening programmes for DS worldwide, its uptake, and impact on pregnancy outcomes. Following NIPT implementation, the proportion of women choosing IPD after high chance biochemical screening decreased from 75% to 43%, an absolute risk reduction of 38%. Liehr T et al. [7] indicates that NIPT is done in China in later weeks of gestation, than in other countries. Besides, here for the first time it is highlighted that false positive NIPT results are less frequent, the earlier the screening is performed. Marton T's [8] research indicates all NIPT methods showed greater sensitivity for the detection of T21, above 97%, than traditional screening tests. For T18 detection, the targeted method with the microarray had a lower sensitivity compared to other tests.

It should be noted that the detection performance of NIPT is significantly influenced by the concentration of fetal cell-free DNA, which is closely related to individual factors such as the mother's height, weight, and age [9-10]. By employing mathematical methods to investigate the variation patterns of fetal cell-free DNA levels and further optimizing personalized optimal testing timing selection strategies, the accuracy and success rate of NIPT can be enhanced, thereby better safeguarding the health of both mothers and fetuses.

This study utilized real-world data from 1,082 male fetuses and 605 female fetuses in China. First, Spearman's correlation coefficient was applied for comprehensive data analysis. Subsequently, K-means clustering was employed to classify pregnant women into four BMI groups. Furthermore, considering factors such as detection errors and the proportion of Y-chromosome concentrations reaching the threshold, a comprehensive BMI grouping model was constructed by combining multivariate clustering with optimization of the mean Y-chromosome concentration. Finally, the L-BFGS-B optimization algorithm was applied to determine the optimal NIPT testing time point for each group.

2. METHODS

2.1. Data Source

The experimental data for this study originates from NIPT-related information concerning a subset of pregnant women in a specific region of China. This encompasses details such as maternal characteristics (height, weight, age, BMI, etc.), various test metrics (chromosomal concentration levels and Z-scores, for instance), routine testing records (including test timing and gestational age), and specialised indicators (such as chromosomal aneuploidy and gene read counts). The data source URL is https://cumcm.cnki.net/cumcm/studentHome/studentHome.

2.2. Correlation Analysis

Following preliminary data analysis, we observed that none of the indicators conformed to a normal distribution (see Table 1 in Section 3 for normality test results). The Spearman correlation coefficient does not require data to satisfy assumptions of linearity or normality. Therefore, we employed it to analyze the relationship between Y chromosome concentration and other indicators.

For a two-dimensional population (X, Y), sample data $x_1, x_2, \cdots, x_n$ and $y_1, y_2, \cdots, y_n$ can be obtained, with their rank statistics being $R_1, R_2, \cdots, R_n$ and $S_1, S_2, \cdots, S_n$ respectively. When X and Y are closely linked, these two sets of rank statistics are also closely related. The Spearman correlation coefficient is defined as the correlation coefficient q_{xy} between these two sets of rank statistics.

$$q_{xy} = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}}, \tag{1}$$

Based on the Spearman correlation coefficient test results (see Table 2 in Section 3), it can be concluded that there is no significant linear relationship between Y chromosome concentration and other relevant indicators.

Therefore, a multiple nonlinear regression method was employed to establish a regression equation, enabling a quantitative analysis of the trend changes among these variables.

$$\mathrm{Y} = \beta_0 + \beta_1 t + \beta_2 b + \beta_3 x + \beta_4 tb + \beta_5 tx + \beta_6 bx \tag{2}$$

Where Y denotes Y chromosome concentration, b represents maternal BMI, x indicates X chromosome concentration, and t signifies gestational age at testing.

Solve for the value of $\beta_0, \beta_1, ..., \beta_6$ using the method of least squares, then assess the model's significance via an F-test.

2.3. Multivariate Clustering Grouping

We employ a multivariate clustering approach to rationally group expectant mothers carrying male fetuses according to their BMI.

2.3.1. Bifactorial Clustering

Using the K-means clustering method to group the two variables of maternal BMI and Y chromosome concentration, dividing the samples into four categories.

For each sample point x_i, two features are available: BMI value b_i and Y chromosome concentration y_i, denoted as $x_i = (b_i, Y_i)$. The objective function for the K-means algorithm is then:

$$J = \sum_{i=1}^{k}\sum_{x \in C_i} \left\| x - \mu_i \right\|^2 = \sum_{i=1}^{k}\sum_{x \in C_i} [(b - \mu_{i,b})^2 + (y - \mu_{i,y})^2] \tag{3}$$

Where $\mu_i = (\mu_{i,b}, \mu_{i,y})$ denotes the centroid of the i-th cluster.

Set K to 4 and perform clustering. The steps are as follows:

Step 1) Initialise Cluster Centres

Using BMI and Y-chromosome concentration as features, randomly selecting K subjects as initial cluster centres.

Step 2) Allocation Procedure

For each sample x_i, assign it to the cluster containing the nearest centroid, with cluster label $z_i^{(t)}$:

$$z_i^{(t)} = \arg\min_{j\in\{1,2,\ldots,k\}} \left\| X_i - \mu_j^{(t)} \right\|^2 \tag{4}$$

The classification obtained at the t-th iteration is as follows:

$$C_j^{(t)} = \left\{ X_i : z_i^{(t)} = j \right\} \tag{5}$$

Step 3) Update Procedure

Based on the previous cluster partition, recalculate the new centroid for each cluster.

$$\mu_j^{(t+1)} = \frac{1}{\left| C_j^{(t)} \right|} \sum_{x\in C_j^{(t)}} X \tag{6}$$

Step 4) Convergence Check

Verify whether the algorithm has converged; if not, return to Step 2 to initiate the next iteration.

Step 5) Projecting Two-Dimensional Clustering Results onto the BMI Dimension

Calculate the range of each cluster within the BMI dimension. Then, identify the boundary points of adjacent clusters within the BMI dimension.

2.3.2. Comprehensive BMI Grouping

The time required for male fetuses to achieve adequate Y chromosome concentration is influenced by multiple factors (height, weight, age, etc.), whereas fetal cell-free DNA content exhibits strong correlation with maternal height and weight, yet weak association with maternal age [11]. Therefore, maternal age was excluded from clustering. Ultimately, ten indicators were selected: maternal height, weight, X chromosome concentration, GC content, read metrics (including raw read count, proportion mapped to the reference genome, proportion of duplicate reads, and number of uniquely mapped reads), BMI, and Y chromosome concentration. These were clustered into four categories, with the clustering results projected onto maternal BMI.

Considering the proportion of samples reaching the Y chromosome concentration threshold across each BMI group, two optimization directions are proposed: Firstly, a higher mean Y chromosome concentration indicates a greater number of samples within that cluster achieving the 4% threshold, which aids in determining the optimal NIPT timing. Second, it is essential to minimize the impact of variations in Y chromosome concentration across different BMI groups on determining the optimal timing for NIPT.

Consequently, we set the constraint that the mean Y-chromosome concentration must exceed 0.65%, with inter-group differences not exceeding 0.1%. The method is as follows:

Step 1) Formulation of the Objective Function

The objective function shall be derived from the standard K-means clustering objective.

$$\min J = \sum_{j=1}^{4} \sum_{x_i\in C_j} \left\| x_i - \mu_j \right\|^2 \tag{7}$$

Step 2) Constraints

$$\bar{y}_j \geq 0.0065 \tag{8}$$

$$max(\bar{y}_j) - min(\bar{y}_j) \le 0.001 \tag{9}$$

Where $\bar{y}_j$ denotes the mean Y-chromosome concentration for category j.

Step 3) Consolidate to obtain the optimised model

$$\begin{aligned} &\min J = \sum_{j=1}^{4} \sum_{x_i \in C_j} \left\| x_i - \mu_j \right\|^2 \\ &st. \begin{cases} max(\bar{y}_j) - min(\bar{y}_j) \le 0.001 \\ \bar{y}_j \ge 0.0065 \end{cases} \end{aligned} \tag{10}$$

Step 4) Employ the L-BFGS-B optimisation algorithm

Utilize L-BFGS-B algorithm package in Python to determine BMI groupings satisfying the mean Y-chromosome concentration constraint.

2.4. Optimal Timing Solution Model

For male fetuses, reliable detection results are only achievable when Y chromosome concentration exceeds 4%. Testing too early may yield inaccurate results due to insufficient Y chromosome concentration below 4%. Delayed testing risks missing the optimal treatment window, thereby increasing health risks. Clinical experience indicates that the optimal gestational period for detecting fetal sex chromosome concentration typically falls between 10 and 25 weeks.

Therefore, we have constructed a minimum risk function encompassing three dimensions: risk of insufficient Y chromosome concentration, risk of testing delay, and risk of concentration variation.

$$R(t) = \alpha R_{concentration}(t) + \beta R_{delay}(t) + \gamma R_{\text{variability}}(t) \tag{11}$$

In this context, t denotes the gestational age, α, β, γ represents the weighting coefficients for each risk dimension ($\alpha + \beta + \gamma = 1$). The following outlines three sub-functions:

(1) Risk of insufficient Y chromosome concentration

First, using a sigmoid function to approximate the risk of concentration deficiency.

$$R_{concentration}(t) \approx P(Y(t) < 4\%) \approx \frac{1}{1 + e^{K(\hat{Y}(t) - 4)}} \tag{12}$$

Where $\hat{Y}(t)$ denotes the predicted Y chromosome concentration at gestational week t, and K represents the slope parameter controlling the steepness of the risk function.

Subsequently, a random forest regression model is employed to forecast Y chromosome concentration:

$$\hat{Y} = f_{RF}(\text{Gestational age}, \text{BMI}) \tag{13}$$

(2) Risk of testing delay

$$R_{delay}(t) = \max(0, \frac{t - t_{ideal}}{t_{\max} - t_{ideal}}) \tag{14}$$

Where t_{ideal} denotes the ideal detection time, and $t_{\max}$ represents the maximum permissible detection time.

(3) Risk of concentration variation

$$R_{\text{var}iability}(t) = \min(1, c\frac{\sigma(Y(t))}{\mu(Y(t))}) \tag{15}$$

Where $\mu(Y(t))$ denotes the mean Y chromosome concentration around gestational week t, $\sigma(Y(t))$ represents the standard deviation of Y chromosome concentration around gestational week t, and c is the scaling factor.

(4) Weight Determination

The minimum risk function is obtained by summing the three subfunctions above with weights α, β, γ. And α, β, γ were determined by averaging the results obtained from three methods: the entropy weighting approach, principal component analysis, and direct weighting based on medical literature and clinical experience.

(5) Use L-BFGS-B to find the minimum-risk value

The optimal timing for NIPT t^* is determined by minimizing the total risk function:

$$t^* = \arg\min_{t\in[10,25]} R(t) \tag{16}$$

Employing L-BFGS-B package in Python to locate the t value corresponding to the minimum risk and the optimal NIPT timing for each BMI interval.

3. RESULTS

3.1. Preliminary Data Analysis

Following normality testing, the results are presented in Table 1. Most indicators did not conform to a normal distribution.

Table 1. Results of Normality Tests for Data

Indicator		Kolmogorov–Smirnov Test		Shapiro-Wilk Test		
BMI	Statistical Measure	Degree of Freedom	Significance	Statistical Measure	Degree of Freedom	Fignificance
Gestational age	0.037	1082	0.002	0.970	1082	0.000
Z-score of the X chromosome	0.019	1082	0.200*	0.998	1082	0.388
X chromosome concentration	0.049	1082	0.000	0.968	1082	0.000

3.2. Correlation Results

The Spearman correlation coefficient indicates that Y chromosome concentration exhibits a strong positive correlation with X chromosome concentration and gestational age, and a negative correlation with maternal BMI. Correlations with other indicators are relatively weak.

The multivariate nonlinear regression equation obtained between Y chromosome concentration, BMI, X chromosome concentration, and gestational age is as follows:

$$Y = -0.0539 + 0.0083t - 0.1416b + 0.4740x + 0.0571tb + 0.2461tx - 0.0157bx \quad (17)$$

Following the F-test, the results revealed overall significance with F=99.61 and p=0.000. Furthermore, the association between gestational age and BMI, as well as X chromosome concentration, was more pronounced, indicating that gestational age and BMI exert an interactive effect on Y chromosome concentration.

Table 2. F-test analysis for multiple non-linear regression

	Coefficient	t-statistic	P-value
b	-0.1416	-5.523	0.000
x	0.4740	18.879	0.000
$b•t$	0.0571	2.321	0.020
$x•t$	0.2461	10.511	0.000
Overall: F=99.61， p=0.000			

3.3. Optimal NIPT Timing

3.3.1. Bifactorial Clustering Results

When BMI clustering groups were obtained by considering only two factors, namely the BMI and Y chromosome concentration, four cluster intervals were identified through Python coding: (20.702, 30.474], (30.474, 32.321], (32.321, 35.267], and (35.267, 46.875]. The clustering results are shown in Figure 1.

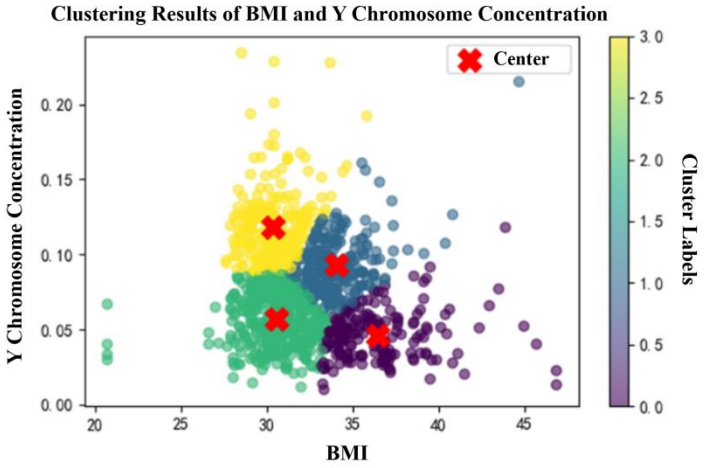

Figure 1. Bifactorial Clustering Results

Using the L-BFGS-B algorithm in Python to solve the minimum risk function, the optimal NIPT timing for each BMI range and the risk function values at that time are shown in Table 3.

(1) The optimal NIPT timing in Table 3 reveals that the optimal NIPT timing is delayed as BMI increases, consistent with the negative correlation between BMI and Y chromosome concentration.

(2) Risk values also increase with higher BMI. Pregnant women with higher BMI may experience slower increases in fetal Y chromosome concentration. To ensure accuracy of NIPT result, the NIPT timing must be delayed, leading to a higher risk of testing delay.

Table 3. Optimal NIPT Timing after Bifactorial Clustering

BMI Interval	Optimal timing (weeks)	Risk value
(20.702, 30.474]	11.51	0.1877
(30.474, 32.321]	12.33	0.2067
(32.321, 35.267]	13.20	0.2226
(35.267, 46.875]	13.83	0.2864

3.3.2. Comprehensive BMI Grouping Results

After comprehensively considering the impact of height, weight, age, detection error, and Y chromosome concentration compliance rate on BMI grouping, the optimized BMI grouping results are: [20.70, 30.21], [30.21, 31.81], [31.81, 33.93], and [33.93, 46.88]. The mean Y-chromosome concentrations for each groups are 0.0823, 0.0812, 0.0774, and 0.0685. The optimal NIPT timing and the risk function values at that time are shown in Table 4.

Compared to the previous bifactorial approach, the optimal timing for each component was delayed and the intervals widened after comprehensively considering multiple factors. Simultaneously, the risk values for Intervals 3 and 4 were significantly reduced compared to the bifactorial results. This indicates that the integrated grouping method effectively balances the risk of insufficient Y chromosome concentration and the risk of testing delay.

Table 4. Optimal NIPT Timing after Comprehensive BMI Grouping

BMI Interval	Optimal timing (weeks)	Risk value
[20.70, 30.21]	12.14	0.1886
[30.21, 31.81]	13.28	0.2082
[31.81, 33.93]	14.35	0.2178
[33.93, 46.88]	14.60	0.2464

3.3.3. Differences Between Bifactorial and Comprehensive Results

Regarding risk function values, compared to the bifactorial clustering, the risks associated with the optimal NIPT timing in intervals 1 and 2 remained nearly unchanged after comprehensive clustering, while the risk values for the optimal NIPT timing in intervals 3 and 4 decreased significantly. This indicates that grouping based on multiple factors can narrow the risk disparities among different groups.

This is because the comprehensive approach prioritizes optimizing the mean Y-chromosome concentration within each group, laying a solid foundation for subsequent timing determination:

First, the comprehensive results show that the mean Y-chromosome concentration is relatively high across all BMI groups (8.23%, 8.12%, 7.74%, and 6.85%, respectively), particularly in the first three intervals. This indicates a high proportion of Y chromosome concentrations meeting the standard in each group, enhancing the accuracy of subsequent NIPT timing determination. Additionally, the mean Y chromosome concentrations across the first three groups show minimal variation. This helps minimize errors in determining the optimal NIPT timing caused by differences in BMI intervals.

4. CONCLUSION

This study is based on real-world data from China's NITP testing. Methods including Spearman's correlation coefficient and K-means clustering were employed to investigate the relationship between Y chromosome concentration and factors such as maternal BMI and gestational age. A

comprehensive BMI grouping model was constructed by combining multivariate clustering with optimization of the mean Y chromosome concentration, followed by L-BFGS-B optimization.

A refined process for determining optimal testing time tailored to pregnant women with different characteristics was established, facilitating broader application and achieving more accurate chromosomal disease screening outcomes.

During modeling analysis, it was further observed that other factors such as geographic location, number of pregnancies, and testing platform may also influence optimal testing time. Therefore, subsequent efforts could involve collecting and constructing a high-quality NITP testing dataset incorporating these factors to enable more refined NITP testing recommendations for pregnant women.

REFERENCES

[1] Zhang Jun, Lo Y.M. Dennis. Clinical Applications of Plasma/Serum Cell-Free Nucleic Acids [J]. Chinese Journal of Clinical Laboratory Science, 2002, 20(z1): 13-17.

[2] JIANG Pei-Yong, Lo Y.M. Dennis. Properties of cell-free RNA and potential applications in noninvasive prenatal testing [J]. Chinese Bulletin of Life Sciences, 2018, 30(2): 127-133.

[3] Jayashankar SS, Nasaruddin ML, Hassan MF, et al. Non-Invasive Prenatal Testing (NIPT): Reliability, Challenges, and Future Directions [J]. Diagnostics, 2023, 13(15): 2570.

[4] Zhou Gongyi. Study on the Application Effect of Non-invasive Prenatal Testing (NIPT) in Screening for Fetal Chromosomal Diseases [J]. Chinese Science and Technology Journal Database (Abstract Edition) Medicine and Health, 2025(3): 060-063.

[5] Dai P, Yang Y, Zhao G, et al. A dPCR-NIPT assay for detections of trisomies 21, 18 and 13 in a single-tube reaction-could it replace serum biochemical tests as a primary maternal plasma screening tool? [J]. Translational Medicine, 2022, 20(1): 269.

[6] Sebire E, Rodrigo CH, Bhattacharya S, et al. The implementation and impact of non-invasive prenatal testing (NIPT) for Down's syndrome into antenatal screening programmes: A systematic review and meta-analysis [J]. PLoS One, 2024, 19(5): e0298643.

[7] Liehr T. Noninvasive prenatal testing (NIPT) results are less accurate the later applied during pregnancy [J]. Taiwan J Obstet Gynecol, 2024, 63(6):892-895.

[8] Marton T, Erdélyi ZR, Takai M, et al. Systematic Review of Accuracy Differences in NIPT Methods for Common Aneuploidy Screening [J]. Journal of Clinical Medicine, 2025, 14(8): 2813.

[9] Zhiheng Lan. Noninvasive Prenatal Testing for Aneuploidy Detection by Size Ratio-Based in Pregnancies [D]. South China University of Technology, 2020.

[10] Wu Lifang, Chen Chong, Xu Xueqin, et al. Study on the Variation Patterns of Fetal Free DNA Content during Pregnancy and after Delivery [C]// Proceedings of the 9th National Academic Conference on Genetic Disease Diagnosis and Prenatal Diagnosis & Symposium on New Technologies in Prenatal Diagnosis and Medical Genetics. 2014: 161-162.

[11] Rui Zhang et al. External Quality Assessment for Detection of Fetal Trisomy 21, 18, and 13 by Massively Parallel Sequencing in Clinical Laboratories [J]. The Journal of Molecular Diagnostics. 2016:244-252.