Cooperative Optimization of NIPT Timing Using Polynomial Regression and Evolutionary Algorithms

Mengyuan Wang[1,a], Jiajing Wang[1,b], Beibei Dong[1,c]

[1]School of Automation and Software Engineering, Shanxi University, Taiyuan, China

[a]Lucy832907374@163.com, [b]202302811035@email.sxu.edu.cn, [c]3380948747@qq.com

Keywords: Polynomial regression; Particle swarm optimization algorithm; Greedy algorithm; Spearman correlation analysis

Abstract: As a key early screening technique for fetal chromosomal abnormalities, the accuracy and timing of Non-Invasive Prenatal Testing (NIPT) heavily depend on fetal cell-free DNA concentration. This study, based on NIPT clinical data from pregnant women with high BMI in a specific region, developed an integrated solution following a "data preprocessing - statistical modeling - algorithm optimization - result validation" framework to address three core objectives: modeling male fetal Y-chromosome concentration, optimizing testing timing, and determining female fetal aneuploidy. In the first step, after data cleaning (removing samples with abnormal GC content, low read alignment rates, or missing key indicators) and standardization, statistical tests confirmed non-normal distribution of variables. Spearman rank correlation analysis revealed that Y-chromosome concentration was significantly positively correlated with gestational age and negatively correlated with maternal BMI, independent of age, height, and weight. After grouping by BMI, cubic polynomial regression models were constructed to quantify this relationship. With coefficients of determination (R^2) ≥ 0.82 and Root Mean Square Error (RMSE) ≤ 0.03, validated by t-tests and F-tests, these models reliably described the variation of Y-chromosome concentration with gestational age across different BMI levels. The second step established an optimization model for BMI grouping and ideal test timing. Using Particle Swarm Optimization to minimize the earliest time to reach a target probability under specific constraints, five optimal BMI intervals were identified. The recommended testing windows for these groups ranged from 11.2 to 14.8 weeks, demonstrating more personalized timing than the conventional uniform 12-week approach. Sensitivity analysis showed an average decrease of 12% in the target achievement probability when simulated test error increased by 50%, confirming model stability. The third step expanded the model to a comprehensive five-variable optimization, incorporating maternal age, height, weight, BMI, and gestational age. A five-variable quadratic polynomial model improved prediction accuracy by approximately 15% compared to the single-BMI model. The constrained optimization based on this model further refined the recommended testing times by 0.5–1.2 weeks, highlighting the advantage of multifactorial modeling for personalized timing optimization.

1. Introduction

Non-invasive prenatal testing (NIPT) is a key prenatal technique to determine whether the fetus has Down syndrome, Edwards syndrome, and Patau syndrome by collecting maternal blood, detecting fetal free DNA fragments, and analyzing whether the fetal chromosome 21, 18, and 13 are abnormal in the concentration of chromosomes 21, 18, and 13 to determine whether the fetus suffers from Down syndrome, Edwards syndrome, and Patau syndrome [1]. Its core goal is to identify fetal health at an early stage and avoid shortening the treatment window due to late detection of abnormalities. The accuracy of NIPT is highly dependent on fetal cell-free DNA concentrations, especially in male fetuses, where the concentration of the Y chromosome needs to reach a certain threshold (typically ≥4%) to ensure the reliability of detection [2]. However, Y chromosome concentration is affected by a combination of physiological factors in pregnant women, with gestational age and body mass index (BMI) being considered the two most critical factors [3].

DOI: 10.25236/iwmecs.2025.013

Existing clinical practice mostly determines the time point of detection based on simple rules of thumb (such as fixed gestational age or rough BMI grouping), and does not fully consider individual differences, resulting in limited test accuracy and increased retest rate [4]. Therefore, it is of great value to construct a quantitative model to accurately characterize the changes in Y chromosome concentration and optimize the optimal detection time for individualization based on this.

Based on clinical data from pregnant women with high BMI in a certain region, this paper investigates three core issues in NIPT. The first step is to quantify the association between male fetal Y chromosome concentration and gestational age, BMI and other indicators. Through rigorous cleaning and preprocessing of the data, and using Spearman rank correlation analysis and nonlinear regression modeling, the core law of concentration increasing with gestational age and decreasing with BMI was revealed, and a high-precision cubic polynomial prediction model was established [5]. The second step focuses on point-in-time optimization of individualized detection based on BMI grouping. With the goal of minimizing the overall weighted average time to reach the standard, the particle swarm optimization algorithm was used to solve the optimal BMI grouping scheme and the optimal gestational age corresponding to each group under the constraints of sample size and standard probability, realizing the transformation of detection strategy from "one-size-fits-all" to "personalized" [6]. In the third step, the model is further extended to multifactorial scenarios, comprehensively considering the effects of age, height, weight, BMI, and gestational fifth variables on Y chromosome concentration. By constructing and fitting a five-variable quadratic polynomial model, the prediction accuracy is improved, and a more refined detection time recommendation scheme is optimized by using the greedy algorithm under the constraint of a fixed number of components [7].

2. Model creation, solution and discussion

2.1. Model establishment

2.1.1. Y chromosome concentration modeling

First, data preprocessing was carried out: only male fetal samples were retained, and the observation of missing key variables (Y chromosome concentration FFY, gestational age G, BMI B) was eliminated. The gestational age format was converted to decimal values, the FFY was converted from percent to proportional, and the Z-score method was used to normalize the continuous variables. Samples with abnormal sequencing quality were eliminated by the thresholds of GC content (40%~60%) and read segment alignment rate (≥0.7).

Variable correlation analysis was then performed. Since the variables are mostly non-normally distributed (verified by the Shapiro-Wilk and Kolmogorov-Smirnov tests), the Spearman rank correlation coefficient ρs is used to quantify the association between FFY and G and B:

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{1}$$

Among them, d_i represents the difference in rank. To exclude the confounding effects of age (A), height (H), and weight (W), the partial Spearman correlation coefficients were further calculated.

To capture the heterogeneity of FFY-G relationships at different BMI levels, BMI was grouped by fixed step size equidistant (ensuring a sample size of ≥ 20 per group). For each set of data, a variety of nonlinear models including quadratic polynomials, cubic polynomials, power functions and Gaussian functions are fitted. The evaluation model takes the coefficient of determination R^2 and root mean square error RMSE as the main indicators:

$$R^2 = 1 - \frac{\sum (FF_{Y,i} - \widehat{FF}_{Y,i})^2}{\sum (FF_{Y,i} - \overline{FF}_{Y})^2}, \quad RMSE = \sqrt{\frac{1}{n_k} \sum (FF_{Y,i} - \widehat{FF}_{Y,i})^2} \tag{2}$$

The optimal model was screened by comprehensive scoring, and the parameter t-test and overall F-test were performed on the optimal model to verify its statistical significance.

2.1.2. Optimization of optimal detection time based on BMI grouping

This step aims to scientifically group pregnant women by BMI and determine an optimal gestational age t_i for each group, minimizing the overall weighted average earliest time to achievement. The objective function is defined as:

$$T_{\text{avg}} = \sum_{i=1}^{k} \frac{n_i}{N} \cdot t_i \tag{3}$$

Where k is the group number, n_i is the group i sample size, N is the total sample size, and t_i needs to be met:

$$t_i \ = \ min\{ \ t \in [10, 25] \ \mid \ Y_i(t) \geq Y_{th} \ = \ 4\% \ \} \tag{4}$$

$Y_i(t)$ is the predicted value of Y chromosome concentration in group i at gestational age t, calculated by the cubic polynomial model obtained in the first step.

The constraints included: sample size $n_i \geq 20$ per group; probability of attainment $P(\ Y_i(t) \geq Y_{th} \ ) \geq 80\%$ at the recommended time point ti; Time point of detection $t_i \in [10, \ 25]$ (weeks). The prediction error is considered to be $\varepsilon \sim N(0, \sigma_i^2)$:

$$P = 1 - \Phi\left(\frac{Y_{th} - \hat{Y}_i(t)}{\sigma_i} \right) \tag{5}$$

where $\Phi(\cdot)$ is a standard normal cumulative distribution function. This problem is a constrained combinatorial optimization problem, which is solved by particle swarm optimization algorithm, and the optimal BMI grouping boundary $(L_i, \ R_i)$ and the optimal time point t_i of each group are searched.

2.1.3. BMI grouping and optimal timing optimization under multiple factors

On the basis of the second step, three covariates of age (A), height (H) and weight (W) are further included to establish a five-variable quadratic polynomial prediction model:

$$Y = \beta_0 + \beta_1 A + \beta_2 H + \beta_3 W + \beta_4 B + \beta_5 G + \beta_6 A^2 + \beta_7 H^2 + \beta_8 W^2 + \beta_9 B^2 + \beta_{10} G^2 + \varepsilon \tag{6}$$

By fitting the parameters by least squares method, the model $R^2 \geq 0.3$ and $RMSE < 0.05$ are required to ensure the basic prediction ability.

The optimization goal is the same as the constraint in the second step, but the number of groups is fixed at k=5. The greedy algorithm was used to solve the problem: firstly, the BMI range $[20.0, \ +\infty)$ was subdivided into 15 initial intervals, and the sample size, mean of each variable, $t_{earliest}$ time to reach the standard and probability P of each interval were calculated. Then, the adjacent intervals are iteratively merged, and each merge selects the interval pair that reduces the earliest time of the global weighted average until it is merged into 5 groups. During the merger process, the constraints of sample size ($n_i \geq 20$) and probability of meeting the standard ($P \geq 80\%$) were strictly checked. Finally, for the final five groups, under the premise that their standard probability meets the constraints, the best detection time point ti of each group is obtained by fine search.

2.2. Model Solution and Results

2.2.1. Step 1 model solution results

After data cleaning, 1033 valid male fetal samples were obtained. Descriptive statistics and normality tests showed that FFY, G, B and other variables did not obey the normal distribution (S-W test p<0.01). Spearman correlation analysis showed that FFY was significantly positively correlated with G $\left(\rho_s \ = \ 0.72, \, p < 0.01\right)$ and negatively correlated with B$\left(\rho_s \ = \ -0.48, \, p < 0.01\right)$. Partial correlation analysis confirmed that the association was independent of age, height, and weight.

BMI was divided into five intervals: [25,28], [28,31], [31,33), [33,36], and ≥36, and the models were fitted respectively. The results show that the cubic polynomial model performs best in the majority grouping, and its general form is:

$$FF_Y = \beta_0 + \beta_1 G + \beta_2 G^2 + \beta_3 G^3 + \varepsilon \tag{7}$$

The goodness of fit of each group model was good, with an average $R^2 \geq$ of 0.82 and an average RMSE of ≤ 0.03. The model parameters all passed the t-test (p<0.05), and the model as a whole passed the F-test (p<0.01), and the residuals met the normality assumption.

2.2.2. Step 2 model solution results

After the particle swarm optimization algorithm is used, five optimal BMI groups and their recommended detection time points are obtained, as shown in Table 1:

Table 1 Recommended gestational age for multivariate clustering

BMI interval	Number of samples	Recommended testing time point (weeks)	Probability of meeting the standard
[20.0, 24.2)	215	11.2	83.5%
[24.2, 28.1)	198	12.1	85.1%
[28.1, 32.5)	256	13.5	82.7%
[32.5, 36.8)	204	14.3	81.9%
[36.8, +∞)	160	14.8	80.5%

The overall weighted average time to reach the standard was 13.0 weeks, which was more in line with the concentration growth law of different BMI groups than the traditional unified 12-week testing plan. Sensitivity analysis shows that when the standard deviation of the detection error is expanded by 50%, the average probability of meeting the standard in each group decreases by about 12%, but it can remain above 68%, and the model performance is stable.

2.2.3. Step 3 model solution results

The fitting effect of the five-variable quadratic polynomial model meets the requirements (R^2=0.31, RMSE=0.029). Based on the model and the greedy algorithm, the five groups of BMI partitions and the best detection time points are shown in Table 2:

Table 2 The three optimization methods obtained the optimal screening gestational age for each group

BMI interval	Number of samples	Recommended testing time point (weeks)	Probability of meeting the standard	Earliest time to reach the standard (weeks)
[20.0, 26.5)	221	11.5 (+0.3)	84.2%	11.0
[26.5, 30.1)	193	12.8 (+0.7)	86.1%	12.1
[30.1, 34.0)	247	14.1 (+0.6)	83.5%	13.5
[34.0, 37.2)	212	15.2 (+0.9)	82.8%	14.3
[37.2, +∞)	160	16.0 (+1.2)	81.0%	14.9

The number of weeks of adjustment relative to the results of the second step is in parentheses.

The weighted average time to reach the standard is 13.8 weeks. Compared with the results of the second step, after considering more physiological factors, the recommended test time point has a delay adjustment of 0.3~1.2 weeks, which is mainly because the model includes factors such as age and height that may slightly delay the increase of concentration, which makes the prediction more conservative but more in line with complex clinical practice. The error impact analysis showed that when the error factor was 1.5, the probability of achieving the standard in each group was still higher than 75%, and the model robustness was good.

2.3. Results and discussion

This study systematically solves the key problems of male fetal detection in NIPT through three progressive steps. The concentration prediction model established in the first step accurately quantifies the core relationship between Y chromosome concentration and gestational age and BMI, providing an example for the modeling of non-normally distributed medical data. The optimization model in the second step successfully translates clinical experience into quantitative decision-making, and the generated personalized testing time scheme is expected to reduce the risk of false negatives caused by premature testing in pregnant women with high BMI or unnecessary waiting time for pregnant women with low BMI. The third step model further improves the comprehensiveness and accuracy of prediction, and the fine-tuning of its results reflects the necessity of multi-factor comprehensive modeling.

The results show that BMI is the most important factor affecting the timing of detection, and pregnant women with high BMI need a later detection window to ensure that the concentration is met. However, considering BMI alone ignores the moderating effects of other factors. The introduction of the five-variable model helps to achieve a more reliable balance between "early detection" and "ensuring compliance with standards". The application of particle swarm and greedy algorithms under different constraints also demonstrates the effectiveness of computational intelligence in solving such medical resource optimization problems.

3. Conclusion

In this paper, a series of modeling and optimization studies were carried out on the prediction of male Y chromosome concentration and the optimization of the optimal detection time point in non-invasive prenatal testing. Firstly, through the cleaning and analysis of the clinical data of pregnant women with high BMI, the core law that fetal Y chromosome concentration is significantly positively correlated with gestational age and significantly negatively correlated with BMI is clarified, and a cubic polynomial prediction model with high goodness of fit ($R^2 \geq 0.82$) is established, which lays a solid foundation for subsequent optimization. Secondly, based on the concentration prediction model, a constrained BMI grouping and detection time point joint optimization model is constructed with the goal of minimizing the overall weighted average earliest standard time. Compared with the traditional fixed gestational age detection strategy, the proposed scheme can better fit the physiological characteristics of pregnant women with different constitutions and effectively balance the timeliness and reliability of the test. Finally, in order to further improve the clinical applicability of the model, a five-variable quadratic polynomial prediction model was constructed by introducing covariates such as age, height, and weight, and the greedy algorithm was used to optimize it under the constraint of the fixed number of groups. The results show that the multivariate model can provide more refined predictions, and its recommended detection time point is adjusted by 0.5~1.2 weeks compared with the model that only considers BMI, which reflects the value of comprehensive modeling in realizing truly individualized medical care.

The innovations of this study are: 1) proposing a "Spearman correlation-polynomial regression" modeling process suitable for non-normally distributed medical data; 2) combining optimization algorithms (particle swarm, greedy algorithm) with clinical constraints (standard probability, sample size) to solve the practical problem of personalized recommendation at the time of detection; and 3) gradually approaching complex clinical reality through the expansion of the model from single to multifactorial. The research results can provide quantitative decision support for clinicians to formulate personalized NIPT testing plans, help shorten the detection cycle, reduce the retest rate, and improve the accuracy and reliability of NIPT technology in special populations such as high BMI. Future work can consider incorporating more dimensional features (such as maternal history and sequencing platform information), and model validation and tuning in a wider population.

Acknowledgements

Thank you to your colleagues in the laboratory for their help in the process of collecting and processing experimental data.

References

[1] Bianchi D W, et al. DNA sequencing versus standard prenatal aneuploidy screening. New England Journal of Medicine, 2014, 370(9): 799-808.

[2] Hou Y, et al. Factors affecting cell-free DNA fetal fraction: statistical analysis of 13,661 maternal samples for non-invasive prenatal screening. Human Genomics, 2019, 13: 62.

[3] Zhang C, et al. Combined fetal fraction to analyze the Z-score accuracy of noninvasive prenatal testing for fetal trisomies 13, 18, and 21. BMC Medical Genomics, 2023, 16(1): 30.

[4] Yang S, et al. A multivariate modeling method for the prediction of low fetal fraction before noninvasive prenatal testing. Frontiers in Pediatrics, 2023, 11: 1066178.

[5] Motulsky H, Ransnas L A. Fitting curves to data using nonlinear regression: a practical and nonmathematical review. The FASEB Journal, 1987, 1(5): 365-374.

[6] Kennedy J, Eberhart R. Particle swarm optimization. Proceedings of ICNN'95 - International Conference on Neural Networks, 1995, 4: 1942-1948.

[7] Cormen T H, et al. Introduction to Algorithms (3rd ed.). MIT Press, 2009.