Research on NIPT Time Point Selection and Fetal Abnormality Determination Based on Multiple Nonlinear Regression Model

Jie Qian[1,a], Xuanyu Lu[1,b]

[1]School of Mathematical Science and Applied Mathematics, Nanjing Normal University Taizhou College, Jiangsu, China

[a]2804380577@qq.com, [b]2962295968@qq.com

Keywords: NIPT; K-Means Clustering; Greedy Algorithm; Multiple Linear Regression

Abstract: With the rapid development of medical technology, NIPT, a non-invasive prenatal testing technology, has become one of the important tools for early detection and determination of fetal health. Based on the K-means clustering and greedy algorithm, this study focuses on the selection of the optimal NIPT time point for different pregnant women. Firstly, the three-dimensional clustering of Y chromosome concentration, gestational age and BMI was grouped by K-means clustering, and a cubic polynomial relationship model between Y chromosome concentration and gestational age and BMI was established. Then, for the problem of BMI grouping of male fetuses and pregnant women, a single-objective optimization model was established and the greedy algorithm was used to solve the best detection gestational age of each BMI grouping, and the results showed that different BMI intervals corresponded to different optimal detection times, and the detection error had an impact of 6.09% on the results. Finally, considering multiple factors such as height, weight, and age, a multiple nonlinear regression model was established and the detection time point was optimized, and the detection error influence was 16.16%. The results show that the optimization model using clustering and greedy algorithms can effectively determine the time point of personalized NIPT detection, providing a theoretical basis for clinical practice.

1. Introduction

The development of non-invasive prenatal testing (NIPT) represents a milestone advancement in the field of prenatal medicine. Leveraging the maturity of high-throughput sequencing technologies and the refinement of bioinformatics analysis methods, NIPT enables early, non-invasive, and high-accuracy screening for fetal chromosomal abnormalities by detecting the minute fraction of cell-free fetal DNA (cffDNA) in maternal plasma, a foundational discovery made by [1]. Since its gradual introduction into clinical practice around 2011, this technology has been rapidly adopted worldwide [2].

A particular challenge exists for pregnant women with high Body Mass Index (BMI). Their blood often contains a lower fraction of cffDNA [3], which can lead to delayed testing windows or even test failure. This not only increases the medical burden of repeat sampling but may also delay the identification of and intervention for fetal abnormalities, as a low fetal fraction is a known factor impacting test performance [4]. Current international clinical practice guidelines, such as those from [5], already recommend considering the cffDNA fraction threshold in relation to factors such as gestational age and BMI. However, there is still a lack of refined, stratified models for optimal testing timing based on large-sample data [6]. Furthermore, research systematically exploring how to balance test accuracy with clinical risks under the interaction of multiple factors remains scarce.

Against this backdrop, this paper uses the challenge of achieving a sufficient fetal Y-chromosome concentration as a starting point. The significance of this extends far beyond the detection of male fetuses—it addresses a core challenge of NIPT in the era of personalized medicine: how to optimize testing strategies based on population characteristics [7]. From a broader perspective, establishing a quantitative relationship between fetal Y-chromosome concentration, gestational age, and BMI is not merely for determining the optimal timing for male fetus detection;

DOI: 10.25236/iwmecs.2025.011

the underlying methodology can also be extended to the screening process for other autosomal aneuploidies.

This study will first analyze the correlation between fetal Y-chromosome concentration and maternal indicators such as gestational age and BMI, establishing a suitable mathematical model to describe this relationship. Subsequently, pregnant women carrying male fetuses will be grouped according to BMI to determine the BMI ranges and optimal testing time for each group. Finally, considering multiple factors such as height, weight, and age, along with measurement errors and target achievement rates, the BMI grouping for women with male fetuses will be re-evaluated to determine the optimal testing time for each new group, thereby minimizing clinical risks. Through this three-step research, we aim to provide a systematic solution for the personalized timing of NIPT.

2. Model creation, solution and discussion

2.1. The relationship model of Y chromosome concentration was established

The relationship between fetal Y chromosome concentration and gestational age, BMI and other indicators was studied. Firstly, the data were preprocessed to eliminate the logical abnormalities and BMI extreme values of gestational age. K-means clustering was used to group the Y chromosome concentration, gestational age and BMI in three-dimensional clustering, and the data were divided into four stages of pregnancy.

Establish a multiple linear regression model:

$$YC = a \cdot GW + b \cdot BMI + c \tag{1}$$

where YC represents the concentration of Y chromosomes, GW represents the gestational age of detection, and a, b, and c are the regression coefficients. Due to the poor fitting effect of the linear regression model, a cubic polynomial nonlinear regression model was further established:

$$YC = a \cdot BMI^3 + b \cdot BMI^2 \cdot GW + c \cdot BMI \cdot GW + dBMI^2 + e \cdot GW + f \cdot BMI + g \tag{2}$$

2.2. A single-objective optimization model based on BMI grouping was established

A single-objective optimization model was established to determine the BMI grouping of male pregnant women and the determination of the optimal time point for NIPT. Objective function:

$$\min Q = -\left|\sum_{t=1}^{n} z(YC_{(t)1}) - \sum_{t=1}^{n} z(GW_{(t)1})\right| \tag{3}$$

Where $YC_{(i)j}$ represents the Y chromosome concentration of the j pregnant women in group i, $GW_{(i)j}$ represents the gestational age of the jth pregnant women in group i, z(·) represents the standardized treatment. The constraints include: $10 \leq GW_{(i)j} \leq 25$, i = 1, 2,...,7 ;$0.4 \leq GC(j) \leq 0.6$;

2.3. Optimization model under the influence of multiple factors

On the basis of the second step, the influence of height, weight, age and other factors was further considered, and a multiple linear regression model was established: $YC_i = a \cdot AG + b \cdot HT + c \cdot WT + d \cdot GW + e \cdot BMI - f$ .where AG represents age, HT represents height, and WT represents weight. The objective function of the optimization model is the same as the second step, but the relational model in the constraint is changed to the above multiple linear regression equation.

2.4. Model solving and discussion

As shown in Table 1, K-means cluster analysis identified four distinct stages of pregnancy based on gestational age, BMI, and Y-chromosome concentration. The cluster centers reveal that late gestational age and post-pregnancy stages are associated with higher average gestational weeks

(21.568 and 22.213, respectively), while BMI and Y-chromosome concentration vary across stages, with post-pregnancy showing the highest average BMI (36.930) and late gestational age exhibiting the highest average Y-chromosome concentration (0.082).

By K-means cluster analysis, the clustering centers of the four stages of pregnancy are as follows:

Table 1 Clustering centers in pregnancy

Category	Detection of gestational age	BMI Y chromosome	concentration of pregnant women
Pregnancies	14.072	30.213	0.078
Pregnant	14.618	34.327	0.073
Late gestational age	21.568	31.052	0.082
Post-pregnancy	22.213	36.930	0.070

Table 2 presents the fitting performance of a nonlinear regression model applied to each of these pregnancy stages. The model demonstrates varying explanatory power across stages, with the pregestational age group having the highest R^2 (0.1595) and the lowest RMSE (0.0268), indicating the best fit. In contrast, the pregnancy stage shows the poorest fit, with the lowest R^2 (0.0426) and a relatively higher RMSE (0.0314). These results suggest that the relationship between the variables is more effectively captured by the nonlinear model in certain pregnancy stages, particularly in pregestational and late gestational age periods.

The fitting effect of the nonlinear regression model is as follows:

Table 2 Fitting effect of nonlinear regression model

Cluster group	R^2	RMSE
Pregestational age	0.1595	0.0268
Pregnancy	0.0426	0.0314
Late gestational age	0.1251	0.0355
Post-pregnancy	0.1411	0.0368

2.5. Single-objective optimization model based on BMI grouping

Based on the cluster analysis conducted earlier, Table 3 further details the BMI-specific groupings and their corresponding optimal time points for NIPT detection. The table categorizes pregnant women into seven distinct groups (Group 1 through Group 7) based on precise BMI intervals, each associated with a recommended detection window ranging from 12.4 to 22.5 weeks of gestation.

The single-objective optimization model was solved by the greedy algorithm, and the best detected gestational age for each BMI group was obtained as follows:

Table 3 BMI grouping and optimal time point for detection

Cluster group	BMI group interval	Best NIPT detection time point (weeks)
Group 3	[26.62,29.55)	12.6
Group 4	[29.55,31.07)	17.5
Group 2	[31.07,32.65)	16.9
Group 6	[32.65,34.38)	18.3
Group 1	[34.38,36.51)	12.4
Group 7	[36.51,40.14)	22.5
Group 5	[40.14,46.88)	16.0

According to the Figure 1, the impact of detection error on the results showed that the overall impact was 6.09%.

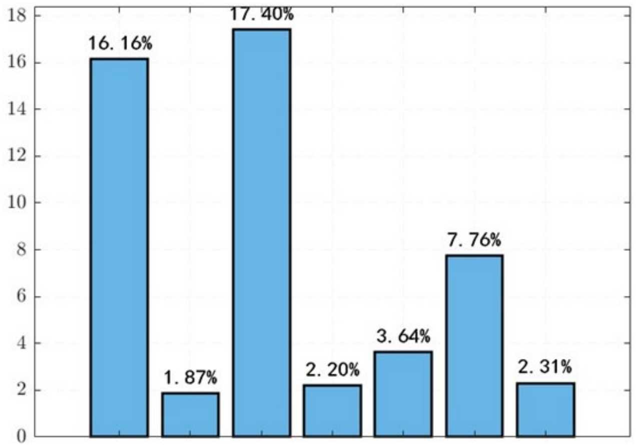

Figure 1 Error impact analysis diagram

2.6. Optimization model solution under the influence of multiple factors

After considering multiple factors, the multiple linear regression model is re-established, and the fitting effect is as follows:

Table 4 Fitting effect of multiple linear regression model

category	R^2	standard regression	equation significance
Group 3	0.175	0.029	0.001
Group 4	0.127	0.028	0.001
Group 2	0.081	0.036	0.013
Group 6	0.112	0.030	0.001
Group 1	0.602	0.039	0.045
Group 7	0.080	0.033	0.001
Group 5	0.217	0.022	0.015

The optimal time points for each BMI group after optimization are shown in Table 5:

Table 5 Optimal detection time after multi-factor optimization

Cluster group	BMI group interval	Best NIPT detection time point (weeks)
Group 3	[26.62,29.55)	20.1
Group 4	[29.55,31.07)	16.9
Group 2	[31.07,32.65)	15.6
Group 6	[32.65,34.38)	19.1
Group 1	[34.38,36.51)	10.6
Group 7	[36.51,40.14)	20.7
Group 5	[40.14,46.88)	15.4

The impact of detection error on the results showed that the overall impact was 16.16%.

In this study, an optimization system for NIPT detection time selection was established through three models. As shown in Table 4, the Y chromosome concentration relationship model provides a basis for subsequent optimization. The single-objective optimization model based on BMI grouping showed that different BMI groups needed different detection times, and the low BMI group needed to be detected earlier, while the high BMI group could be appropriately delayed. As shown in Table 4, the optimization model under the influence of multiple factors showed that after considering more factors, the optimal detection time distribution showed new characteristics, and the high BMI group needed to be detected earlier, while the moderate BMI group could be delayed to about 20 weeks. From the perspective of error analysis, the influence of detection error in single-factor optimization on the moderate BMI interval is more obvious, while the detection error in multi-factor optimization has a more significant impact on the low BMI and high BMI intervals. This indicates that under the influence of multiple factors, the distribution pattern of detection error has changed, and a more personalized point in time selection strategy is required. The K-means clustering algorithm used in this study can effectively group pregnant women, and the greedy algorithm can quickly solve the optimal detection time.

3. Conclusion

Based on the K-means clustering and greedy algorithm, a systematic optimization model is established for the time point selection problem of NIPT detection. Firstly, the quantitative relationship between Y chromosome concentration and gestational age and BMI was determined by K-means cluster analysis, and it was found that Y chromosome concentration was significantly positively correlated with gestational age and significantly negatively correlated with BMI, and a cubic polynomial relationship model was established. Then, for the problem of BMI grouping of male fetuses and pregnant women, a single-objective optimization model was established, and the greedy algorithm was used to solve the optimal detection age of each BMI group, and the results showed that different BMI intervals corresponded to different optimal detection times, and the detection error had an impact of 6.09% on the results. Finally, considering multiple factors such as height, weight, and age, a multiple linear regression model was established and the detection time point was optimized, and the detection error was 16.16%. The main contributions of this study are: first, a grouping method of pregnant women based on cluster analysis is proposed, which can be reasonably grouped according to the characteristics of pregnant women; second, a time-of-time optimization model considering multiple factors is established, which can balance detection accuracy and time risk; Third, the greedy algorithm is used to effectively solve the optimal detection time, providing an actionable solution for clinical practice. The limitation of this research is that the linear programming model requires both objective functions and constraints to be linear, while most of the practical problems are nonlinear relationships, which require more complex optimization techniques. In addition, the K-means clustering algorithm is only suitable for cases where the clustering mean is meaningful, and the effect is poor for datasets containing symbolic attributes. Future research can be further expanded to consider more influencing factors, such as the genetic background of pregnant women, environmental factors, etc., and more complex optimization algorithms, such as genetic algorithms, particle swarm algorithms, etc., can be explored to further improve the accuracy and practicability of the model. The results of this study can provide a theoretical basis for clinicians to formulate personalized NIPT testing plans, which can help improve the success rate of testing and reduce the risk of pregnant women.

Acknowledgements

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

References

[1] Lo, Y. M., Corbetta, N., Chamberlain, P. F., Rai, V., Sargent, I. L., Redman, C. W., & Wainscoat, J. S. (1997). Presence of fetal DNA in maternal plasma and serum. *The Lancet*, 350(9076), 485-487.

[2] Bianchi, D. W., & Chiu, R. W. (2018). Sequencing of circulating cell-free DNA during pregnancy. *New England Journal of Medicine*, 379(5), 464-473.

[3] Ashoor, G., Syngelaki, A., Poon, L. C., Rezende, J. C., & Nicolaides, K. H. (2013). Fetal fraction in maternal plasma cell-free DNA at 11–13 weeks' gestation: relation to maternal and fetal characteristics. *Ultrasound in Obstetrics & Gynecology*, 41(1), 26-32.

[4] Canick, J. A., Palomaki, G. E., Kloza, E. M., Lambert-Messerlian, G. M., & Haddow, J. E. (2013). The impact of maternal plasma DNA fetal fraction on next generation sequencing tests for common fetal aneuploidies. *Prenatal Diagnosis*, 33(7), 667-674.

[5] Gregg, A. R., Skotko, B. G., Benkendorf, J. L., Monaghan, K. G., Bajaj, K., Best, R. G., ... & Watson, M. S. (2021). Noninvasive prenatal screening for fetal aneuploidy, 2021 update: a position statement of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine*, 23(10), 1797-1814.

[6] Larsen, S. O., Hatt, L., Brasen, C. L., Hauge, S., Nørgaard, P., Hviid, K. V. F., ... & Vogel, I. (2021). Prediction of fetal fraction in non-invasive prenatal testing based on maternal and gestational characteristics. *Prenatal Diagnosis*, 41(10), 1279-1287.

[7] Hui, L., & Bianchi, D. W. (2017). Noninvasive prenatal DNA testing: the journey from technology to clinical care. *Annual Review of Medicine*, 68, 229-236.