Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 19, No. 3, 2025

NIPT Testing Timing Decision Algorithm Based on K-means Clustering and Multi-Objective Risk Optimization

Yufei Li, Sitong Liu, and Haiyuan Liu

College of Big Data and Computer Science, Shanxi Institute of Science and Technology, Shanxi, China

Abstract: The accuracy of non-invasive prenatal testing (NIPT) is highly dependent on fetal cell-free DNA concentrations, especially the 4% threshold for male Y chromosome concentrations. To optimize the timing of detection and minimize clinical risks, this paper proposes a data-driven decision-making framework. Firstly, the relationship between Y chromosome concentration and 21 characteristics such as gestational age and BMI was explored by Spearman rank correlation analysis and random forest regression model, and it was found that X chromosome concentration had the most significant effect. Secondly, a K-means clustering method based on standardized BMI was proposed for male fetal samples, and a comprehensive objective function combining early detection risk, failure risk, error risk and delay punishment was constructed to determine the optimal gestational age detection in each group. The results showed that the recommended testing time point for the low-to-medium BMI group (20.7-35.9) was about 14.9 weeks, while the high BMI group (36.3-45.7) needed to be delayed to 18.2 weeks, and the overall robust success rate was 0.849. Finally, the population was subdivided into four groups, with the optimal time point distributed from 14.1 to 18.5 weeks, and the overall robust success rate reached 0.863, of which the "standard compliance ratio" was identified as the most critical feature. This study provides a systematic algorithm solution for personalized NIPT detection time point planning.

Keywords: Non-invasive prenatal testing; Point-in-time optimization; K-means clustering; Risk function; Random forest.

1. Introduction

With the implementation of Chinas "three-child policy" and the intensification of population aging, ensuring fetal health and achieving eugenics have become important public health issues. Non-invasive prenatal testing (NIPT), as a new technology for screening fetal chromosomal aneuploidy abnormalities by analyzing fetal cell-free DNA in the peripheral blood of pregnant women, has been widely used in clinical practice due to its non-invasive, safe, and high accuracy [1]. However, the detection efficacy of NIPT is not static, and its accuracy is heavily dependent on the relative concentration of fetal cfDNA in maternal blood, i.e., the fetal fraction [2]. For pregnant women with male fetuses, Y chromosome concentration is often used as a reliable surrogate for fetal fraction, and it is usually clinically required to be no less than 4% to ensure the reliability of the test results [3].

The cfDNA concentration of fetuses changes dynamically with the increase of gestational age, and is affected by multiple factors such as maternal body mass index (BMI), age, and weight [4]. If the test is premature, the fetal score may not reach the threshold, resulting in test failure or false negative results [5]; If the detection is too late, the diagnosis and intervention of abnormal fetuses will be delayed, and the valuable clinical treatment window will be shortened [6]. Therefore, how to scientifically determine the optimal time point for NIPT testing according to the individual characteristics of pregnant women to minimize potential clinical risks while ensuring the success rate of testing is a research problem of great practical significance. At present, clinical practice often determines the detection time based on experience or rough BMI grouping, and lacks refined and data-driven decision support.

In order to solve the above problems, this study aims to construct a systematic algorithm framework for optimizing the time point of NIPT detection. This study is mainly divided into three steps: The first step is to explore the key factors affecting the concentration of male Y chromosomes. We used Spearman rank correlation analysis to screen significant variables, and constructed a random forest regression model to quantify the importance of each feature to provide a basis for subsequent grouping. The second step focuses on univariate grouping optimization based on BMI. The pregnant women were divided into groups with high homogeneity by K-means clustering, and a multi-objective function was constructed for each group, which combined the risk of early detection failure, operational error and diagnosis delay, and the recommended gestational age to minimize the group risk was solved. The third step is to further consider the multi-factor synergy effect. More dimensional features such as standard ratio and GC content are introduced, and multi-factor clustering is used for population segmentation, and time-based optimization is carried out using weighted risk functions in order to obtain more accurate and personalized recommendations [7].

In recent years, machine learning and data mining technologies have been increasingly used in medical decision support systems [8]. Clustering algorithms such as K-means have been used for patient stratification [9], while ensemble learning models such as random forests perform well in feature screening and prediction [10]. This study integrates these advanced algorithms to solve the specific clinical problem of timing optimization of NIPT detection, aiming to provide quantifiable algorithmic tools and theoretical references for the precise and personalized implementation of prenatal screening.

2. Methods

2.1. Analysis of influencing factors of Y chromosome concentration

In order to explore the correlation between fetal Y chromosome concentration and gestational age, BMI and other indicators, and to identify key influencing factors, data analysis and modeling were first carried out.

In order to ensure that the gestational age data is continuously comparable, the "weeks + days" format is uniformly converted to the continuous gestational age value t_i. Only data from male fetal samples with available Y chromosome concentrations were retained for analysis. In order to capture the possible nonlinear monotonic relationship between variables, the Spearman rank correlation coefficient was used for analysis. For variables X (e.g., Y concentration, gestational age t, BMI b), their rank $R_{X(i)}$ were treated with parallel values by the median rank method.

The Spearman correlation coefficient between Y concentration and gestational age t is calculated as follows:

$$\rho_{Y,t}^{(S)} = \frac{\sum_{i\in I}(R_Y(i)-\overline{R_Y})(R_t(i)-\overline{R_t})}{\sqrt{\sum_{i\in I}(R_Y(i)-\overline{R_Y})^2}\sqrt{\sum_{i\in I}(R_t(i)-\overline{R_t})^2}} \quad (1)$$

Similarly, the correlation coefficient of Y concentration with BMI b $\rho_{Y,b}^{(S)}$ is calculated. To control for confounding factors, the correlation between Y and gestational age was further calculated after controlling for BMI $\rho_{Y,t|b}^{(S)}$. Significance was assessed by permutation test, and the false finding rate of multiple comparisons was controlled using the Benjamin-Hochberg method.

In order to quantify the prediction contribution of each feature to Y concentration and capture the complex nonlinear relationship, a stochastic forest regression model was constructed. The eigenvector X_i included gestational age, BMI, age, height, weight, and various sequencing quality indicators.

The model is defined as:

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \quad (2)$$

where $f_b(\cdot)$ is the b-th decision tree, and B is the total number of trees. Model performance is evaluated by out-of-pocket error. Feature importance is measured by calculating the increment of the out-of-pocket mean square error before and after the displacement of the eigenvalue. In addition, the mean marginal effect of key features (e.g., gestational age, BMI) on Y concentration was visualized through bias-dependent graphs.

2.2. Detection time optimization based on BMI univariate grouping

This step aims to group pregnant women according to their BMI and determine an optimal time point for NIPT testing for each group to minimize the combined clinical risk.

For each male fetal sample, the gestational age in which its Y chromosome concentration reaches or exceeds 4% for the first time is defined as the "earliest time to reach" T_i. In order to eliminate the influence of dimensions on clustering, BMI was normalized with zero mean and unit variance, and $\tilde{b}$ was obtained.

On the standardized BMI space, the K-means algorithm is used for clustering, and the goal is to minimize the sum of squares within the cluster:

$$\min_{\{C_g,c_g\}}\sum_{g=1}^{k}\sum_{i\in C_g}||\tilde{b}_i - c_g||^2 \quad (3)$$

The optimal cluster number k was determined by the elbow method (observing the square sum and inflection point within the cluster) and the contour coefficient. After clustering, they were renumbered in ascending order of the mean BMI of each cluster to obtain clinically explainable grouping.

A comprehensive risk function $R_g(t)$ is designed for each BMI group, which integrates four risks: (1) early detection risk (probability of failure to meet the standard at the time of detection); (2) failure risk (complementarity of success rate); (3) Error risk (considering the uncertainty of the detection time point and the time of reaching the standard); (4) Delay punishment (secondary punishment for detecting too late).

$$R_g(t) = a_1\Pr(T>t) + a_2(1-\Pr(T\le t)) + a_3\frac{1}{n_g}\sum_{i\in g}\Phi\left(\frac{T_i-t}{\sigma_e}\right) + a_4(\max\{0, t-\overline{T}_g-\delta\})^2 \quad (4)$$

where, a_1 to a_4 is the weight coefficient, $\Phi(\cdot)$ is the standard normal distribution function, $\overline{T}_g$ is the average earliest time in the group, and δ is the buffer period. In the clinically feasible interval $[\max(10,\overline{T}_g-3), \min(20,\overline{T}_g+2)]$, the optimal detection time point t_g^{opt} for minimizing the risk function was solved by one-dimensional bounded optimization.

In order to evaluate the stability of the recommended time point under the actual operation error, Monte Carlo simulation was adopted. Gaussian noise was added to the planned detection time point t and the individual standard time T_i, and the robust success rate $S_g^{rob}(t)$ was calculated after simulating M times:

$$S_g^{rob}(t) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_g}\sum_{i\in g}\mathbf{1}(T_{i,m}\le \tilde{t}_m) \quad (5)$$

2.3. Joint optimization of grouping and time point under multi-factor fusion

In addition to BMI, feature vectors including age, height, weight, GC content, number of pregnancies, births, and individual "standard attainment ratio" r_i (the proportion of Y concentration ≥4% in historical testing) were constructed. The random forest model was used to evaluate the importance of each feature in predicting the time to reach the standard,

and the subset S of features with the highest importance was selected for subsequent clustering.

After normalizing the selected features, multivariate K-means clustering was performed. In each clustering group, the ordinary empirical standard distribution $F_{g(t)}$ and the empirical distribution $F_g^w(t)$ weighted by the standard ratio r_i were calculated, respectively.

The extended comprehensive objective function \(\mathcal{J}_g(t)\ is constructed, and in addition to the risks of early detection, failure, error, and delay, additional risk items based on weighted distribution and inconsistency of the proportion of compliance within the penalty group are added.

$$J_g(t) = 0.25(1-F_g(t)) + 0.15(1-F_g(t)) + 0.25(1-F_g^w(t)) + 0.20E_g(t) + 0.10D_g(t) + 0.05C_g \quad (6)$$

where $E_{g(t)}$ is the weighted error risk term, $D_{g(t)}$ is the delay penalty, and C_g is the consistency penalty.

In the feasible domain defined by the weighted average time to reach the standard T_g^w, the optimal time point t_g^* of each group is optimized. The robust success rate $\hat{S}_g^{rob}$ is also calculated by Monte Carlo simulation. Finally, according to the sample size weighting of each group, the overall robust success rate and the overall risk of early detection are calculated.

3. Results and discussion

3.1. Identification of key influencing factors of Y chromosome concentration

Table 1. Performance of Stochastic Forest Model and Importance of Key Features (Top 5)

Ranking	Features	Spearman correlation coefficient (r)	Random forest importance weight
1	X chromosome concentration	0.4744	0.3962
2	Y chromosome Z value	-0.1876	0.0983
3	Comparison ratio	-0.2749	0.0868
4	X chromosome Z value	-0.1095	0.0570
5	Gestational week	0.1034	0.0329

A total of 1082 male fetal samples were included for analysis. Spearman correlation analysis showed that 14 of the 21 initial features were significantly correlated with Y chromosome concentration (FDR < 0.05). Among them, the X chromosome concentration had the strongest positive correlation with the Y concentration (r=0.4744, p<0.001), which is consistent with some synergistic effects of X and Y chromosomes in sequencing in biology. Gestational age was weakly positively correlated with Y concentration. On the contrary, pregnant womens BMI, weight, Z value of chromosome 18 and comparison ratio were negatively correlated with Y concentration, suggesting that factors such as high BMI may "dilute" fetal cfDNA and delay the increase of its concentration.

To further quantify the importance of features, the constructed stochastic forest regression model showed excellent fitting performance (R^2=0.9468, RMSE=0.0078, MAE=0.0058). Table 1 feature importance ranking confirms that X chromosome concentration is the most predictive variable (importance weight 0.396), which is much higher than other features. Y chromosome Z value, alignment ratio, X chromosome Z value and gestational age ranked second to fifth, respectively. BMI has an importance weight of about 0.025, ranking in the middle. The comparison chart between the predicted value and the real value and the residual diagnosis showed that the model fit well and the residuals were randomly distributed. The results of this step not only validate the clinical experience, but more importantly, provide a key feature selection basis for subsequent grouping optimization.

3.2. Optimization results based on BMI grouping

Table 2. Optimized results based on BMI univariate grouping

Group	BMI range	Number of samples	Average time to reach the standard (weeks)	Recommended testing time point (weeks)	Theoretical success rate	Robust success rate	Risk of early detection
Group 1	[20.7, 31.4]	134	13.3 ± 2.4	14.8	0.858	0.856	0.142
Group 2	[31.6, 35.9]	105	13.4 ± 2.2	14.9	0.867	0.868	0.133
Group 3	[36.3, 45.7]	21	16.7 ± 4.3	18.2	0.714	0.714	0.286
Overall	-	260	-	-	-	0.849	0.150

260 male fetal samples with a clear earliest time to reach the standard were analyzed. The optimal clustering number k=3 was determined by the elbow method and the contour coefficient. The clustering results divided pregnant women into three groups according to BMI: group 1 [20.7, 31.4], group 2 [31.6, 35.9], and group 3 [36.3, 45.7].

By optimizing the comprehensive risk function of each group, the recommended detection time point is obtained. The average time to reach the standard in group 1 and group 2 was similar (about 13.3-13.4 weeks), and the optimal time point was concentrated in 14.8-14.9 weeks, which was slightly later than the average time to reach the standard, reflecting the avoidance of the risk of "early inspection failure". The success rate of theory and robustness in both groups was as

high as about 0.86-0.87, and the risk of early detection was about 0.14. The average time to reach the standard in group 3 (high BMI group) was significantly delayed to 16.7 weeks, and the dispersion was also greater. The recommended test time point was optimized to 18.2 weeks, with a success rate of 0.714 and an early detection risk of 0.286 (see Table 2). Monte Carlo simulation showed that after introducing reasonable operating errors, the success rate of robust in each group decreased very little (<0.2%), indicating that the recommended scheme had good stability. In the end, the overall robust success rate was 0.849 and the overall early detection risk was 0.150. The results clearly suggest that delaying NIPT testing is a safer strategy for pregnant women with high BMI.

3.3. Multi-factor fusion optimization results

Table 3. Multi-factor fusion grouping optimization results

Group	Description of main characteristics	Recommended detection time point (weeks)	Robust success rate	Risk of early detection	Influence of error
Group 1	Multi-feature equilibrium	14.1	0.808	0.192	-0.5%
Group 2	BMI is moderate, and the proportion of compliance is good	15.2	0.861	0.139	-0.8%
Group 3	The proportion of compliance is high, and the GC content is advantageous	14.5	0.939	0.061	-3.3%
Group 4	High BMI and low proportion of compliance with standards	18.5	0.656	0.344	-1.1%
Overall	-	-	0.863	0.161	-

In the 186 samples that included multiple factors such as compliance ratio and GC content, the randomized forest importance assessment showed that "compliance ratio" was the most critical feature (weight≈0.64), followed by GC content (≈0.19) and BMI (≈0.14). According to the importance of features, multivariate clustering was carried out, and the population was finally divided into four groups by combining the contour coefficient and elbow method.

The optimized recommended detection time and performance are shown in Table 3. The recommended gestational age from group 1 to group 3 was between 14.1 and 15.2 weeks, and the robust success rate ranged from 0.808 to 0.939, of which group 3 had the highest success rate but was slightly sensitive to the point error during testing. Group 4 was characterized by a higher BMI and a low proportion of compliance with the standard, and its recommended testing time point was significantly delayed to 18.5 weeks, but the robust success rate was also the lowest (0.656). The visualization results showed that the multi-factor grouping was not a strict BMI interval cutting, but there was overlap in BMI between the groups, reflecting the characteristics of multi-feature co-decision-making. The overall robust success rate and the overall risk of early detection were 0.161, which was slightly improved compared with the single-factor grouping. This suggests that the introduction of characteristics that reflect the testing history of individuals, such as the "compliance ratio", can further refine population stratification and optimize point-in-time decision-making, especially helping to identify high-risk delayed testing populations that require special attention (such as group 4).

4. Conclusion

In this study, a complete algorithm framework from factor analysis, population grouping to point-in-time decision-making was constructed to address the risk of detection failure and diagnosis delay caused by insufficient fetal DNA concentration in NIPT testing. Firstly, the key factors affecting the concentration of Y chromosomes in male fetuses were systematically identified by Spearman correlation and random forest model, among which X chromosome concentration, sequencing quality indicators (such as comparison ratio) and gestational age played the most prominent roles, which provided data support for understanding the feasibility of detection. Secondly, a grouping strategy based on K-means clustering is proposed, which develops the traditional single BMI threshold division into a data-driven group division method with higher homogeneity within the group. An innovative comprehensive risk objective function that integrates early detection failure, operational error and clinical delay penalty is designed, and the optimization of the detection time point is transformed into a solvable mathematical planning problem, so as to recommend personalized optimal detection of gestational age for different characteristic populations.

The experimental results show that when grouping based on BMI alone, the high BMI group (>36.3) needs to postpone the testing time from about 15 weeks to more than 18 weeks to maintain the overall robust success rate at about 85%. When the "standard compliance ratio", which reflects the stability of individual testing history, and multi-dimensional features such as GC content are further introduced, the algorithm can further refine the high-risk subgroups (such as low compliance ratio combined with high BMI), and give more differentiated time point suggestions (14.1-18.5 weeks), so that the overall optimization performance can be improved. The Monte Carlo simulation verifies the robustness of the recommended scheme under reasonable operating errors.

The contribution of this study is to transform clinical empirical rules into quantifiable and optimizable data-driven models, providing algorithmic tools for the accurate implementation of NIPT detection. However, this study still has certain limitations: the model is based on retrospective data, and its effectiveness needs to be further verified in a prospective clinical cohort. more potential confounding factors such as the race and disease history of pregnant women were not taken into account; The model is currently statically optimized, and in the future, it can be explored to integrate into the time series prediction model to achieve more dynamic time point recommendation. Future work also includes extending the framework to multiple pregnancy

scenarios and exploring the feasibility of simplifying the model in healthcare resource-constrained settings to facilitate the application of the algorithm in broader clinical settings.

Acknowledgements

This paper was supported by my teacher Professor Lu.

References

[1] Gil M M, et al. Analysis of cell-free DNA in maternal blood in screening for fetal aneuploidies: updated meta-analysis. Ultrasound Obstet Gynecol, 2020, 55(3): 302-314.

[2] Hui L, et al. Clinical implementation of cell-free DNA-based aneuploidy screening: perspectives from a multinational survey. Prenat Diagn, 2022, 42(8): 975-986.

[3] Yaron Y. The implications of non-invasive prenatal testing failures: a review of an under-discussed phenomenon. Prenat Diagn, 2020, 40(6): 677-683.

[4] Wang E, et al. Gestational age and maternal weight effects on fetal cell-free DNA in maternal plasma. Prenat Diagn, 2021, 41(7): 810-817.

[5] Hudecova I, et al. Maternal plasma fetal DNA fractions in pregnancies with low and high maternal BMI: implications for noninvasive prenatal testing. Prenat Diagn, 2022, 42(1): 54-61.

[6] Norton M E, et al. Cell-free DNA analysis for noninvasive examination of trisomy. N Engl J Med, 2015, 372(17): 1589-1597.

[7] Esteva A, et al. A guide to deep learning in healthcare. Nat Med, 2019, 25(1): 24-29.

[8] Deo R C. Machine Learning in Medicine. Circulation, 2015, 132(20): 1920-1930.

[9] López-Moral A, et al. Machine learning techniques for non-invasive prenatal testing: a systematic review. Bioinform Biol Insights, 2023, 17: 11779322231152420.

[10] Breiman L. Random Forests. Machine Learning, 2001, 45(1): 5-32.