NIPT Time Point Stratified Modeling and Fetal Anomaly Determination Analysis Based on Multivariate Nonlinear Regression

Min Chen[1,a], Bikun Song[1,b], Yuke Wang[1,c]

[1]School of Mathematics and Statistics, Northeastern University at Qinhuangdao, Qinhuangdao, Hebei, China

[a]1627742767@qq.com, [b]1570311650@qq.com, [c]2315611690@qq.com

Keywords: NIPT; multiple linear regression; random forest; Cox returns; risk function; Cluster analysis

Abstract: The accuracy of non-invasive prenatal testing (NIPT) is highly affected by the concentration of free DNA in the fetus, and the male fetus needs to rely on the concentration of Y chromosome to ensure the reliability of the test. In this paper, statistical regression and cluster optimization methods are used to systematically solve the problems of Y chromosome concentration modeling and optimal detection time recommendation. In the first step, the Spearman rank correlation coefficient was used to analyze the relationship between Y chromosome concentration and gestational age, BMI and other indicators, and multiple linear and nonlinear regression models were established. The results showed that gestational age was significantly positively correlated with Y chromosome concentration, and BMI was significantly negatively correlated, and the goodness of fit R^2 of the nonlinear model reached 0.095, which was better than the linear model and passed the significance test. In the second step, a data-driven hierarchical model based on BMI is constructed, and K-means clustering is used to balance the risk of detection failure and delay by weighted risk function, and the optimal detection time point is optimized. The results showed that the high BMI group (group 1 and group 2) needed 16 weeks to meet the standard, and the rest group took 13 weeks. In the third step, multidimensional features such as age and X chromosome concentration were fused, and the weights (X chromosome 0.42, BMI 0.31, age 0.27) were determined by recursive feature elimination and entropy weight method, and five groups were determined by clustering with K-means. The model is verified to be reliable by stability analysis.

1. Introduction

Non-Invasive Prenatal Testing (NIPT), as an important technology of modern prenatal screening, realizes early and non-invasive screening of chromosomal aneuploidy by analyzing fetal cell-free fetal DNA (cffDNA) in the peripheral blood of pregnant women. This technique is widely used in clinical practice due to its safety and high sensitivity, but its detection accuracy is highly dependent on the relative concentration of fetal DNA in the maternal blood (fetal concentration). Studies have shown that when the fetal concentration is below 4%, the reliability of NIPT detection is significantly reduced, and false negative or false positive results may occur [1].

For male fetuses, Y chromosome concentration is a key indicator for evaluating fetal concentration, and reaching or exceeding 4% is considered a key threshold to ensure the reliability of detection. A large number of clinical observations have shown that the concentration of male Y chromosome is closely related to the gestational age and body mass index (BMI) of pregnant women. As gestational age increases, fetal concentration usually rises; An increase in BMI in pregnant women may reduce fetal concentration through blood thinning [2]. This complex interaction makes it particularly important to determine the optimal timing of testing: early testing can lead to test failure due to insufficient fetal concentration, while late testing may delay the diagnosis and intervention of abnormal fetuses.

At present, in clinical practice, the selection of NIPT detection time points is mostly based on empirical recommendations, and there is a lack of individualized and accurate decision-making

  DOI: 10.25236/iwmecs.2025.004

support. Especially for pregnant women with different BMI characteristics, it is often difficult to achieve risk optimization with a unified testing point in time. The detection failure rate of pregnant women with high BMI is significantly higher than that of the normal weight group, and the success rate of re-sampling is limited, which poses a serious challenge to clinical diagnosis [3]. Therefore, it is of great theoretical value and clinical significance to establish a mathematical model based on real clinical data, scientifically divide the group of male pregnant women and determine the optimal time point for NIPT detection.

Based on large-scale clinical test data, this paper uses statistical regression and machine learning to systematically solve the two core problems of Y chromosome concentration modeling and optimal detection time recommendation. The research content mainly includes three key steps: The first step is to establish a regression model of Y chromosome concentration, gestational age, and BMI to reveal the quantitative relationship and related characteristics between them. We used Spearman rank correlation analysis to explore the strength of the association between variables, and constructed multiple linear and nonlinear regression models for comparative analysis to determine the optimal modeling strategy [4].

The second step focuses on the time-of-detection optimization method based on BMI stratification. In view of the detection differences in different BMI groups, we use the K-means clustering algorithm to achieve objective grouping of pregnant women, construct a weighted risk function to comprehensively balance the risk of detection failure and delay, and solve the optimal NIPT detection time point for each BMI group through the optimization algorithm [5]. This method effectively solves the limitations of the traditional "one-size-fits-all" strategy and provides a scientific basis for personalized testing.

The third step is to further expand the feature dimension and optimize the grouping strategy by integrating multi-dimensional features such as age and X chromosome concentration. Recursive feature elimination (RFE) is used for feature selection, and the objective weight of each feature is determined in combination with the entropy weight method, and a more refined grouping system is established, thereby improving the detection timeliness and reliability [6]. This innovative approach not only considers the impact of BMI but also incorporates other important physiological and detection indicators, making the model more clinically applicable.

This study provides reliable data support and decision-making basis for the scientific selection of clinical NIPT detection time through systematic modeling, analysis and optimization calculation. The results not only help improve the accuracy and reliability of NIPT testing, but also provide practical tools for clinicians to develop personalized prenatal screening programs, ultimately achieving the public health goals of eugenics and eugenics [7]. In addition, the methodological framework established in this study can also provide reference for other biomarker-based time-of-detection optimization problems.

2. Model creation, solution and discussion

2.1. Model establishment

2.1.1. Y chromosome concentration regression model

In order to explore the quantitative relationship between Y chromosome concentration and gestational age, BMI and other key indicators, the data were preprocessed first. The gestational age data is uniformly converted to days for easy calculation, and the missing values are supplemented by the reverse calculation method. Spearman rank correlation coefficient analysis showed that Y chromosome concentration was positively correlated with gestational age and negatively correlated with BMI, which provided a theoretical basis for subsequent modeling.

Based on the results of correlation analysis, a multiple linear regression model was constructed as a benchmark model:

$$Y = \beta_0 + \beta_1 \times h + \beta_2 \times BMI + \varepsilon \tag{1}$$

Y represents the concentration of Y chromosomes, h is the number of gestational weeks, BMI is

the body mass index of pregnant women, and ε is the random error term.

Considering that there may be nonlinear relationships between variables in the biological process, in order to further improve the fitting ability of the model, a cubic polynomial regression model is constructed by introducing nonlinear terms on the basis of the multiple linear regression model:

$$Y = \beta_0 + \beta_1 \times h + \beta_2 \times BMI + \beta_3 \times h^2 + \beta_4 \times BMI^2 + \beta_5 \times (h \times BMI) + \beta_6 \times h^3 + \beta_7 \times BMI^3 + \varepsilon \quad (2)$$

By introducing higher-order terms and interaction terms, the model can better capture the complex nonlinear relationships between variables, which is closer to the actual biological process.

2.1.2. Optimal detection time-based model based on BMI stratification

In clinical practice, there are significant differences in fetal cell-free DNA concentrations in pregnant women with different BMIs, which directly affects the reliability of NIPT detection. In order to develop personalized testing strategies, it is necessary to scientifically stratify pregnant women according to BMI. In this study, the K-means clustering algorithm was used to divide pregnant women into five subgroups with significant differences based on BMI values, and each cluster represented a group of pregnant women with similar BMI characteristics and detection risk characteristics.

On the basis of hierarchy, a time-of-detection optimization model with minimal risk is constructed. NIPT detection risks mainly include the risk of detection failure and delay: the former involves the reliability of the test at the technical level, and the latter is related to the timeliness of the clinical intervention time window. The weighting method is used to integrate the two types of risks and construct the overall risk function:

$$Risk(h, B) = 0.7R_{fail}(h, B) + 0.3R_{delay}(h) \quad (3)$$

where $R_{fail}(h, B)$ represents the probability that the fetal Y chromosome concentration does not reach the 4% threshold under the condition of gestational age h and BMI of B; $R_{delay}(h)$ represents the limited effect of the clinical intervention time window caused by the delay in detection. The weight allocation reflects the difference in risk importance, with the risk of detection failure accounting for 0.7 and the risk of delay accounting for 0.3.

The optimization goal is to minimize the weighted risk function, balance technical feasibility and clinical timeliness, and determine the optimal detection point for each BMI group.

2.1.3. Detection time-of-point optimization model for multi-feature fusion

In order to further improve the grouping accuracy and the accuracy of the detection time recommendation, multidimensional features such as age and X chromosome concentration were introduced on the basis of BMI. Firstly, the random forest regression model was used to evaluate the importance of each feature to the time of reaching the Y chromosome concentration, and the key feature sets were screened out by recursive feature elimination (RFE), and finally the three core characteristics of pregnant women's BMI, age and X chromosome concentration were finally determined.

In order to objectively determine the weights of each feature, the entropy weight method is used for weight analysis. Based on the information entropy theory, the entropy weight method can effectively reduce the influence of subjective factors by calculating the information contribution of each feature to determine the weight. The results showed that the weight of X chromosome concentration was the highest at 0.42, and the information contribution was the largest, followed by BMI in pregnant women, accounting for 0.31, and the weight of age was 0.27.

Based on the weight analysis results, a comprehensive scoring function is constructed:

$$J = W_{BMI} \times R_{BMI} + W_{Xconc} \times R_{Xconc} + W_{Age} \times R_{Age} \quad (4)$$

Among them, $W_{BMI} = 0.31, W_{Xconc} = 0.42, W_{Age} = 0.27$ are the weights calculated by the entropy

weight method, and R_{BMI}, R_{Xconc}, *and* R_{Age} represent the specific values of the corresponding features.

On this basis, the K-means clustering algorithm was used to divide the pregnant women into five groups, and the optimal detection time of each group was determined by the Y chromosome concentration compliance rate of ≥95%.

2.2. Model Solution and Results

2.2.1. Step 1 model solution results

By comparing the fitting effect and statistical significance of linear and nonlinear regression models, the following results are obtained:

Table 1 Comparison of performance between linear regression and random forest model

Model Type	R^2	Adjusted R^2	F-Value	F-Test P-Value	Main significant variables
Linear model	0.047	0.044	15.47	2.76e-07	gestational age (t=4.53, p<0.001) BMI (t=-3.86, p<0.001)
Nonlinear model	0.095	0.082	7.26	4.51e-10	gestational age (t=2.30, p=0.022) BMI (t=3.82, p<0.001)

As shown in Table 1, in the performance comparison between the linear and nonlinear models, the linear model exhibits a goodness of fit (R^2) of 0.047, an adjusted R^2 of 0.044, and an overall F-test statistic of 15.47, with a corresponding p-value of 2.76×10^{-7}, indicating that the model is statistically significant overall. Among the variables, gestational age ($t = -4.53$, $p < 0.001$) shows a significant positive correlation with Y-chromosome concentration, while BMI ($t = -3.86$, $p < 0.001$) demonstrates a significant negative correlation, which is consistent with physiological expectations. In contrast, the nonlinear model achieves an improved R^2 of 0.095 and an adjusted R^2 of 0.082, representing a notable enhancement over the linear model. Its overall F-statistic is 7.26, with a corresponding p-value of 4.51×10^{-10}, also reaching a high level of statistical significance. Based on these three indicators, the nonlinear model significantly outperforms the linear model in terms of goodness of fit and exhibits stronger explanatory power and practical applicability in describing the relationship between Y-chromosome concentration and gestational age as well as BMI.

2.2.2. Step 2 model solution results

Based on the K-means clustering algorithm, 603 pregnant women were divided into 5 subgroups with significant differences according to BMI values, and the statistical characteristics and optimal detection time points of each group were as follows:

Table 2 Recommended gestational age for different BMI groups

Group number	Number of samples	BMI range	Mean± standard deviation	Best time point of detection (days)
0	104	26.62-29.64	28.7±0.66	93.1
1	117	33.69-36.70	34.8±0.85	111.8
2	43	36.81-44.70	38.6±1.86	118.9
3	176	29.66-31.48	30.5±0.48	96
4	163	31.57-33.63	32.4±0.62	96.8

As presented in Table 2, the optimal detection time points for the low to moderate BMI groups (Groups 0, 3, and 4) are concentrated between 90–100 days of gestation, while in the high BMI group (Group 1) and the very high BMI group (Group 2), the recommended detection timing is significantly delayed to approximately 112 days and 119 days, respectively. These results indicate that elevated BMI delays the time required for fetal Y-chromosome concentration to reach a detectable threshold, thereby postponing the window for reliable NIPT testing. Sensitivity analysis further revealed that the extreme BMI groups (Groups 0 and 4) exhibit strong decision robustness, whereas the intermediate groups—particularly Group 2—are sensitive to penalty weighting, underscoring the value and necessity of stratified modeling based on BMI categories.

2.2.3. Step 3 model solution results

After fusing multiple features, pregnant women were divided into five groups by K-means clustering, and the key characteristics and optimal detection time points of each group were as follows:

Table 3 Recommended gestational age for multivariate clustering

Cluster group	BMI interval (kg/m^2)	Mean X chromosome concentration	Mean age (years)	Best time point	Compliance rate threshold
0	31.65-36.89	0.032	34.2	13 weeks (90 days)	≥95%
1	26.62-33.21	0.028	29.8	11 weeks (77 days)	≥95%
2	36.36-46.88	0.041	41.5	11 weeks (79 days)	≥95%
3	27.92-32.45	0.025	28.3	11 weeks (77 days)	≥95%
4	31.32-36.25	0.035	32.7	11 weeks (80 days)	≥95%

Table 3 shows that, compared to the BMI-based stratification results in Table 2, the optimal detection time point for most cluster groups after incorporating multiple features (age and X-chromosome concentration) advances to 11 weeks (77–80 days), with only Group 0 requiring 13 weeks (90 days). This indicates that the inclusion of additional maternal characteristics allows for finer stratification of pregnant women and further optimizes the timing of NIPT detection. Stability analysis reveals that the model's adjusted average Rand index (ARI) is 0.6953, indicating moderate-to-high consistency across multiple K-means runs. Moreover, across varying error levels (5%–20%), the fluctuation range of the optimal detection time point remains limited, demonstrating good robustness of the proposed multivariate clustering model.

2.3. Results and discussion

The first step showed that the nonlinear regression model was significantly better than the linear model in fitting the relationship between Y chromosome concentration, gestational age and BMI, and the R^2 increased from 0.047 to 0.095. This finding confirms that there is a complex nonlinear relationship between Y chromosome concentration, gestational age and BMI, and it is difficult to fully capture the interaction between variables using linear models alone. Gestational age is positively correlated with Y chromosome concentration, which is consistent with the physiological process of fetal development with increasing gestational age, and the proportion of fetal free DNA in maternal blood increases. BMI is negatively correlated with Y concentration, which may be due to a decrease in the proportion of fetal DNA due to blood-thinning effects or metabolic differences in obese pregnant women.

In the second step, the optimal detection time point for different groups was determined through BMI stratification, and a data-driven personalized detection strategy was established. The results showed that BMI had a significant effect on the timing of detection, and the optimal detection time point in the low BMI group (group 0) was 93.1 days, while the optimal detection time point in the high BMI group (group 2) was delayed to 118.9 days. This finding has important guiding significance for clinical practice, and it is recommended to appropriately postpone the NIPT test time for pregnant women with high BMI to improve the success rate of testing. Sensitivity analysis revealed the heterogeneous responses of different BMI groups to risk weights, which provided a basis for personalized setting of clinical parameters.

After the fusion of multiple features in the third step, the optimal detection time point for most groups was advanced to within 11 weeks, and the detection efficiency was significantly improved. This suggests that relying solely on BMI for stratification may ignore the influence of other important factors, such as age and X chromosome concentration. The multi-feature fusion model can more comprehensively reflect the physiological state of pregnant women and provide more accurate detection time recommendations. The good stability of the model further proves the feasibility of the method in clinical practice, which can provide reliable support for personalized time point selection for NIPT detection.

3. Conclusion

In this paper, a systematic mathematical model and optimization method are established for the modeling of Y chromosome concentration and the determination of the optimal detection time in NIPT detection. The first step was to reveal the quantitative relationship between Y chromosome concentration and gestational age and BMI through multiple nonlinear regression model, and the goodness of fit of the nonlinear model was significantly better than that of the linear model, which provided a theoretical basis for understanding the variation law of fetal cell-free DNA concentration. In the second step, a risk-minimized detection time optimization model was established based on BMI stratification, and pregnant women were divided into five subgroups with significant differences through K-means clustering, and the optimal detection time point for each group was determined, providing a personalized solution for solving the problem of detection failure caused by BMI differences in clinical practice. In the third step, multi-dimensional features such as age and X chromosome concentration were further integrated, and the entropy weight method was used to determine the feature weights, and a more refined grouping system was constructed, which advanced the optimal detection time point for most pregnant women to within 11 weeks, significantly improving the detection efficiency. The results show that the data-driven hierarchical modeling method can effectively balance the risk of detection failure and delay, and provide reliable technical support for personalized recommendation at the time of clinical NIPT testing. Future studies can further incorporate more clinical features, such as pregnancy history, genetic factors, etc., to refine the model and improve its clinical applicability.

Acknowledgements

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

References

[1] Bianchi DW, et al. DNA sequencing versus standard prenatal aneuploidy screening. N Engl J Med 2014;370(9):799-808.

[2] Wang E, et al. Gestational age and maternal weight effects on fetal cell-free DNA in maternal plasma. Prenat Diagn 2013;33(7):662-666.

[3] Suzumori N, et al. Factors affecting cell-free DNA fetal fraction and the consequences for test accuracy. J Matern Fetal Neonatal Med 2018;31(11):1485-1490.

[4] Hastie T, et al. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer; 2009.

[5] Hartigan JA, Wong MA. A K-means clustering algorithm. Journal of the Royal Statistical Society 1979;28(1):100-108.

[6] Guyon I, et al. An introduction to variable and feature selection. J Mach Learn Res 2003; 3:1157-1182.

[7] American College of Obstetricians and Gynecologists. Cell-free DNA screening for fetal aneuploidy. Committee Opinion No. 640. Obstet Gynecol 2015;126:e31-37.