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

Optimization and anomaly judgment of NIPT detection based on multivariate statistical model and machine learning

Ziqi Zhao[1], Yangsai Zhou[1], and Qianziyi Guo[2]

[1] Qiusuo Honors College, Tianjin Foreign Studies University, Tianjin, China
[2] International Business College, Tianjin Foreign Studies University, Tianjin, China

Abstract: In this study, focusing on non-invasive prenatal detection technology, using the individual characteristics and sequencing data of pregnant women, a multivariate statistical model and machine learning algorithm were constructed to systematically analyze the relationship between fetal sex chromosome concentration changes and the optimal detection time. Firstly, through Spearman rank correlation analysis and multiple linear regression model, the significant effects of gestational age and BMI on Y chromosome concentration were revealed. Secondly, the pregnant women were divided into three groups according to BMI by K-means clustering, and the optimal time points for each group were determined by linear fitting, which were 12 weeks and 2 days (low BMI group), 16 weeks and 4 days (medium BMI group) and 18 weeks and 5 days (high BMI group). The thin plate spline interpolation model was further introduced to capture the nonlinear interaction effect between age, height and Y concentration, and the X chromosome concentration was identified as the main source of technical error. The results show that the NIPT time-point recommendation strategy based on BMI grouping and multivariate modeling can significantly improve the reliability of detection and provide a theoretical basis for clinical personalized detection.

Keywords: Non-invasive prenatal testing; Multiple linear regression; Cluster analysis; Spline interpolation; Error analysis; Machine learning.

1. Introduction

With the rapid development of high-throughput sequencing technology, non-invasive prenatal testing (NIPT) has become an important means of early screening for fetal chromosomal abnormalities. NIPT can achieve non-invasive and early detection of fetal chromosomal aneuploidy by analyzing fetal cell-free DNA in the peripheral blood of pregnant women [1]. However, the accuracy of the test is affected by a combination of biological and technical factors, including the gestational age of the pregnant woman [2], body mass index (BMI) [3], sequencing data quality [4], GC content [5], and fetal DNA ratio [6]. Therefore, it is of great significance to scientifically determine the optimal detection time, identify the main sources of error, and establish a personalized detection model to improve the clinical applicability and reliability of NIPT [7].

In recent years, many studies have focused on optimizing NIPT detection strategies. For example, Chen et al. (2021) confirmed a positive correlation between gestational age and fetal DNA concentration through a large-scale cohort study [8]; Liu et al. (2022) pointed out that BMI is a key factor affecting the proportion of fetal DNA [9]. Wang et al. (2023) further used machine learning methods to improve the accuracy of abnormal chromosome detection [10]. However, most of the existing studies focus on single-factor analysis, and there is a lack of systematic modeling of multivariate interaction effects and nonlinear relationships [11], especially in the analysis of point-in-time recommendations and error mechanisms based on BMI grouping [12].

This paper aims to construct a complete modeling framework to systematically solve the problems of time point selection and error control in NIPT detection through three-step progressive analysis. In the first step, a quantitative relationship model between gestational age, BMI and Y chromosome concentration was established. In the second step, based on BMI cluster analysis, the best detection time recommendation strategy for grouping is proposed. In the third step, more biological and technical variables are introduced to identify nonlinear effects and sources of error using plate spline interpolation [13]. This study not only provides a scientific basis for clinical NIPT detection, but also provides a method reference for multivariate medical data modeling [14][15].

2. Methods

2.1. Modeling of the relationship between gestational age, BMI and Y chromosome concentration

In order to explore the relationship between Y chromosome concentration, gestational age and BMI, the data were preprocessed, female fetal samples and missing values were removed, gestational age was converted into decimal form, and GC content and other indicators were standardized. Spearmans rank correlation analysis was used to test the correlation between variables, and the nonparametric method was suitable for nonlinear relationship testing:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (1)$$

Where d_i is the rank difference of the two variables, and n is the sample size.

On the basis of correlation testing, a multiple linear regression model is established to quantify the influence of each variable:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{11} x_{11} + \varepsilon \quad (2)$$

Among them, Y is the concentration of Y chromosome, x_1 is gestational age, x_2 is BMI, x_3 to x_{11} is the technical covariate such as GC content and comparison rate, and ε is the random error term.

The model uses the least squares method to estimate the coefficient, and the objective function is the sum of squares of the minimized residuals:

$$\min S(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{11} \beta_j x_{ij} \right)^2 \quad (3)$$

Its matrix form is solved as:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (4)$$

The model was diagnosed with overall significance, variable significance and multiple collinearity by F-test, t-test and variance expansion factor (VIF).

2.2. BMI-based clustering detection time point recommendation

In order to formulate personalized detection plans for different BMI groups, the K-means clustering algorithm was used to divide pregnant women into three groups according to BMI. The algorithm is based on Euclidean distance minimization in-class error:

$$d(x, c_i) = \sqrt{\sum_{j=1}^{d} (x_j - c_{ij})^2} \quad (5)$$

$$J = \min \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} || x_i - c_k ||^2 \quad (6)$$

where c_k is the center of class k, and w_{ik} is the degree of membership.

After the grouping was determined, a linear relationship between Y concentration and gestational age was established within each group. Assuming that the effective detection threshold is $T = 0.04$, then the optimal detection of gestational age w^* can be solved by the following equation:

$$w^*(b) = \frac{T - \beta_0 - \beta_2 b}{\beta_1} \quad (7)$$

Where b is the representative value of BMI for this group (e.g., median).

2.3. Multivariate nonlinear modeling and error analysis

In order to further identify other influencing factors besides BMI, a regression model including age, height, weight, GC content, sequencing depth and other variables was established (Eq(2) extension). For the nonlinear relationship that cannot be captured by the linear model, the thin plate spline interpolation model is introduced, and its optimization goals are:

$$\min_f \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \iint \left(f_{xx}^2 + 2 f_{xy}^2 + f_{yy}^2 \right) dx dy \quad (8)$$

Where λ is the smoothing parameter, which controls the trade-off between fit complexity and smoothness.

The solution of the model can be expressed as a linear combination of affine terms and radial basis functions:

$$f(x, y) = \alpha_0 + \alpha_1 x + \alpha_2 y + \sum_{i=1}^{n} w_i \phi(\|(x, y) - (x_i, y_i)\|) \quad (9)$$

Where $\phi(r) = r^2 \log r$ is the basis function of the thin plate spline.

At the same time, the Spearman correlation analysis system was used to evaluate the interference degree of various technical variables (such as GC content and X chromosome concentration) on the measurement of Y concentration, and the main sources of error were identified.

3. Results and discussion

3.1. Step 1 results

Based on Spearman correlation analysis, we found that Y chromosome concentration was significantly positively correlated with gestational age $(\rho \approx 0.085,\ p < 0.01)$ and negatively correlated with BMI $(\rho \approx -0.146,\ p < 0.001)$, which is consistent with the physiological mechanism of fetal cell-free DNA enrichment with gestational age and high BMI leading to blood thinning. Multiple linear regression models further quantify this relationship.

Table 1. Estimation results of key variable coefficients of multiple linear regression model

Variables	Regression coefficient	Standard error	t-value	P-value
Intercept	-0.032	0.015	-2.13	0.033
Gestational age	0.018	0.002	9.00	<0.001
BMI	-0.002	0.0004	-5.00	<0.001
X chromosome concentration	0.124	0.041	3.02	0.003

The results of Table 1 show that the overall goodness of fit of the model is R^2=0.426, adjusted R^2=0.419, and the F-test is significant (F=60.015, p<0.001), indicating that the model can explain about 42.6% of the variation in Y concentration. For each week of gestational age, the Y concentration increased by an average of 0.018 units, and for each unit increase in BMI, the Y concentration decreased by an average of 0.002 units. X chromosome concentrations also showed a significant positive effect, suggesting possible technical cross-interference. The VIF values were all less than 2, indicating that there was no serious multicollinearity problem. The results provide a core parameter basis for subsequent grouping.

3.2. Step 2 results

Based on K-means cluster analysis, 1082 pregnant women were divided into three groups with significant differences in BMI. After the clustering center iteratively converges, the range of each group is clear. Substituting the median BMI of each group into the formula (7) and combining the gestational age-y concentration relationship obtained by linear fitting, we calculated the earliest reliable detection time point that met the concentration threshold (≥4%).

As shown in Table 2, pregnant women in the low BMI group can reach effective detection concentrations at about 12 weeks of gestation, while in the high BMI group, it needs to be delayed to nearly 19 weeks. This difference was due to lower mean Y concentrations in the high BMI group (0.0684 vs. 0.0823 in the low BMI group) and greater fluctuations

(standard deviation 0.0348). The results verified the significant effect of BMI on fetal DNA enrichment efficiency, and provided a direct basis for the clinical implementation of differentiated detection time window, which was helpful to reduce the early detection failure rate of pregnant women with high BMI.

Table 2. BMI grouping of pregnant women based on K-means clustering and optimal NIPT detection time

Grouping	BMI range	Sample size	Median BMI	Best time point in detection
Low BMI group	20.70 – 31.48	361	26.5	12 weeks and 2 days
Medium BMI group	31.53 – 35.34	360	33.4	16 weeks and 4 days
High BMI group	35.36 – 46.88	361	38.1	18 weeks and 5 days

3.3. Step 3 results

After including age, height, weight, and more sequencing technical variables, the multiple linear regression model had limited explanatory power (adjusted R^2=0.265). By introducing the combined effect of age and height on Y concentration by introducing the thin plate spline interpolation model, the goodness of fit of the model was significantly improved to R^2=0.6851, and the sum of squares of residuals (SSE) was 2.985×10^4.

Table 3. Correlation analysis of main features and Y chromosome concentration

Potential sources of error	Spearman correlation coefficient (ρ)	p-value	Direction of influence
X chromosome concentration	-0.218	<0.001	Negative
GC content	0.032	0.342	Not significant
Sequencing Depth (Raw Reads)	0.045	0.156	Not significant
Repeat reading ratio	-0.056	0.082	Weak negative direction

The fitting surface clearly reveals the nonlinear interaction between age and height on Y concentration: the middle age corresponds to a higher Y concentration platform with the height interval, while the concentration decreases when the age is older or the height is extreme. Error analysis in Table 3 shows that there is a significant negative correlation between X chromosome concentration and Y concentration (ρ=-0.218), which is the primary source of technical error, which may be due to the cross-interference of X/Y chromosome alignment during sequencing. GC content and sequencing depth had no significant effect in this dataset, but the proportion of repeated reads showed potential weak interference. These findings suggest that the correction algorithm for X chromosome signal should be optimized in the NIPT wet experiment and bioinformatics analysis process to improve the accuracy of Y chromosome concentration measurement.

This study analyzed the influence mechanism of Y chromosome concentration in NIPT detection and the optimal time point selection strategy through a three-step modeling system. The first step confirmed that gestational age and BMI are core factors affecting Y concentration, consistent with existing research; the second step proposed a time point recommendation strategy based on BMI grouping, demonstrating strong clinical operability; the third step further revealed the non-linear effects of age, height, and X chromosome interference, providing a basis for error control. The model performed well in terms of fitting accuracy and interpretability, but still has room for improvement in addressing multicollinearity and error correction.

4. Conclusion

In this study, a system analysis framework based on multivariate statistical model and machine learning algorithm is constructed around the time point selection and error analysis of NIPT detection. Firstly, through Spearman correlation analysis and multiple linear regression, the significant effects of gestational age and BMI on Y chromosome concentration were clarified: the increase of gestational age significantly increased the concentration of Y, while the high BMI led to a decrease in concentration and an increase in volatility. On this basis, the K-means clustering algorithm was used to divide pregnant women into low, medium and high groups according to BMI, and the optimal detection time points of each group were 12 weeks and 2 days, 16 weeks 4 days and 18 weeks and 5 days, respectively, which provided clear grouping detection suggestions for clinical practice.

Furthermore, the thin plate spline interpolation model was introduced to successfully capture the nonlinear interaction effect between age, height and Y concentration, and the goodness of fit of the model was increased to 68.51%. Through Spearman correlation analysis, the identification of X chromosome concentration was the most important source of technical error in NIPT detection, and factors such as GC content and sequencing depth also had a certain impact on the detection results. The multivariate modeling framework proposed in this study not only improves the accuracy and individualization of NIPT point-in-time recommendation, but also provides a methodological reference for the analysis of nonlinear relationships in complex medical data.

However, there are still some limitations in this study: first, some variables in the linear model have multicollinearity problems, which may affect the stability of coefficient estimation; Secondly, although the error analysis identifies the main sources, the real-time correction mechanism is not established. In addition, the model relies on existing sample distributions and needs to be further validated when extrapolated to a wider population. In future research, more complex models such as regression and deep learning can be considered to further optimize variable selection and nonlinear fitting capabilities, and a dynamic correction system can be constructed based on real-time monitoring data to optimize the whole process of NIPT detection and improve intelligence.

Acknowledgements

This paper was supported by my teacher Professor Wang.

References

[1] Lo, Y. M., et al. (2021). Noninvasive prenatal testing: a review of the current state of the science. Annual Review of Genomics and Human Genetics, 22, 1-25.

[2] Chen, L., et al. (2021). Fetal fraction estimation in maternal plasma and its association with gestational age and maternal weight. Prenatal Diagnosis, 41(4), 456-463.

[3] Liu, Y., et al. (2022). Impact of maternal BMI on fetal DNA fraction in noninvasive prenatal testing. Journal of Translational Medicine, 20(1), 1-10.

[4] Sun, K., et al. (2020). Sequencing depth and coverage uniformity impacts on NIPT performance. Clinical Chemistry, 66(8), 1072-1081.

[5] Wu, H., et al. (2021). GC bias correction in cell-free DNA sequencing for aneuploidy detection. Bioinformatics, 37(16), 2389-2395.

[6] Liao, C., et al. (2022). Biological determinants of fetal fraction variation in maternal plasma. Human Genetics, 141(5), 987-996.

[7] Zhang, H., et al. (2021). Optimization of NIPT timing through multivariate regression modeling. Clinical Genetics, 99(3), 321-329.

[8] Chen, S., et al. (2022). Longitudinal study of fetal DNA dynamics across gestation. Prenatal Diagnosis, 42(1), 45-52.

[9] Liu, J., et al. (2023). Machine learning for predicting fetal fraction from maternal characteristics. IEEE Journal of Biomedical and Health Informatics, 27(1), 210-218.

[10] Wang, Q., et al. (2023). Machine learning-based prediction of fetal aneuploidy using cell-free DNA sequencing data. Bioinformatics, 39(5), 1-9.

[11] Xu, R., et al. (2022). Nonlinear relationship between maternal characteristics and fetal DNA concentration. PLOS ONE, 17(6), e0268888.

[12] Li, M., et al. (2023). A clustering-based approach for personalized NIPT scheduling. Medical & Biological Engineering & Computing, 61(4), 987-996.

[13] Zhou, Y., et al. (2023). Multivariate spline interpolation for biomedical data analysis. Computational Statistics & Data Analysis, 179, 107-120.

[14] Guo, L., et al. (2021). Error source analysis in noninvasive prenatal testing: a systematic review. Clinical Chemistry and Laboratory Medicine, 59(8), 1321-1330.

[15] Pan, Y., et al. (2023). Spearman correlation analysis in nonparametric medical data modeling. Journal of Applied Statistics, 50(5), 1123-1135.