Model Construction and Inference of Fetal Y-Chromosome Fraction in Male Pregnancies for NIPT

Xiaodong Zhao[1], Jingfang Chu[1], Yueyang Li[1], Mingming Gong[2,*]
[1]School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, Henan, China
[2]iFLYTEK Co., Ltd., Hefei, Anhui, China
**Corresponding Author*

Abstract: Non-invasive prenatal testing NIPT) has become an important method for screening fetal chromosomal abnormalities; however, determining optimal testing time and accounting for individual maternal differences remain significant challenges in clinical practice. This study investigates the relationship between fetal Y-chromosome fraction and maternal characteristics and proposes a mathematical modeling framework for optimizing individualized NIPT testing strategies. Based on clinical NIPT data, Pearson and Spearman correlation analyses were performed to evaluate the relationships among gestational age, maternal body mass index BMI), and fetal Y-chromosome fraction. Multiple regression models, including linear, polynomial, exponential, and sigmoid models, were constructed and compared to identify the most suitable predictive model. The results indicate that gestational age is positively correlated with fetal Y-chromosome fraction, whereas maternal BMI shows a negative correlation. Among the tested models, the cubic polynomial regression achieved the best fitting performance, and piecewise optimization further improved the model with a maximum coefficient of determination R^2) of 0.4556. These findings provide a quantitative basis for determining individualized NIPT testing time and improving the reliability of prenatal screening.

Keywords: Non-Invasive Prenatal Testing (NIPT); Fetal Y-Chromosome Fraction; Piecewise Nonlinear Model; Testing Time Optimization; Model Validation

1. Introduction

With the rapid development of non-invasive prenatal testing (NIPT) technology [1], prenatal screening has become safer and more accurate for detecting fetal chromosomal abnormalities [2]. However, several challenges remain in clinical practice, including differences among pregnant women, the complexity of selecting appropriate testing time points, and ensuring high diagnostic accuracy.

Previous studies indicate that current NIPT protocols often do not fully consider maternal factors such as age, body mass index BMI), and pregnancy conditions. Ignoring these factors may lead to inaccurate test results, which can affect fetal health assessment and potentially reduce the available window for medical intervention. Furthermore, with the increasing demand for NIPT and limited clinical resources, improper selection of testing time may result in sequencing failures or repeated tests, thereby reducing the reliability and convenience of prenatal screening.

In clinical practice, the fetal Y-chromosome fraction in maternal plasma is significantly influenced by both gestational age and maternal BMI. However, many clinical protocols rely on fixed gestational windows or empirical grouping methods, which may not adequately reflect individual differences. As a result, it is difficult to balance testing success rates and diagnostic timeliness.

Recent studies have attempted to optimize NIPT testing timing through various approaches [3], including BMI stratification and gestational stage selection[4]. In addition, machine learning techniques such as regression analysis and neural networks have gradually been introduced to improve predictive accuracy. Nevertheless, many existing models still adopt general frameworks that may lead to premature testing or misinterpretation of results. Therefore, more suitable modeling approaches are required to improve individualized decision-making.

To address these issues, this study focuses on two main objectives. First, it investigates the relationships between fetal Y-chromosome fraction and maternal characteristics, particularly gestational age and BMI, and constructs corresponding mathematical models. Second, based on these models, an individualized prediction method for optimal NIPT testing time is proposed. Through data-driven analysis and modeling, the study aims to reveal the underlying relationships among key factors and provide a more precise decision-support framework for clinical prenatal testing.

2. Research Methods

The main objective of this study is to establish a data-driven mathematical model to support individualized decision-making in non-invasive prenatal testing (NIPT). The overall research framework is illustrated in Figure 1.

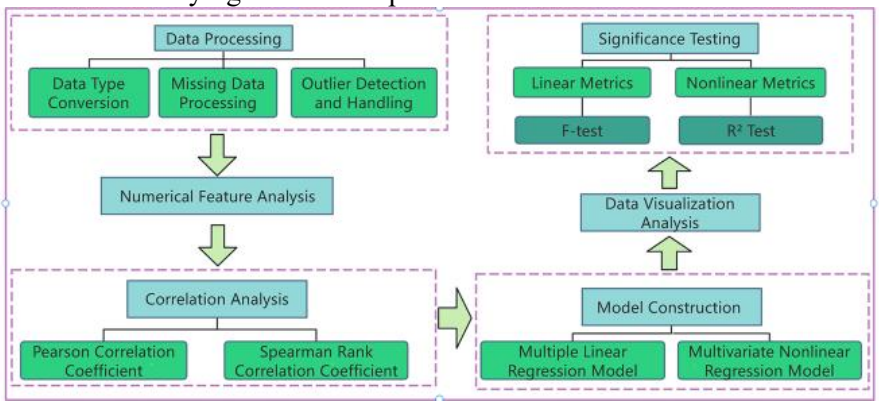

Figure 1. NIPT Individualized Decision-Making Framework

First, data preprocessing was performed. Missing values were examined and supplemented where necessary, irrelevant features were removed, and the data format was standardized. All variables were converted into numerical form to ensure compatibility with subsequent modeling procedures.

Subsequently, exploratory data analysis was conducted to investigate the distribution characteristics of the dataset and the relationships among variables. Visualization techniques were used to identify potential patterns and underlying trends in the data. Pearson and Spearman correlation coefficients were then calculated to evaluate the relationships between fetal Y-chromosome fraction and the explanatory variables. The results of these correlation analyses provided guidance for selecting appropriate model structures.

Finally, both linear and nonlinear regression models were constructed to fit the data, and statistical significance tests were conducted to evaluate the reliability and effectiveness of the models.

2.1 Data Processing

2.1.1 Dataset and preprocessing

The dataset used in this study was obtained from real clinical NIPT testing records collected from a regional hospital. Prior to model construction, the dataset was carefully examined to ensure its reliability and completeness.

The results of the data inspection indicate that the dataset has high integrity. Key variables including maternal BMI, gestational age at the time of testing, and fetal Y-chromosome fraction contain no missing values; therefore, additional data imputation was not required.

Previous studies have shown that fetal cell-free DNA concentrations remain relatively stable during 10–25 weeks of gestation, allowing reliable detection of fetal abnormalities. In addition, maintaining the GC content within the range of 0.4–0.6 (40%–60%) helps ensure sequencing quality and prevents data distortion caused by abnormal GC composition[5].

Based on these conditions, samples with a high risk of sequencing failure were removed, and a valid dataset was constructed for subsequent analysis.

To facilitate the analysis, the key variables used in this study are summarized in Table 1.

Table 1. Symbol Definitions

Symbol	Description	Unit
BMI	Maternal body mass index	kg/m^2
yz	Gestational age at testing	week
Y	Y-chromosome fraction	ng/μL
F	F-test statistic	/
P	Probability value	/

2.1.2 Data transformation

To ensure consistency in data representation, the expression of gestational age was standardized. All gestational age records were converted into decimal week values.

The conversion formula is defined as follows:

$$\text{Gestational Age}=\text{Whole Weeks}+\frac{\text{Additional days}}{7} \quad (1)$$

After data preprocessing, three variables were extracted from the male fetal dataset for analysis:

maternal BMI

gestational age at testing

fetal Y-chromosome fraction

The statistical characteristics of these variables are summarized in Table 2.

Table 2. Statistical Results of Key Variables

Variable	BMI	Gestational Age	Y-Chromosome Fraction
Sample	1082.00	1082.00	1082.00
Maximun	46.8750	29.0000	0.2342
Minimum	20.7031	11.0000	0.0100
Median	31.8116	16.0000	0.0751
Mean	32.2888	16.8457	0.0772
Standard deviation	2.9724	4.0763	0.0335

These statistical indicators [6] provide a basic overview of the dataset and serve as the foundation for subsequent correlation analysis and model construction.

2.2 Correlation Analysis of Factors Affecting the Y-Chromosome Fraction

Before constructing the regression models, it is necessary to analyze the correlation relationships among variables. This step helps determine both the direction and strength of the effects of gestational age and maternal BMI on the fetal Y-chromosome fraction, thereby providing a theoretical basis for selecting an appropriate model structure.

In this study, two commonly used correlation coefficients were employed: the Pearson correlation coefficient [7] and the Spearman rank correlation coefficient [8].

2.2.1 Pearson correlation coefficient

The Pearson correlation coefficient is used to measure the degree of linear correlation between two continuous variables and is denoted by r. The calculation formula is

$$r_{X,Y}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\cdot\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (2)$$

where X and Y represent two variables (in this study, BMI or gestational age and the Y-chromosome fraction), and n represents the sample size.

The value of r ranges from −1 to 1:

r>0 indicates a positive correlation

r<0 indicates a negative correlation

a value closer to 1 or −1 indicates stronger correlation.

The Pearson coefficient assumes that variables follow an approximately normal distribution and that their relationship is linear; therefore, it is mainly suitable for analyzing linear relationships.

2.2.2 Spearman rank correlation coefficient

When the relationship between variables may be nonlinear but monotonic, the Spearman rank correlation coefficient, denoted byρ, can be used. It is calculated based on the rank of variables rather than their actual values.

$$\rho=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)} \quad (3)$$

where di represents the difference between the ranks of the two variables for the i-th sample.

Similar to the Pearson coefficient, the value ofρalso ranges from −1 to 1, and the sign and magnitude have similar interpretations.

Compared with the Pearson coefficient, the Spearman coefficient does not require a normal distribution assumption and is less sensitive to outliers. Therefore, it is more suitable for datasets that may contain measurement noise or fluctuations.

2.2.3 Correlation analysis results

The male fetal dataset, including maternal BMI, gestational age, and fetal Y-chromosome fraction, was analyzed. The relationship between gestational age and Y-chromosome fraction is shown in Figure 2.

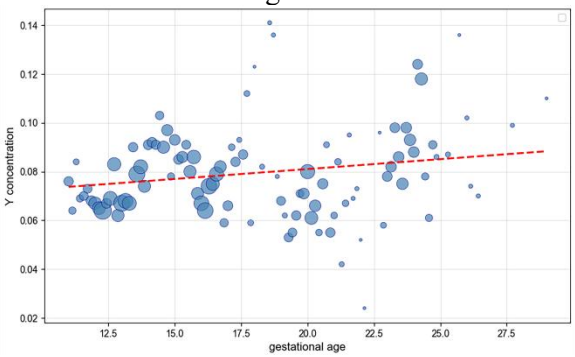

Figure 2. Correlation between Gestational Age and Y Concentration

Figure 2 indicates that the mean Y-chromosome fraction generally increases with gestational age, showing a significant positive correlation. Although fluctuations exist within the gestational period of 10–25 weeks, the overall

concentration remains within a stable and detectable range. This provides an intuitive basis for determining appropriate NIPT testing time points.

The relationship between maternal BMI and Y-chromosome fraction is illustrated in Figure 3.

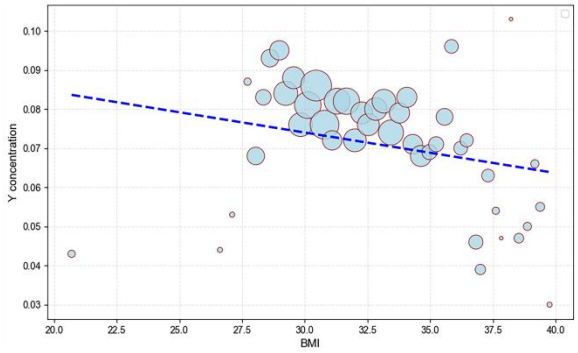

Figure 3. Correlation between BMI and Y Concentration

The results show a clear negative correlation between maternal BMI and the mean fetal Y-chromosome fraction. As BMI increases, the average concentration gradually decreases. However, within the same BMI interval, individual differences remain large, resulting in a relatively dispersed data distribution. This suggests that BMI alone is insufficient to accurately predict the Y-chromosome fraction.

This negative correlation provides important evidence for constructing a predictive model that incorporates both BMI and gestational age.

The calculated correlation coefficients are summarized in Figure 4.

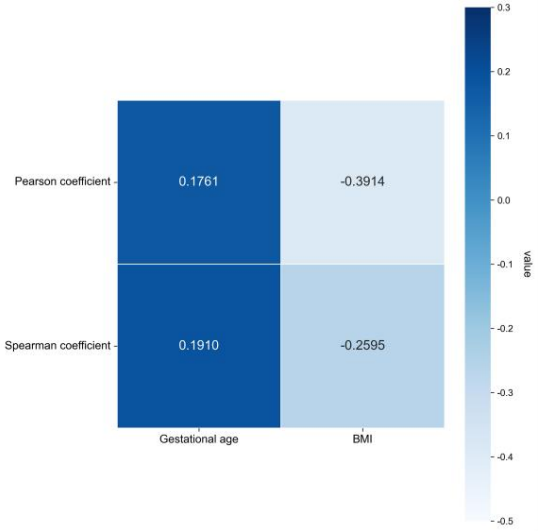

Figure 4. Calculation Results of Correlation Coefficient

From the correlation heatmap, gestational age shows a positive correlation with fetal Y-chromosome fraction (Pearson coefficient 0.1761, Spearman coefficient 0.1910). In contrast, maternal BMI shows a negative correlation with Y-chromosome fraction (Pearson coefficient -0.3914, Spearman coefficient -0.2595).

These results indicate that the Y-chromosome fraction tends to increase with gestational age, while it decreases as maternal BMI increases. Moreover, BMI exhibits a relatively stronger influence on the Y-chromosome fraction, providing a clear basis for constructing a multivariable predictive model.

3. Model Construction

Based on the available dataset, the fetal Y-chromosome fraction was selected as the dependent variable, while gestational age (yz) and maternal BMI were used as the independent variables. To ensure the reliability of the constructed model, statistical tests including the F-test, t-test [9], and heteroscedasticity test were conducted. The results indicate that the proposed models are statistically significant and exhibit good robustness.

To comprehensively explore the relationship between fetal Y-chromosome fraction and maternal characteristics, several mathematical models with different functional forms were established and compared. These models range from simple linear models to more complex nonlinear regression models.

3.1 Quadratic Polynomial Model

The quadratic polynomial regression model incorporates squared terms and interaction terms between the independent variables. The model is expressed as

$$Y=\beta 1\cdot BMI2+\beta 2\cdot yz2+\beta 3\cdot BMI\cdot yz+\beta 4\cdot BMI+\beta 5\cdot yz+\beta 6 \quad (4)$$

where:

Y represents the fetal Y-chromosome fraction, BMI represents maternal body mass index, yz represents gestational age at testing, $\beta 1, \beta 2, \ldots, \beta 6$ are regression coefficients.

This model captures basic nonlinear relationships between the independent variables and the dependent variable.

3.2 Exponential Model

The exponential model assumes that the Y-chromosome fraction grows exponentially with respect to the independent variables. The mathematical form of the model is

$$Y=\alpha\cdot e^{(\beta_1\cdot BMI+\beta_2\cdot yz)}+\gamma \quad (5)$$

where:

α_1, β_1, β_2, and γ are model parameters, e denotes the base of the natural logarithm.

This model is suitable for describing situations in which the dependent variable exhibits exponential growth patterns.

$$Y=\beta_1\cdot BMI^3+\beta_2\cdot yz^3+\beta_3\cdot BMI^2\cdot yz+\beta_4\cdot BMI\cdot yz^2+\beta_5\cdot BMI^2+\beta_6\cdot yz^2+\beta_7\cdot BMI\cdot yz+\beta_8\cdot BMI+\beta_9\cdot yz+\beta_{10} \quad (6)$$

Compared with lower-order models, the cubic polynomial model provides stronger nonlinear fitting capability and can capture more complex interactions between maternal BMI and gestational age.

3.4 Sigmoid Model

The Sigmoid model is based on a logistic function and is commonly used to describe biological growth processes and saturation effects. The model is expressed as

$$Y=\frac{\alpha}{1+e^{-(\beta_1\cdot BMI+\beta_2\cdot yz+\beta_3)}}+\delta \quad (7)$$

where:

α_1, β_1, β_2, β_3, and δ are model parameters.

This model is particularly useful for describing situations where the response variable approaches an upper limit.

3.5 Model Evaluation Criteria

To evaluate the fitting performance of different models, two main statistical indicators were used: the coefficient of determination (R^2) and the F-statistic.

The value of R^2ranges from 0 to 1. A value close to 1 indicates that the model explains a large proportion of the variance in the dependent variable, while a value close to 0 suggests weak explanatory power.

The F-statistic is used to test the overall significance of the regression model. Under the null hypothesis H0, the regression model has no explanatory power, and the regression sum of squares (SSreg) should be relatively small. If the model is effective, SSregwill be much larger than the residual sum of squares (SSres), resulting in a large F value.

Therefore, a model with higher R^2 values and larger F-statistics (with small p-values) is considered to have stronger explanatory ability and better statistical significance.

By combining these evaluation metrics, the performance of different regression models can be systematically compared and the most suitable model for describing the relationship between maternal characteristics and fetal Y-chromosome fraction can be identified.

3.3 Cubic Polynomial Model

To further capture complex nonlinear relationships, a cubic polynomial regression model was constructed. This model includes all third-order polynomial combinations of the independent variables and is defined as

4. Experimental Results

Using multiple mathematical models, the relationship between fetal Y-chromosome fraction and maternal characteristics such as gestational age and BMI was fitted and analyzed. The fitting results of different models are summarized in Table 3.

Table 3. Model Fitting Performance

Mathematical Model	Quadratic Polynomial	Exponential Model	Cubic Polynomial	Sigmoid Model
R^2	0.0494	-0.0354	0.1182	0.0981
F value	5.5015	-1.7785	7.8322	6.6782
P value	$5.5015e^{-5}$	0.9719	$7.0875e^{-11}$	$4.0978e^{-7}$

As shown in Table 3, the cubic polynomial model demonstrates better fitting performance compared with other models. However, the overall goodness-of-fit remains relatively low when applied to the entire dataset, indicating limited explanatory capability.

To further improve the model performance, this study optimized the cubic polynomial model using a piecewise fitting strategy [10]. Specifically, based on both the data distribution and clinical considerations, the ranges of gestational age and BMI were each divided into three intervals. Separate cubic polynomial models were then fitted within each interval.

The results of the piecewise fitting are shown in Table 4.

Table 4. Piecewise Fitting Results of the Cubic Polynomial Model

Interval	R^2	F value	P value
(0,0)	0.3011	0.7744	0.6403
(0,1)	0.3005	2.0045	0.0629
(0,2)	0.3144	2.1915	0.0416
(1,0)	0.1175	0.6951	0.7097
(1,1)	0.1310	0.8378	0.5852
(1,2)	0.3302	2.7388	0.0111
(2,0)	0.2799	1.5977	0.1521
(2,1)	0.4556	5.1136	0.1171
(2,2)	0.3190	3.1233	0.0038

After applying the piecewise fitting approach, the performance of the cubic polynomial model improved significantly. The highest coefficient of determination reached 0.4556 in interval (2, 1), while the maximum F value increased to 5.1136.

The three-dimensional scatter plots of the fitted piecewise cubic polynomial models are shown in Figure 5, where different colors represent different fitted model equations.

Overall, the results indicate that the variation patterns of fetal Y-chromosome fraction differ significantly across BMI groups. In general, the Y-chromosome fraction is relatively low during early pregnancy and gradually increases as gestational age progresses. However, this increasing trend appears earlier in the low-BMI group, while it is delayed in the high-BMI group, and in some gestational periods the concentration remains relatively low.

Further regression analysis suggests that higher BMI leads to a later gestational age at which the Y-chromosome fraction reaches the 4% threshold. This finding indicates that maternal body weight plays a significant role in determining the optimal testing time for NIPT.

Although certain fluctuations exist in the scatter distribution across different groups, which may be influenced by measurement errors, experimental conditions, or individual biological differences, the overall trend remains stable. These findings support the conclusion that BMI is an important factor affecting the timing of fetal cell-free DNA detection.

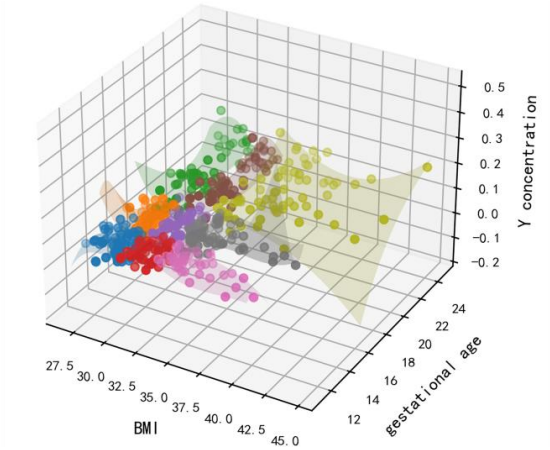

Figure 5. Piecewise Fitting of Three-Dimensional Scatter Plot Using Cubic Polynomial Model

5. Conclusion

This study addresses the limitations of conventional NIPT clinical testing strategies that often ignore individual differences among pregnant women. Such limitations may lead to inappropriate testing timing or diagnostic inaccuracies. To overcome these issues, a modeling framework based on piecewise nonlinear regression analysis was established to quantitatively analyze the relationship between maternal characteristics and fetal Y-chromosome fraction.

Experimental results demonstrate that gestational age shows a weak positive correlation with fetal Y-chromosome fraction (Pearson coefficient 0.1761, Spearman coefficient 0.1910), while maternal BMI exhibits a moderate negative correlation (Pearson coefficient −0.3914, Spearman coefficient −0.2595). Among the studied variables, BMI plays a key role in determining the timing at which the Y-chromosome fraction reaches clinically detectable levels.

Overall, this study proposes an integrated analytical framework that combines correlation analysis, regression modeling, and piecewise optimization. The framework provides a quantitative basis for individualized decision-making in NIPT testing and helps reduce the risk of detection failure while extending the available clinical intervention window.

In addition, the proposed modeling strategy demonstrates good adaptability and practical applicability. The framework may be extended to other prenatal screening scenarios or medical detection tasks involving low-concentration nucleic acids, providing a general methodological reference for solving personalized. medical decision-making problems.

Future work will focus on further improving the model by incorporating additional factors such as environmental noise and data uncertainty to enhance robustness and generalization ability. Moreover, integrating the model with clinical diagnostic workflows and developing visualization-based decision support tools may facilitate the practical application of this research in real clinical environments.

References

[1] Liang P. Research progress of non-invasive prenatal testing using fetal cell-free DNA in maternal peripheral blood. Chinese Medical Science Database, 2023.

[2] Yang C G, Xu L J, Zeng Y F. Influence of NIPT-plus combined with fetal cell-free DNA enrichment technology on the early diagnostic accuracy of chromosomal microdeletion and microduplication syndromes. Chinese Journal of Eugenics and Genetics, 2025, 33(10): 2220-2227.

[3] Ma F, Liu Y. Construction and solution of the association model for fetal Y-chromosome concentration in non-invasive prenatal testing. Frontiers in Medical Science Research, 2025, 7(6).

[4] Li T. Regression analysis-based investigation of factors influencing male fetal Y chromosome concentration and stratified optimization of optimal timing for non-invasive prenatal testing. Journal of Pharmaceutical and Medical Research, 2025, 7(3).

[5] Jia Q, Wu H T, Zhou X J, et al. The regulatory role of GC-rich DNA fragments in ultra-high gene expression in mammalian cells. Science China Life Sciences, 2010, 40(2): 159–165.

[6] van Beek D M, Straver R, Weiss M M, et al. Comparing methods for fetal fraction determination and quality control of NIPT samples. Prenatal diagnosis, 2017, 37(8): 769-773.

[7] Mo J T. Research on group consensus models of non-reciprocal judgment matrices based on the Pearson correlation coefficient. Guangxi University, 2025.

[8] Li Y, Liu Y T, Feng L W. Nonlinear dynamic process feature extraction and fault detection based on Spearman correlation analysis. Journal of Shandong University of Science and Technology (Natural Science Edition), 2023, 42(2): 98–107.

[9] Jin, T. L., & Zhang, B. Q. (2009). Further Discussion on the Relationship between t-test and F-test in Regression Analysis. Statistics & Decision, (21), 7–9.

[10] Tian C F, Li J J, Weng G J, et al. Improved local extremum-based piecewise polynomial fitting algorithm for accurate correction of Raman spectral baseline. Spectroscopy and Spectral Analysis, 2024, 44(4): 1073–1080.

