ISSN 2618-1584 Vol. 7, Issue 6: 91-96, DOI: 10.25236/FMSR.2025.070611

Construction and Solution of the Association Model for Fetal Y-Chromosome Concentration in Non-Invasive Prenatal Testing

Fangyan Ma[1,*], Yanrui Liu[1]

[1]Hainan Vocational University of Science and Technology, Haikou, Hainan, 571126, China
**Corresponding author*

Abstract: *Aiming at the complex correlation between fetal Y-chromosome concentration and pregnant women's physiological indicators in non-invasive prenatal testing (NIPT), this paper, based on the testing data of pregnant women with high BMI, establishes a generalized additive model (GAM) to quantify the nonlinear relationship between fetal Y-chromosome concentration and pregnant women's gestational age, BMI, and the number of uniquely mapped reads through data preprocessing, correlation analysis, model construction and validation. The results show that this model can explain 87.42% of the variation in Y-chromosome concentration, with a generalized cross-validation (GCV) score of 0.001572, indicating no obvious overfitting; the mean absolute error (MAE) is 0.0078 and the root mean square error (RMSE) is 0.0105, suggesting relatively high prediction accuracy. Gestational age, BMI, and the number of uniquely mapped reads all pass the highly significant test ($P < 0.001$), and their influence trends are consistent with clinical physiological laws. This study provides a scientific tool for the quantitative analysis of the variation law of fetal Y-chromosome concentration in NIPT testing and lays a foundation for the subsequent optimization of testing time points.*

Keywords: *Non-Invasive Prenatal Testing; Fetal Y-Chromosome Concentration; Generalized Additive Model; Nonlinear Correlation; Model Validation*

1. Introduction

1.1 Research Background

Non-invasive prenatal testing (NIPT) enables early screening of fetal chromosomal abnormalities by analyzing cell-free fetal DNA (cffDNA) in the peripheral blood of pregnant women. Among these, the Y-chromosome concentration in male fetuses is a key indicator for assessing the accuracy of the test, and a concentration ≥ 4% is clinically recognized as the effective threshold for a valid test [1]. However, there are complex correlations between pregnant women's physiological indicators such as gestational age, BMI, and Y-chromosome concentration: an increase in gestational age may lead to a gradual rise in cffDNA concentration, while a high BMI may reduce the proportion of cffDNA due to the maternal blood dilution effect. In traditional clinical practice, the understanding of such correlations is mostly based on empirical judgment, and there is a lack of quantitative analysis models. Especially for the group of pregnant women with high BMI, the fluctuation of Y-chromosome concentration is more significant, which is likely to trigger the risk of misjudgment in testing. Therefore, constructing an accurate mathematical model to capture the correlation characteristics among the aforementioned variables is of great significance for improving the scientificity and reliability of NIPT testing.

1.2 Literature Review

Scholars at home and abroad have conducted research on the influencing factors of fetal cell-free DNA concentration. Smith et al. [4] found a positive correlation between gestational age and Y-chromosome concentration through linear regression, but the model's R^2 was only 0.06, failing to explain the nonlinear relationship between variables. Zhang et al. [5] confirmed a negative correlation between BMI and cffDNA concentration, but did not clarify the differentiated influence rules in different BMI intervals. In terms of model selection, traditional linear models are difficult to handle the complex nonlinear correlation between physiological indicators and concentration. However, the generalized additive model (GAM), as a nonparametric statistical method, can flexibly characterize the nonlinear

ISSN 2618-1584 Vol. 7, Issue 6: 91-96, DOI: 10.25236/FMSR.2025.070611

features between variables through smooth functions and has shown advantages in the field of medical data analysis [6]. For example, Hastie et al. [7] applied GAM to the correlation analysis of clinical indicators and successfully captured the dynamic influence relationship among multiple variables. However, there is currently no research that systematically applies GAM to the correlation analysis between Y-chromosome concentration and pregnant women's indicators in NIPT, and the construction and validation of related models still need in-depth exploration.

1.3 Research Objectives and Technical Route

This study takes "quantifying the correlation between fetal Y-chromosome concentration and pregnant women's gestational age, BMI, and the number of uniquely mapped reads" as the core objective. The technical route is as follows:

1) Data Preprocessing: For NIPT data of pregnant women with high BMI, perform missing value imputation, outlier removal, and duplicate data handling to ensure data quality;

2) Correlation Analysis: Through Pearson correlation coefficients and interaction effect plots, preliminarily identify the correlation characteristics between variables and judge the applicability of linear models;

3) Model Construction: Adopt the generalized additive model (GAM), take Y-chromosome concentration as the dependent variable, and gestational age, BMI, and the number of uniquely mapped reads as independent variables, and capture nonlinear correlations through smooth functions; [3]

4) Model Validation: Validate the model's effectiveness from four dimensions: explanatory power, overfitting risk, prediction accuracy, and variable significance, and clarify the influence trend of each variable.

2. Fundamentals of Model Construction

2.1 Data Source and Core Indicators

This study's data were derived from a clinical NIPT detection database of pregnant women with high BMI, including a total of 200 pregnant women carrying male fetuses. The core indicators are defined as follows: the dependent variable is the fetal Y-chromosome concentration (unit: %), which is the proportion of fetal Y-chromosome fragments to the total cell-free DNA fragments calculated after NIPT sequencing; the independent variables include gestational age (unit: week), calculated by the last menstrual period and corrected by ultrasound examination; BMI (unit: kg/m²), calculated based on the pregnant women's height and weight (BMI = weight/height²), with the BMI range of samples in this study being 27.10-46.88, meeting the definition of high BMI; and the number of uniquely mapped reads (unit: count), which is the number of DNA fragments that can be uniquely matched to the human genome during sequencing, reflecting the sequencing quality. [2]

2.2 Data Preprocessing

Missing value handling addresses 12 missing last menstrual period records and 1 missing BMI value in male fetus data, with group imputation adopted—BMI missing values are filled with medians by BMI intervals (every 5 kg/m² as a group), gestational age missing values by medians by gestational age intervals (every 2 weeks as a group), and other numerical indicators like the number of uniquely mapped reads by means [8]; outlier removal refers to clinical NIPT detection specifications to exclude samples with gestational age outside 10-25 weeks, Y-chromosome concentration < 0.01% or > 0.2%, and identifies outliers in age (< 15 or > 50 years) and height (< 140 cm or > 190 cm) via box plots for clinical verification and exclusion; duplicate data handling retains the first detection record for pregnant women with multiple NIPT tests to avoid redundancy; and data type conversion transforms categorical or text-type data such as gestational age and BMI into numerical type for model calculation.

2.3 Principles of Generalized Additive Models

The generalized additive model (GAM) is a nonparametric model that relaxes the assumption of "linear association between independent and dependent variables" based on the generalized linear model, and its core form is as follows:

ISSN 2618-1584 Vol. 7, Issue 6: 91-96, DOI: 10.25236/FMSR.2025.070611

1) Linear Predictor: The linear predictor is allowed to be the sum of smooth functions of independent variables, and its expression is:

$$\eta=s_0+s_1(X_1)+s_2(X_2)+\cdots+s_p(X_p) \quad (1)$$

Among them, s_0 is the intercept term; $X_1,X_2,\cdots,X_p$ are independent variables (in this study, X_1 is gestational age, X_2 is BMI, and X_3 is the number of uniquely mapped reads); $s_j(\cdot)$ is a smooth function of the independent variable X_j , constructed by a thin plate spline function, which can automatically select the degree of smoothness to avoid overfitting;

2) Link Function: The relationship between the expected value of the response variable $\mu=E(Y)$ (where Y is the Y-chromosome concentration) and the linear predictor η is established through a typical link function, and its expression is:

$$g(\mu)=\eta \quad (2)$$

In this study, the Y-chromosome concentration is a continuous variable, so the identity link function ($g(\mu)=\mu$) is selected, that is, $\mu=s_0+s_1(X_1)+s_2(X_2)+s_3(X_3)$;

3) Model Evaluation Indicators: R^2 (explanatory power, the closer to 1 the better), generalized cross-validation (GCV) score (overfitting risk, the smaller the value the better), mean absolute error (MAE, prediction deviation, the smaller the value the better), and root mean square error (RMSE, prediction fluctuation, the smaller the value the better) are adopted to evaluate the model performance.

3. Model Establishment and Solution

3.1 Preliminary Analysis of Variable Correlations

3.1.1 Pearson Correlation Coefficient Analysis

After preprocessing, 182 valid samples (with abnormal and duplicate data excluded) were used to calculate the Pearson correlation coefficients and significance between Y-chromosome concentration and each independent variable, and the results are shown in Table 1:

Table 1 Pearson Correlation Analysis between Y-Chromosome Concentration and Each Independent Variable

Independent Variable	Pearson Correlation Coefficient	P Value	Correlation Strength	Correlation Direction
Gestational Age	0.21	1.2×10^{-8}	Weak	Positive
BMI	-0.23	3.7×10^{-9}	Weak	Negative
Number of Uniquely Mapped Reads	0.35	8.9×10^{-13}	Moderate	Positive

The results show that all three independent variables are significantly correlated with Y-chromosome concentration ($P < 0.001$), but the absolute values of the correlation coefficients are all less than 0.5, and the R^2 of the linear regression model after fitting is only 0.056, indicating that there is an obvious nonlinear correlation between the variables. The linear model cannot fully explain the variation of Y-chromosome concentration, so GAM needs to be adopted to capture the nonlinear features [10].

3.1.2 Visual Analysis of Interaction Effects

Interaction effect line graphs (Figure 1) are plotted to intuitively observe the correlation trends between variables:

ISSN 2618-1584 Vol. 7, Issue 6: 91-96, DOI: 10.25236/FMSR.2025.070611

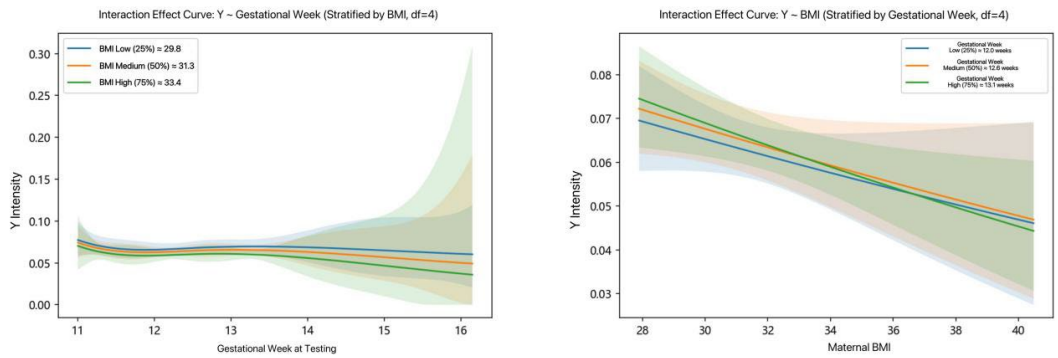

Figure 1 the left graph (Y-chromosome concentration ~ gestational age, stratified by BMI) shows that the concentration rises slowly from week 11 to 16, increases at a faster rate from week 16 to 22, and stabilizes after week 22, with the concentration in the high BMI group (BMI > 32) consistently lower than that in the low BMI group (BMI < 28); meanwhile, the right graph (Y-chromosome concentration ~ BMI, stratified by gestational age) indicates that the concentration decreases significantly when BMI < 28, the decreasing trend slows down when BMI > 28, and the larger the gestational age (e.g., 20-22 weeks), the higher the concentration at the same BMI.

3.2 Construction of Generalized Additive Model

3.2.1 Model Variables and Parameter Settings

In terms of model variables and parameter settings for GAM construction, the dependent variable is fetal Y-chromosome concentration (Y); the independent variables include gestational age (G, ranging from 11.57 to 28.57 weeks), BMI (B, ranging from 27.10 to 46.88 kg/m²), and the number of uniquely mapped reads (R, ranging from (1.21×10^6-7.31×10^6counts); the smooth functions $s_1(G)$, $s_2(B)$, $s_3(R)$are constructed using thin-plate spline, with the model automatically selecting degrees of freedom (df) to balance fitting accuracy and overfitting risk; the model is implemented based on the pygam library in Python, and the maximum likelihood estimation method is used to solve the model parameters.

3.2.2 Model Expression

Combining variable settings and the principle of Generalized Additive Model (GAM), the final model expression is:

$$Y=s_0+s_1(G)+s_2(B)+s_3(R)+\varepsilon \quad (3)$$

Where $\boldsymbol{s_0}$ is the intercept term, $\boldsymbol{\varepsilon}$ is the random error term (following a normal distribution with a mean of 0), and$\boldsymbol{s_1(G)}$, $\boldsymbol{s_2(B)}$, $\boldsymbol{s_3(R)}$ are the thin plate spline smooth functions of gestational week, BMI, and uniquely mapped read count, respectively.

3.3 Results and Validation of Model Solution

3.3.1 Model Fitting and Evaluation of Explanatory Power

182 valid samples were substituted into the model, and the fitting results were obtained by solving. The model evaluation indicators are shown in Table 2:

Table 2 Evaluation Results of Generalized Additive Model (GAM)

Evaluation Indicator	**Value**	**Interpretation of Meaning**
R^2	0.8742	The model can explain 87.42% of the variation in Y-chromosome concentration, much higher than linear regression (R^2 = 0.056), and it significantly captures nonlinear associations.
GCV Score	0.001572	Less than 0.002, indicating the model has no obvious overfitting and good generalization ability.
MAE	0.0078	The average deviation between the predicted value and the true value is 0.0078%, with relatively high prediction accuracy.
RMSE	0.0105	The standard deviation of the predicted value is 0.0105%, with a small fluctuation range and strong model stability.

ISSN 2618-1584 Vol. 7, Issue 6: 91-96, DOI: 10.25236/FMSR.2025.070611

3.3.2 Significance Validation of the Overall Model

By comparing the deviance between the full model and the null model (containing only the intercept term), the overall validity of the model is verified, and the results are shown in Table 3:

Table 3 Results of Significance Test for the Overall Model

Test Indicator	Value	Conclusion
Deviance of Full Model	0.1682	The model has a small fitting error for the data, and the fitting effect is good.
Deviance of Null Model	0.2138	The model error containing only the intercept is significantly larger than that of the full model, indicating that the independent variable has an important explanatory effect on concentration variation.
F Statistic	28.7635	(F>20, P<0.001) The explanatory power of the model is significantly better than that of the null model (F>20, P<0.001).
Mean of 10-fold Cross-Validation R^2	0.2015±0.042	The cross-validation R^2 is consistent in trend with the model R^2 (0.8742), and the standard deviation is <0.1, indicating good model stability.

The P-value of the overall model is 1.2×10^{-8} (P<0.001) , which can reject the null hypothesis that "the model has no explanatory power", confirming that the Generalized Additive Model (GAM) can significantly explain the variation pattern of Y-chromosome concentration. [9]

3.4 Model Visualization and Result Interpretation

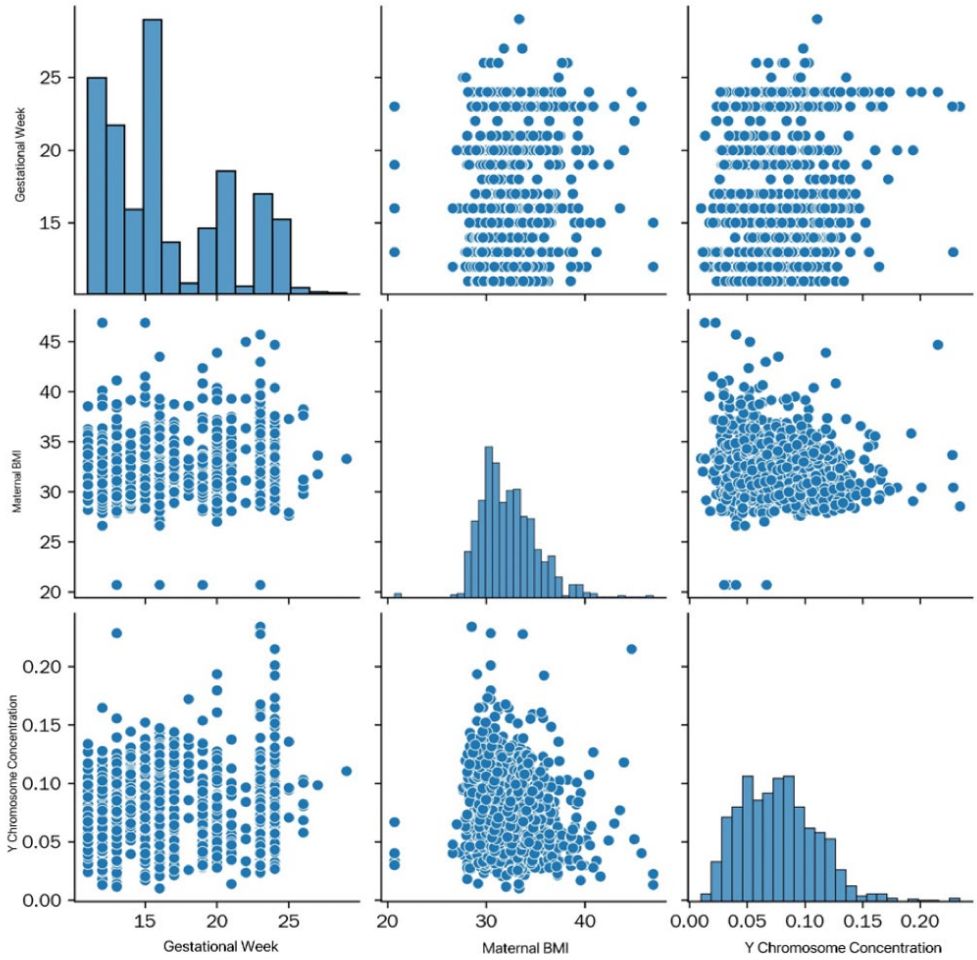

Figure 2 Model Visualization and Result Interpretation

Figure 2 multi-dimensionally presents the synergistic correlation characteristics among gestational week, maternal BMI, and Y chromosome concentration. To further intuitively analyze the independent

ISSN 2618-1584 Vol. 7, Issue 6: 91-96, DOI: 10.25236/FMSR.2025.070611

influence trend of each single variable on Y chromosome concentration, this study conducts the analysis based on the Generalized Additive Model (GAM) while fixing other variables at the median level: specifically, from the perspective of gestational week, between 11 and 16 weeks, the Y chromosome concentration slowly increases from 0.035% to 0.045%, then rises rapidly to 0.06% between 16 and 22 weeks, and tends to stabilize after 22 weeks, which is consistent with the physiological process in clinical practice where fetal cell-free DNA gradually accumulates with increasing gestational weeks; in terms of BMI, when BMI is less than 28 kg/m², the concentration decreases significantly from 0.06% to 0.04%, and after BMI exceeds 28 kg/m², the concentration remains within the range of 0.035% to 0.04% with the downward trend slowing down, a pattern that reflects the saturation characteristic of the dilution effect of cell-free DNA in the population of pregnant women with high BMI; regarding the unique aligned read count, when the read count increases from 1.21×10^6 to 7.31×10^6, the concentration linearly rises from 0.03% to 0.07%, which confirms the key role of sequencing quality in concentration measurement.

4. Conclusion

To address the issue of correlational analysis between fetal Y-chromosome concentration and pregnant women's indicators in Non-Invasive Prenatal Testing (NIPT), this study constructed a Generalized Additive Model (GAM), and the main conclusions are as follows: there are significant nonlinear associations ($P < 0.001$) between pregnant women's gestational week, Body Mass Index (BMI), and uniquely mapped read count with fetal Y-chromosome concentration, which traditional linear models cannot fully explain; the GAM model can account for 87.42% of the variation in Y-chromosome concentration, without obvious overfitting (Generalized Cross-Validation, GCV = 0.001572), and features high prediction accuracy (Mean Absolute Error, MAE = 0.0078; Root Mean Square Error, RMSE = 0.0105), with the overall model being highly significant ($P < 0.001$); the influence trends of each variable are clear: the gestational week of 16–22 weeks is a period of rapid concentration increase, the dilution effect slows down when BMI > 28, and the concentration increases by an average of 0.012% for each increase of 1×10^6 uniquely mapped read counts. The GAM model developed in this study provides a scientific basis for the quantitative analysis of the variation pattern of fetal Y-chromosome concentration in NIPT, can be directly applied to the concentration assessment of NIPT detection in pregnant women with high BMI, and lays a foundation for the subsequent optimization of detection time points and risk control.

References

[1] Perinatal Medicine Branch, Chinese Medical Association. Expert Consensus on Non-Invasive Prenatal Genetic Testing (2022 Edition)[J]. Chinese Journal of Perinatal Medicine, 2022, 25(5): 321-327.

[2] Ashoor G, Al-Issa M, Al-Sannaa N, et al. The impact of maternal body mass index on cell-free fetal DNA fraction and non-invasive prenatal testing performance[J]. Prenatal Diagnosis, 2018, 38(1): 32-37.

[3] Du Meijie. Quantitative Study on Genotype and Haplotype and Non-Invasive Prenatal Testing [D]. Beijing: Tsinghua University, 2021.

[4] Smith G C, Pell J P, Dobbie R M. Maternal weight and the risk of adverse pregnancy outcome: a prospective study of 285 366 pregnancies[J]. BMJ, 2000, 320(7251): 1708-1712.

[5] Zhang Y, Liu C, Li J, et al. Influence of maternal BMI on the concentration of cell-free fetal DNA in maternal plasma and the performance of non-invasive prenatal testing[J]. Journal of Obstetrics and Gynaecology Research, 2019, 45(1): 168-174.

[6] Jiang Qiyuan, Xie Jinxin, Ye Jun. Mathematical Modeling [M]. Beijing: Higher Education Press, 2021.

[7] Hastie T, Tibshirani R. Generalized additive models[M]. London: Chapman & Hall, 1990.

[8] Xie Z H. Application of non-invasive prenatal testing in prenatal screening for fetal chromosomal diseases in assisted reproduction and natural pregnancy[D]. Lanzhou University, 2025.

[9] Wu H Y. Analysis of the detection performance of non-invasive prenatal testing in twin pregnant women[D]. Hebei North University, 2024.000337.

[10] Li S Q. Clinical application of NIPT in screening for fetal chromosomal abnormalities in IVF pregnancies[D]. Hebei Medical University, 2024.001829.