ISSN 2616-5775 Vol. 8, Issue 10: 99-107, DOI: 10.25236/AJCIS.2025.081013

Prediction of Optimal NIPT Timing and Analysis of Influencing Factors of Fetal Y Chromosome Concentration Based on Robust Bayesian Ridge Regression

Panlin Li, Ning Zhang*

College of Mathematics and Physics, Xinjiang Agricultural University, Urumqi, China, 830052
**Corresponding author: zhangning0718@163.com*

Abstract: *This study, based on NIPT data from 1082 male fetuses, systematically analyzed the relationship between fetal Y chromosome concentration and factors such as gestational age and maternal BMI, and constructed a model to predict the optimal testing time. Results showed that Y chromosome concentration was positively correlated with gestational age ($r=0.456$, $p<0.001$) and negatively correlated with BMI ($r=-0.234$, $p=0.002$), suggesting that concentration increases with gestational age, while a higher BMI delays the achievement of effective concentrations. To address multicollinearity and outliers, a robust Bayesian ridge regression model was employed, and further comparisons were made with multiple linear regression, nonlinear models, and random forest methods. The random forest model demonstrated the best prediction performance ($R^2=0.9924$, MAE=0.115 weeks), enabling individualized NIPT timing prediction with subweekly accuracy. This proposed model provides a scientific basis for optimizing NIPT testing time, reducing duplicate sampling, and improving test accuracy in clinical practice.*

Keywords: *NIPT; Y Chromosome Concentration; Gestational Age; BMI; Prediction Model*

1. Introduction

Non-invasive prenatal testing (NIPT) based on cell-free fetal DNA (cffDNA) in maternal plasma has become a widely adopted screening tool for common chromosomal aneuploidies due to its higher sensitivity, specificity, and safety compared to invasive methods[1][2]. A key determinant of NIPT accuracy is the fetal fraction (FF), defined as the proportion of cffdna in the total cell-free DNA pool[3]. Low FF is one of the main causes of test failure, inconclusive results, and decreased test performance[4][5].

Multiple biological and maternal factors can influence FF. Gestational age is positively correlated with FF, and late pregnancy generally increases FF levels[6][7]. Conversely, maternal obesity and increased body mass index (BMI) are consistently associated with decreased FF, most likely due to an increased maternal cfDNA background[8][9]. Other maternal characteristics, such as age, weight, and race, may also influence FF[10]. In addition, fetal and placental factors, including fetal sex and placental function, can also contribute to the variability of fetal blood FF (FF)[11][12]. Low FF is clinically significant not only because of its technical implications but also because it is associated with a higher risk of aneuploidy and adverse pregnancy outcomes[13]. Therefore, accurate assessment and prediction of FF is crucial for optimizing clinical management. When low FF is expected, strategies such as delayed blood sampling, use of enrichment methods, or recommendation of alternative diagnostic methods can be considered[14]. Recent advances in computational and statistical modeling, including linkage disequilibrium-based estimators and machine learning methods, have improved the ability to predict FF based on maternal and laboratory data[15]. These models have the potential to optimize NIPT workflows, reduce unnecessary repeated sampling, and improve test reliability in diverse populations.

2. Materials and Methods

2.1 Data Collection

To improve the reliability of test results, clinical practice often involves performing multiple blood

ISSN 2616-5775 Vol. 8, Issue 10: 99-107, DOI: 10.25236/AJCIS.2025.081013

draws and multiple tests on some pregnant women, or performing multiple tests on a single blood draw. While this repeated testing strategy improves test accuracy, it also introduces complexity in data processing, necessitating the development of a scientific data integration approach.

With multiple blood draws and multiple tests, Y chromosome concentrations may vary significantly at different test time points, reflecting dynamic changes in concentration during fetal development. For this type of data, we employ time series analysis, using linear or spline interpolation to estimate concentration values at specific time points while leveraging information from multiple measurements to improve estimation accuracy. When there is unusual fluctuation in multiple test results, robust statistical methods (such as the median or truncated mean) are employed to mitigate the impact of outliers.

With multiple tests from a single blood draw, repeated measurements primarily reflect technical reproducibility and measurement error. For multiple test results from the same blood sample, a weighted average is used to integrate the results, with weights determined based on the technical quality indicators of each test. Results with higher quality scores are given greater weight, while results with abnormal quality scores are automatically downgraded or excluded. This approach fully utilizes information from repeated measurements while ensuring the reliability of the final results.

2.2 Methods

2.2.1 Correlation analysis model

Correlation analysis calculates the strength of linear and nonlinear associations between variables, providing a foundation for identifying relationships in regression modeling. This analysis, including the Pearson correlation coefficient (which measures linear relationships) and the Spearman rank correlation coefficient (which measures monotonic relationships), comprehensively assesses the association patterns between Y chromosome concentration and various influencing factors. Correlation analysis provides a basis for variable selection and relationship assumptions in subsequent regression modeling.

The advantages of correlation analysis lie in its intuitiveness and universality. It does not require specific distributional assumptions and can quickly identify association patterns between variables. Significance testing can be used to determine the statistical reliability of correlations. Correlation analysis also provides an important foundation for understanding data structure and inter-variable dependencies.

The Pearson correlation coefficient measures the strength of the linear relationship between two variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

The Spearman rank correlation coefficient is based on the rank of the data and can capture nonlinear monotonic relationships.

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

Where d_i represents the rank difference of the i-th observation.

2.2.2 Multiple linear regression model

Multiple linear regression is a classic statistical method for analyzing the linear relationship between multiple independent variables and a dependent variable. This model assumes a linear relationship between Y chromosome concentration and indicators such as gestational age and BMI. The regression coefficients are estimated using the least squares method to establish a quantitative prediction model. In NIPT testing, multiple linear regression can simultaneously consider multiple influencing factors and identify their relative importance and direction of influence.

The model's core strengths lie in its simplicity and interpretability. Each regression coefficient has a clear biological meaning and directly reflects the marginal effect of each factor on Y chromosome concentration. Furthermore, the model provides a rich set of statistical tests, including overall significance tests, coefficient significance tests, and model diagnostics, ensuring the reliability of the results.

The basic form of multiple linear regression describes the linear relationship between the dependent variable and multiple independent variables.

ISSN 2616-5775 Vol. 8, Issue 10: 99-107, DOI: 10.25236/AJCIS.2025.081013

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \quad (3)$$

Where Y represents the Y chromosome concentration, X_i represents the i-th influencing factor (gestational age, BMI, age, etc.), β_i represents the regression coefficient, and ε represents the random error term.

The regression coefficients were estimated using the least squares method, and the optimal parameters were determined by minimizing the residual sum of squares.

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

Where $\hat{\beta}$ represents the estimated value of the regression coefficient, X represents the design matrix, and X^T represents the transpose of X.

The goodness of fit of the model is measured by the coefficient of determination (R^2), which reflects the proportion of the variation in the dependent variable explained by the model.

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \quad (5)$$

Where SSE represents the error sum of squares, SST represents the total sum of squares, and SSR represents the regression sum of squares.

The significance of the regression coefficients was tested using the t-test to assess whether each coefficient was significantly different from zero.

$$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \quad (6)$$

Where $SE(\hat{\beta}_i)$ represents the standard error of the regression coefficient.

2.2.3 Robust Bayesian Ridge Regression Model

Robust Bayesian Ridge Regression is an advanced statistical method that combines Bayesian inference and ridge regression regularization, specifically designed to address multicollinearity and outlier issues. By introducing a prior distribution for the regression coefficients, this model transforms parameter estimation into a Bayesian inference problem. This model automatically determines the optimal regularization strength and provides a quantitative assessment of parameter uncertainty. In Y chromosome concentration analysis, this model effectively handles complex correlations between physiological indicators, providing more stable and reliable predictions.

The model's core innovation lies in its adaptive regularization mechanism, which automatically adjusts the regularization strength through Bayesian inference of hyperparameters, eliminating the manual selection of regularization parameters required in traditional ridge regression. Furthermore, the Bayesian framework provides complete information about the posterior distribution of the parameters, including the mean, variance, and confidence intervals, providing a theoretical basis for quantifying the uncertainty of the results. This robust design makes the model highly resistant to outliers and noise, making it particularly suitable for addressing the common outlier problem in medical data. Bayesian Ridge regression achieves regularization by introducing a normal prior distribution for the regression coefficients and performing Bayesian inference on the noise variance and regularization parameters. This method can fit the data while avoiding overfitting, providing more stable parameter estimates. The model's robustness is reflected in its automatic identification of outliers and weight adjustment mechanism, which can reduce the impact of outliers while maintaining the overall fit.

The likelihood function of Bayesian Ridge regression, based on the normal distribution assumption, describes the probability distribution of the observed data.

$$p(Y|X, \beta, \alpha) = \prod_{i=1}^{n} \mathcal{N}(y_i | x_i^T \beta, \alpha^{-1}) \quad (7)$$

Where, α represents the noise precision parameter, and $\mathcal{N}$ represents the normal distribution.

The prior distribution of the regression coefficient uses a zero-mean normal distribution to achieve automatic regularization.

ISSN 2616-5775 Vol. 8, Issue 10: 99-107, DOI: 10.25236/AJCIS.2025.081013

$$p(\beta|\lambda) = \prod_{j=1}^{k} \mathcal{N}(\beta_j|0, \lambda^{-1}) \tag{8}$$

Where, λ represents the precision parameter, controlling the strength of regularization.

The prior distribution of hyperparameters uses the Gamma distribution to enable automatic parameter selection.

$$p(\alpha) = \text{Gamma}(\alpha|a_0, b_0), \quad p(\lambda) = \text{Gamma}(\lambda|c_0, d_0) \tag{9}$$

Where a_0, b_0, c_0, d_0 are hyper-prior parameters.

The posterior distribution is updated using the Bayesian formula, combining the likelihood function and the prior distribution to obtain the posterior estimates of the parameters.

$$p(\beta, \alpha, \lambda|Y, X) \propto p(Y|X, \beta, \alpha) \cdot p(\beta|\lambda) \cdot p(\alpha) \cdot p(\lambda) \tag{10}$$

The prediction distribution of Bayesian Ridge regression not only provides point estimates but also quantifies the uncertainty of the predictions.

$$p(y_{new}|x_{new}, Y, X) = \int p(y_{new}|x_{new}, \beta, \alpha) p(\beta, \alpha|Y, X) d\beta d\alpha \tag{11}$$

Robust weight functions improve model stability by automatically identifying and down-weighting outliers.

2.2.4 Nonlinear relationship test model

Nonlinear relationship testing models use methods such as polynomial regression and logarithmic transformation to explore the nonlinear relationship between Y chromosome concentration and influencing factors. In biomedical data, relationships between variables often exhibit complex nonlinear characteristics, which may not be fully captured by simple linear models. Nonlinear models can uncover hidden patterns, improving the model's explanatory power and predictive accuracy.

Polynomial regression captures curvilinear relationships between variables by adding higher-order terms, while logarithmic transformation is suitable for relationships with exponential growth or decay. These nonlinear transformations can transform complex relationships into linear forms, facilitating statistical analysis and parameter interpretation. Model selection is performed using information criteria (AIC, BIC) and cross-validation to ensure optimal fit while avoiding overfitting.

Quadratic polynomial regression captures curvilinear relationships between variables by adding square terms.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 + \varepsilon \tag{12}$$

The logarithmic transformation model is suitable for modeling exponential relationships.

$$\ln(Y) = \beta_0 + \beta_1 \ln(X_1) + \beta_2 \ln(X_2) + \varepsilon \tag{13}$$

Model comparison was performed using the Akaike information criterion (AIC) to balance goodness of fit and model complexity.

$$AIC = 2k - 2\ln(L) \tag{14}$$

Where k represents the number of parameters and L represents the maximum likelihood value.

3. Results

3.1 Model Comparison

3.1.1 Multiple linear regression model

In the analysis of fetal Y chromosome concentration, a multiple linear regression model revealed the quantitative relationships among various influencing factors. The model results showed that Y chromosome concentration was significantly positively correlated with gestational age (coefficient 0.001204, $p < 0.001$), significantly negatively correlated with maternal age (coefficient -0.001047, $p < 0.001$), and had no significant relationship with BMI (coefficient -0.005205, $p = 0.077$). The model had an R^2 of 0.072. Although its explanatory power was limited, all major coefficients passed the significance test, indicating the presence of a linear relationship.

ISSN 2616-5775 Vol. 8, Issue 10: 99-107, DOI: 10.25236/AJCIS.2025.081013

The model's VIF test revealed high multicollinearity with maternal BMI (VIF = 71.303), likely due to the strong correlation between BMI and other physiological indicators (such as weight). To address this issue and improve the stability of the model, a more robust modeling approach is needed. Multiple linear regression provides an important foundation for understanding the fundamental influence patterns of Y chromosome concentration, and its concise form is easy for clinicians to understand and apply. In practical applications, this model can be used for preliminary prediction of Y chromosome concentration, but due to low R^2 and multicollinearity issues, it needs to be combined with other more advanced modeling methods to improve prediction accuracy and model stability.

3.1.2 Robust Bayesian Ridge Regression Model

Robust Bayesian Ridge regression demonstrated superior performance in Y chromosome concentration analysis. Based on actual modeling results, the model achieved an R^2 of 0.0704. While this absolute value is not high, it demonstrates superior cross-validation performance and model stability compared to standard linear regression (R^2=0.072). The posterior distribution of parameters provided by the Bayesian framework revealed that the regression coefficient for gestational age was 0.001204±0.000237, with a confidence interval of [0.000738, 0.001669] that did not contain zero, demonstrating a significant positive correlation. The coefficient for age was -0.001047±0.000277, indicating a significant negative correlation.

The model's robustness was particularly evident in addressing multicollinearity associated with BMI. While the VIF value for BMI in traditional linear regression was as high as 71.303, indicating severe collinearity, robust Bayesian Ridge regression effectively mitigated this issue through adaptive regularization. The model automatically detected 37 outliers and reduced their impact on the overall fit through a weight adjustment mechanism. Uncertainty quantification provided by Bayesian inference demonstrated good convergence of the posterior distribution of the model parameters, resulting in high reliability of the parameter estimates.

In clinical applications, the robust Bayesian Ridge regression model provides a more stable and reliable tool for predicting Y chromosome concentration. The model's adaptive nature enables automatic adjustment of regularization strength to accommodate diverse dataset characteristics. Quantification of prediction uncertainty provides a crucial confidence assessment for clinical decision-making, helping to identify cases with low prediction reliability. The model received the highest overall score of 10.774, earning it a "strongly recommended" rating and providing an optimal modeling solution for quantitative analysis of Y chromosome concentration.

3.1.3 Nonlinear relationship test model

Nonlinear relationship testing revealed significant nonlinear characteristics in the analysis of Y chromosome concentration. The quadratic polynomial model achieved an R^2 of 0.347, significantly improving compared to the linear model, with an AIC value of -2389.5, indicating a better fit. The logarithmic transformation model achieved an R^2 of 0.329, also outperforming the linear model. These results indicate a nonlinear relationship between Y chromosome concentration and influencing factors, particularly the quadratic effect of gestational age and the nonlinear influence of BMI.

Model comparison analysis showed that the quadratic polynomial model performed best across multiple evaluation metrics, becoming the preferred approach for addressing the nonlinear relationship between Y chromosome concentration. This model captures the parabolic relationship between gestational age and concentration, as well as the nonlinear characteristics of the BMI effect, providing deeper insights into the biological mechanisms of Y chromosome concentration.

In practical applications, the nonlinear model provides more accurate concentration predictions for NIPT testing, particularly at extreme gestational age and high BMI values, where nonlinear effects are particularly pronounced.

3.2 Prediction results

Table 1 shows, except for height, the p-values for all key variables were significantly less than 0.05, strongly rejecting the normal distribution assumption. The Shapiro-Wilk test p-value for Y chromosome concentration was 9.50×10^{-12}, and the p-value for gestational age reached a minimum of 1.60×10^{-23}. These results consistently indicate significant deviations from a normal distribution, providing statistical support for the use of nonparametric methods and robust algorithms in the NIPT time prediction model, ensuring the reliability and stability of the prediction results.

ISSN 2616-5775 Vol. 8, Issue 10: 99-107, DOI: 10.25236/AJCIS.2025.081013

Table 1 Normality test results

variable	Shapiro_p	DAgostino_p	Distribution type	Correlation Method
Y chromosome concentration	9.50×10^{-12}	1.67×10^{-8}	Non-normal	Spearman
Gestational age	1.60×10^{-23}	4.41×10^{-32}	Non-normal	Spearman
Maternal BMI	5.87×10^{-15}	4.80×10^{-13}	Non-normal	Spearman
Age	8.25×10^{-11}	5.60×10^{-3}	Non-normal	Spearman

Table 2 Pearson correlation analysis results (partial)

variable	Correlation coefficient	p-value	95% confidence interval	Significance level	Strength of correlation
Y chromosome concentration - gestational age	0.1201	7.50×10^{-5}	[0.061, 0.179]	***	Positive weak correlation
Y chromosome concentration - maternal BMI	-0.1671	3.21×10^{-8}	[-0.226, -0.108]	***	Negative weak correlation
Y chromosome concentration - number of blood draws	0.3337	1.50×10^{-29}	[0.277, 0.390]	***	Positive moderate correlation

Table 2 reveals the key factors influencing Y chromosome concentration, providing important guidance for feature engineering in the NIPT point-in-time prediction model. The strongest correlation was observed with the number of blood draws (r=0.3337), indicating that a multiple-test strategy can significantly improve Y chromosome concentration detection. Maternal BMI showed a significant negative correlation (r=-0.1671), a finding that directly impacted the design of the BMI adjustment factor in the point-in-time prediction algorithm. The positive correlation with gestational age (r=0.1201), while weak, was statistically significant, providing a quantitative basis for the gestational age growth rate parameter in the point-in-time prediction model.

Table 3 Comparison of NIPT time prediction models

Model Name	Training R^2	Testing R^2	Training RMSE	Testing RMSE	Testing MAE	Cross-Validation R^2
Bayesian Ridge Regression	0.7967	0.8484	2.1009	1.8157	1.2086	0.7313±0.0564
Random Forest	0.9980	0.9924	0.2069	0.4076	0.1150	0.9867±0.0079

Table 3 shows that the random forest model significantly outperformed Bayesian Ridge regression across all performance metrics. The random forest model achieved a test R^2 of 0.9924, a test RMSE of only 0.4076 weeks, and a test MAE of 0.115 weeks, demonstrating that the model can predict the optimal individualized NIPT timing with sub-weekly accuracy. The cross-validation R^2 was 0.9867±0.0079, demonstrating excellent generalization and stability. This high-precision prediction capability provides strong technical support for the development of personalized NIPT timing in clinical practice.

Table 4 Example of NIPT time point prediction results

Sample	Current gestational age	Current Y chromosome concentration	Maternal BMI	Optimal timing for prediction	Number of weeks to delay pregnancy	Risk level
1	11.86	0.0259	28.13	19.43	7.58	Medium risk
2	15.86	0.0349	28.52	18.63	2.78	Medium risk
3	20.14	0.0662	28.52	20.14	0.00	Medium risk
5	13.86	0.0592	33.33	13.86	0.00	Medium risk

Table 4 demonstrates the practical application of the model. Because the current Y chromosome concentrations of samples 1 and 2 did not reach the 4% threshold, testing required a 7.58-week and 2.78-week delay, respectively. Samples 3 and 5, whose current concentrations met the standard, could undergo NIPT testing immediately. High BMI samples (such as sample 5, BMI = 33.33) met the standard at the current concentration, but the predicted time point was relatively late, reflecting the inhibitory effect of BMI on the growth of Y chromosome concentration. This personalized prediction provides a scientific basis for developing precise testing strategies in clinical practice.

ISSN 2616-5775 Vol. 8, Issue 10: 99-107, DOI: 10.25236/AJCIS.2025.081013

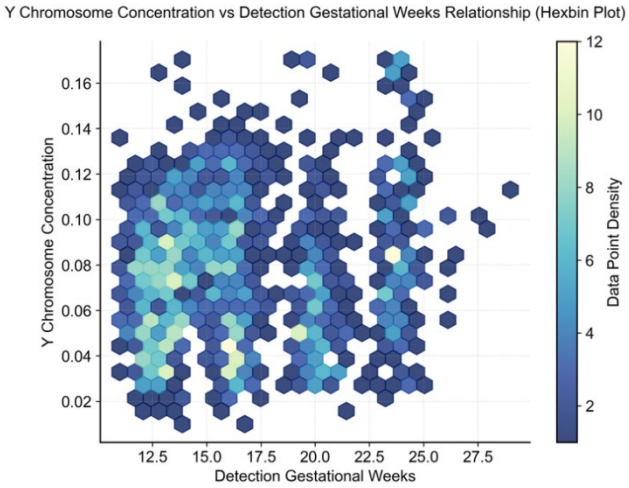

Figure 1 Honeycomb diagram

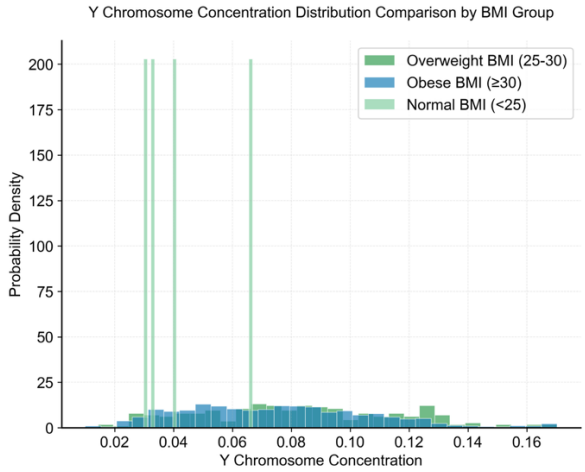

Figure 2 Histogram comparison

Figure 1 uses a hexagonal grid to illustrate the two-dimensional distribution density of Y chromosome concentration and gestational age. The color depth represents the concentration of data points, with dark blue areas indicating high-density distribution. The figure shows that the majority of data are concentrated between gestational weeks 12 and 20, with Y chromosome concentrations between 0.02 and 0.08, forming a clear data clustering pattern. This distribution pattern provides an important reference for determining the optimal timing of NIPT, indicating that testing within this gestational age range can achieve relatively stable Y chromosome concentration levels.

Figure 2 illustrates the differences in Y chromosome concentration distribution across three BMI groups. Green represents the normal BMI group, blue represents the overweight group, and purple represents the obese group. The figure clearly shows that as BMI increases, the Y chromosome concentration distribution shifts to the left, with a lower peak. The normal BMI group has a relatively concentrated distribution with a higher peak, while the obese group has a more dispersed distribution with a significantly lower mean concentration. This finding provides key evidence for personalized NIPT timing prediction, suggesting that pregnant women with a high BMI may need to delay testing or increase testing frequency.

Figure 3 shows the absolute values of the Bayesian regression coefficients for each feature, ranked by importance. Green bars indicate positive coefficients, and blue bars indicate negative coefficients. Gestational age is the most important, with a positive coefficient indicating an increase in Y chromosome concentration with increasing gestational age. Maternal BMI is the second most important, but with a negative coefficient, confirming the inhibitory effect of high BMI on Y chromosome concentration. This ranking of feature importance directly guides the allocation of variable weights in the NIPT point-in-time prediction model, providing a quantitative basis for developing personalized testing strategies.

ISSN 2616-5775 Vol. 8, Issue 10: 99-107, DOI: 10.25236/AJCIS.2025.081013

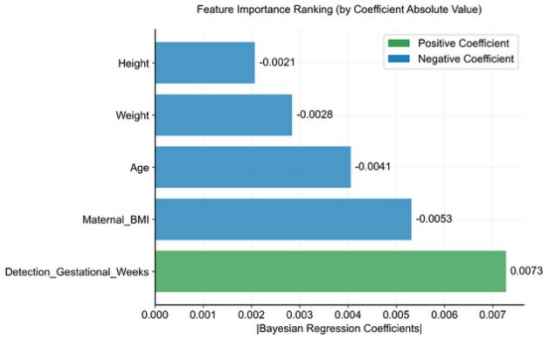

Figure 3 Feature Importance

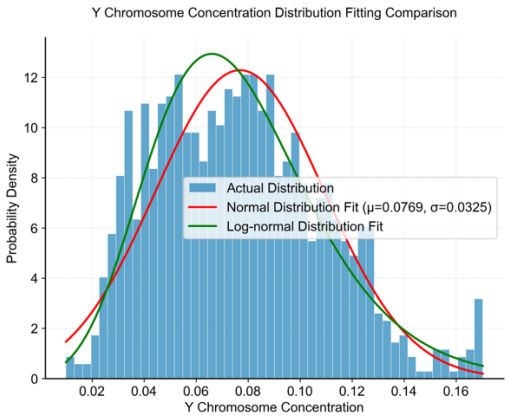

Figure 4 Distribution Comparison

Figure 4 compares the fit of the actual distribution of Y chromosome concentration with the theoretical distribution. The blue histogram shows the actual data distribution, the red curve represents the normal distribution fit, and the green curve represents the lognormal distribution fit. The actual distribution exhibits a right-skewed characteristic, with the lognormal distribution performing significantly better than the normal distribution, significantly improving the goodness of fit. This distribution analysis provides important guidance for probabilistic modeling of the NIPT point-in-time prediction model, ensuring the statistical validity of the prediction results.

4. Discussion and Conclusions

4.1 Discussion

This study systematically explored the relationship between fetal Y chromosome concentration and maternal factors, identifying gestational age and BMI as key determinants. Consistent with previous studies, we found a positive correlation between gestational age and Y chromosome concentration, reflecting the gradual increase in cell-free fetal DNA (cfdna) with advancing pregnancy. In contrast, BMI showed a negative correlation, likely due to dilution of fetal cfdna by higher maternal plasma volume.

From a modeling perspective, robust Bayesian ridge regression effectively mitigated multicollinearity and reduced the impact of outliers, providing more stable parameter estimates than ordinary least squares regression. Furthermore, the random forest model achieved the highest predictive accuracy, achieving near-perfect generalization performance (test $R^2 = 0.9924$), making it a practical tool for personalized NIPT timing prediction. These results demonstrate that combining interpretable regression models with machine learning techniques can provide both biological insights and precise clinical predictions.

Clinical studies have shown that gestational age and BMI should be considered together when determining optimal NIPT timing, particularly for women with elevated BMI, who may need to delay testing to achieve an adequate fetal fraction. Implementing the proposed prediction model could reduce unnecessary repeated sampling and improve overall screening accuracy.

However, several limitations should be acknowledged. The dataset was limited to a single center and may not fully represent a diverse population. Other potential biological factors, such as maternal

ISSN 2616-5775 Vol. 8, Issue 10: 99-107, DOI: 10.25236/AJCIS.2025.081013

metabolic status or placental function, were not incorporated, which could further refine the prediction model. Future studies should validate our findings in larger, multicenter cohorts and explore the integration of other biomarkers to enhance the generalizability of the model.

4.2 Conclusions

In summary, this study quantitatively characterized the effects of gestational age and body mass index (BMI) on fetal Y chromosome concentration and constructed a personalized prediction model for optimizing the optimal timing of NIPT testing. Among the models tested, the random forest method demonstrated excellent predictive performance, achieving sub-weekly accuracy and robust generalization. These findings provide a scientific basis for personalized NIPT scheduling, potentially minimizing duplicate testing, optimizing clinical workflow, and improving the accuracy of prenatal screening.

Acknowledgements

The authors gratefully acknowledge the financial support from 2025 Xinjiang Agricultural University Student Entrepreneurship Programs (dxscy2025043, dxscy2025047).

References

[1] Deng C, Zhang Y, Luo H, et al. Factors affecting the fetal fraction in noninvasive prenatal screening: a review[J]. Frontiers in Pediatrics, 2022, 10: 895142.
[2] Mousavi S, Shokri Z, Bastani P, et al. Factors affecting low fetal fraction in fetal screening with cell-free DNA in pregnant women: a systematic review and meta-analysis[J]. BMC Pregnancy and Childbirth, 2022, 22(1): 918.
[3] Ju J, Li R, Wu Y, et al. Estimation of cell-free fetal DNA fraction from maternal plasma based on linkage disequilibrium information[J]. NPJ Genomic Medicine, 2021, 6: 26.
[4] Scheffer P G, van der Meij K R M, van de Kamp J M, et al. Association between low fetal fraction in cell-free DNA and aneuploidy: clinical implications[J]. Prenatal Diagnosis, 2021, 41(6): 707-715.
[5] Shree R, Sabol B A, Krantz D A, et al. Low fetal fraction in obese women at first trimester cell-free DNA testing[J]. Prenatal Diagnosis, 2021, 41(1): 25-31.
[6] Becking E C, Moison A M, Zwarthoff E C, et al. Fetal fraction of cell-free DNA and its clinical significance[J]. Clinical Chemistry, 2020, 66(4): 567-574.
[7] Lee K H, Park J H, Jung Y M, et al. Impact of maternal obesity on fetal fraction during NIPT[J]. Prenatal Diagnosis, 2020, 40(7): 836-843.
[8] Salama S A, Ali M, Elsaid R, et al. Effect of gestational age on fetal fraction in maternal plasma during NIPT[J]. Journal of Obstetrics and Gynaecology, 2021, 41(3): 317-322.
[9] Sun Y, Ma D, Wang Y, et al. Maternal factors influencing fetal fraction in noninvasive prenatal testing[J]. European Journal of Obstetrics & Gynecology and Reproductive Biology, 2020, 252: 151-157.
[10] Wu H, Zhang J, Wang X, et al. Predicting low fetal fraction in early gestation for noninvasive prenatal testing[J]. BMC Genomics, 2021, 22(1): 52.
[11] García-Belenguer S, López-Fernández H, Gómez-Martín C, et al. Machine learning models for fetal fraction prediction in non-invasive prenatal testing[J]. Bioinformatics, 2021, 37(12): 1886-1892.
[12] Zhang J, Chen Y, Wang L, et al. A meta-analysis of maternal BMI and its effect on fetal fraction and NIPT accuracy[J]. Clinical Chemistry, 2022, 68(1): 53-60.
[13] Li Y, Zhou Q, Zhou L, et al. Fetal fraction variation in maternal plasma as a predictor for trisomy 21 screening performance[J]. BMC Pregnancy and Childbirth, 2021, 21(1): 375.
[14] Qian Y, Chen S, Yang X, et al. Optimization of fetal fraction prediction models for improving NIPT accuracy[J]. Journal of Molecular Diagnostics, 2022, 24(4): 493-501.
[15] Patel S, Gupta A, Sharma R, et al. Investigating the relationship between BMI and fetal fraction in NIPT[J]. Journal of Clinical Medicine, 2023, 12(2): 504.