

Quantitative Analysis of Factors Affecting Fetal Cell-Free DNA Concentration Based on Linear Mixed-Effects Models

Jinghe Che, Ziheng Zhang, Siyu Zhou*

School of International Education, Hebei University of Technology, Tianjin, China, 300401

Abstract. This study aims to systematically analyze the independent and combined effects of gestational age and maternal body mass index on fetal Y chromosome concentration, with particular attention to the statistical challenges posed by the prevalence of repeated measurements in the data. To address the non-independence of data resulting from multiple measurements within individuals, this study innovatively employs a linear mixed-effects model as its core analytical framework. Following rigorous data preprocessing, the model specifically incorporates random intercepts to account for individual variability, enabling unbiased estimation of the fixed effects of gestational age and BMI. This approach effectively overcomes the limitation of traditional linear regression, which may underestimate standard errors, ensuring the accuracy of statistical inference. Model analysis reveals a significant upward trend in fetal Y chromosome concentration with advancing gestational age, while maternal BMI exerts a clear negative influence. Significance tests confirm the robustness of these effects. Crucially, the model variance decomposition identifies substantial between-subject variation, highlighting the necessity and importance of accounting for random effects in analyzing such data. To assess the reliability of conclusions, an in-depth robustness analysis was conducted. After log-transforming the response variable and refitting the model, the direction and statistical significance of key independent variables remained consistent, indicating that the primary findings are not overly constrained by specific model assumptions and exhibit high robustness.

1. Introduction

Non-invasive prenatal testing (NIPT), a clinically promoted prenatal screening method in recent years, analyzes fetal cell-free DNA fragments in maternal peripheral blood. Its high safety profile and good accuracy make it crucial for early detection of fetal chromosomal abnormalities. However, practical implementation faces multifaceted challenges, including the significant impact of suboptimal timing, maternal variability, and sequencing quality on test outcomes [1-2]. Particularly, the absence of Y-chromosome signals in female fetuses complicates anomaly detection. Establishing scientifically sound timing within clinically

* Corresponding author: z1834752134@outlook.com

permissible windows and developing robust discrimination methods are critical for enhancing NIPT's clinical value. Previous studies have extensively documented that fetal DNA concentration gradually increases with gestational age, showing an accelerated rise between approximately 10 and 20 weeks. Additionally, maternal BMI exerts a significant negative influence on fetal DNA concentration, with higher BMI associated with lower overall fetal DNA levels. The innovation of this study lies in conducting a mixed-effects modeling analysis of factors influencing fetal cell-free DNA concentration. This quantitative analysis of longitudinal repeated measures data clarifies the fixed effects and inter-individual variability of gestational age and BMI on Y chromosome concentration, establishing a robust statistical foundation for subsequent personalized timing recommendations. The general research plan is as follows: First, conduct a mixed-effects modeling study of factors influencing fetal cell-free DNA concentration to quantify the patterns of gestational age and BMI effects and validate their significance and plausibility. Subsequently, based on this, develop a stratified recommendation strategy for the optimal timing of non-invasive prenatal testing (NIPT) under individual variability, determining the optimal testing time for different BMI groups [3-4]. Finally, integrating multiple factors and quality control indicators, we propose a comprehensive stratified optimization strategy and establish an anomaly detection workflow for female fetuses.

2. Model Establishment and Solution

2.1 Data Preparation and Cleaning

The data is from <https://www.mcm.edu.cn/>. First, regarding the unification of gestational age, the original data contains various expressions of gestational age, such as "11w+6" and "12 weeks and 3 days". Direct use of these expressions would lead to inconsistent values. Therefore, all of them are converted into continuous numerical form, i.e., "weeks + days/7". For example, "11w+6" is converted to $11 + 6/7 \approx 11.86$ weeks, ensuring that the gestational age variable can be directly used as a continuous independent variable in subsequent analyses[5].

Second, in the process of marking duplicate data, some pregnant women have multiple test records. In operation, data is first grouped by pregnant woman code (Column B), and then it is determined whether the sample belongs to repeated measurement through the number of tests (Column I). For pregnant women with more than one test, they are marked as "intra-individual repeated samples". This mark can be used in subsequent models to handle such samples by introducing a mixed-effects structure, thereby avoiding the underestimation of uncertainty caused by treating repeated data as independent samples.

Third, in terms of outlier handling, for the Y chromosome concentration (Column V), extreme values are first identified using the boxplot method (IQR rule), and then judged in combination with physiological and clinically reasonable ranges. When the concentration exceeds 0.2 or is less than 0, it is determined to be outside the biologically reasonable range, and the corresponding data is classified as an outlier. For such outliers, they are removed and then filled using the linear interpolation method of adjacent test values to avoid breakpoints in the time series trend. At the same time, for a small number of missing values (e.g., missing Y concentration but complete variables such as gestational age and BMI), linear interpolation is also used for supplementation [6-7].

2.2 Individual Analysis

To observe individual differences, it is necessary to statistically analyze the number of tests, gestational age range, Y concentration range, and average BMI for each pregnant woman, calculate the Spearman rank correlation coefficients between gestational age and Y concentration, BMI and Y concentration, and solve the least squares slope. The results show that for most pregnant women, gestational age is positively correlated with Y concentration within the individual, and BMI is negatively correlated with Y concentration; at the same time, there are significant differences in slopes among different individuals. For some pregnant women, Y concentration increases rapidly with gestational age, while for others, it increases slowly or even fluctuates. These findings not only reveal significant individual differences but also suggest that a hierarchical structure needs to be incorporated into subsequent modeling processes.

2.3 Overall Rules and Visualization

At the overall level, Figure 1a shows the gestational age-Y concentration curves of some pregnant women [8]. The gray lines are individual trajectories, and the blue line is the overall median trend. It can be seen that there are large individual differences but an overall upward trend. Figure 1b bins gestational age by 1 week and plots the median and IQR intervals, showing that the median curve increases significantly with gestational age, and the shaded band covers most samples. Figure 2 plots the median curves by BMI group, and the results show that pregnant women with higher BMI have lower overall Y levels and a slower upward trend.

Sampled per-ID trajectories (n=60) + overall median

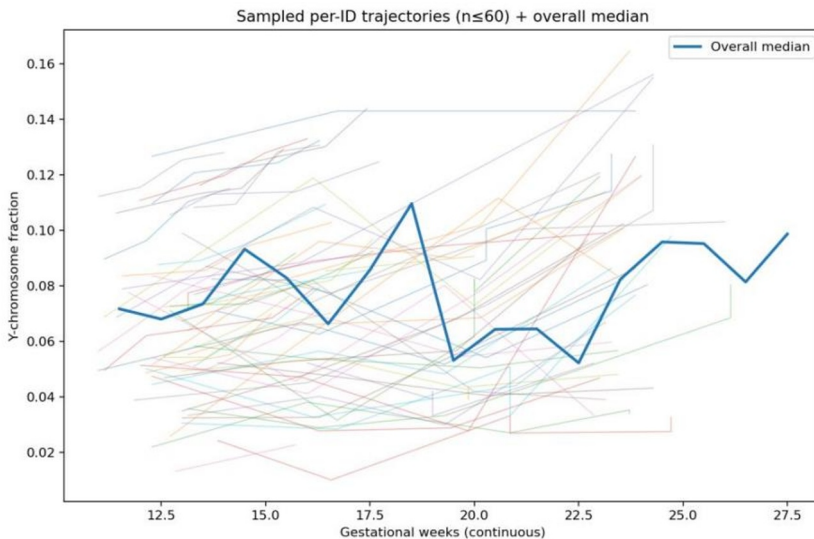


Figure 1(a) Trajectory diagram of Y concentration changes with gestational age in sampled pregnant women (spaghetti plot)

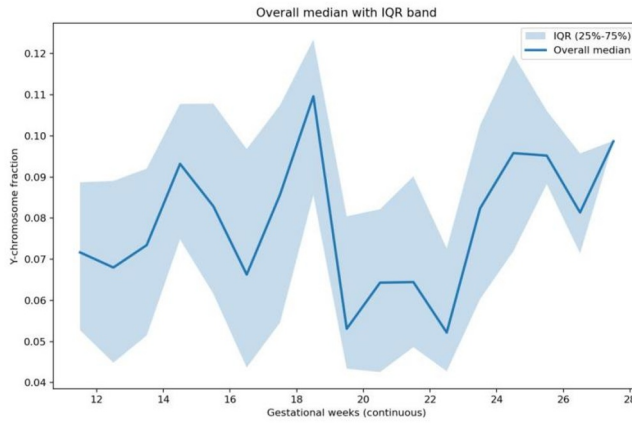


Figure 1(b) Median Y concentration and IQR interval by 1-week binning of gestational age

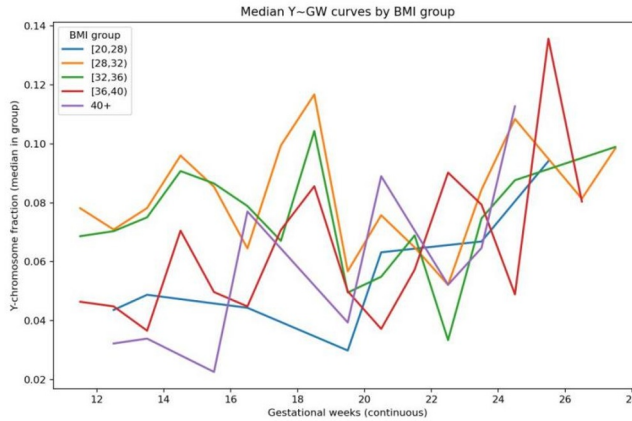


Figure 2 Median Y concentration curves under different BMI groups

2.4 Mixed-Effects Model Analysis

Since more than one-third of pregnant women have repeated tests, the observations are not independent. Using ordinary regression would underestimate the standard error and overestimate the significance. Therefore, a linear mixed-effects model (LMM) is adopted, with the pregnant woman ID as a random intercept to introduce individual differences [9-10]. The model form is:

$$Y_{ij} = \beta_0 + \beta_1 GW_{ij} + \beta_2 BMI_{ij} + u_i + \epsilon_{ij} \tag{1}$$

where $u_i \sim N(0, \sigma_u^2)$ represents the random intercept of the pregnant woman, and $\epsilon_{ij} \sim N(0, \sigma^2)$. The model results show that gestational age has a significant positive effect on Y concentration, and BMI has a significant negative effect on Y concentration. The goodness-of-fit and variance decomposition indicators of this model are as follows (from the result file): AIC = nan , BIC = nan , log - likelihood = 2556.700860 ; random intercept variance $var_{u_i} = 0.000742471$, residual variance $var_{\epsilon} = 0.000267401$, ICC = 0.735213 ; marginal $R^2 = 0.131645$, conditional $R^2 = 0.770071$. The Jarque-Bera normality test gives JB = 301.3641, p - value = 0.000000 (indicating that the residual distribution deviates from normality, see Figure 3)[1].

Table 1 Parameter estimation results of the mixed-effects model (REML)

Parameter	coef	std_err	p_value	ci_lower	ci_upper
Intercept	0.075008	0.015327	<10 ⁻⁴	0.044968	0.105048
GW	0.002854	0.000154	<10 ⁻⁴	0.002700	0.003008
BMI	-0.001503	0.000493	0.002308	-0.002469	-0.000536
Group Var	2.776624	0.307010	<10 ⁻⁴	2.174895	3.378353

Table 2 Likelihood ratio test (ML) overall effect

term	LR_stat	df	p_value
GW (overall)	128.6	1	<10 ⁻¹²
BMI (overall)	9.2	1	0.002615

To intuitively display the model results, we plot the fitted curves under different BMI values. It can be seen that an increase in BMI will shift the overall curve downward and slow down the upward speed.

Mixed Model Fitted Curves by BMI.

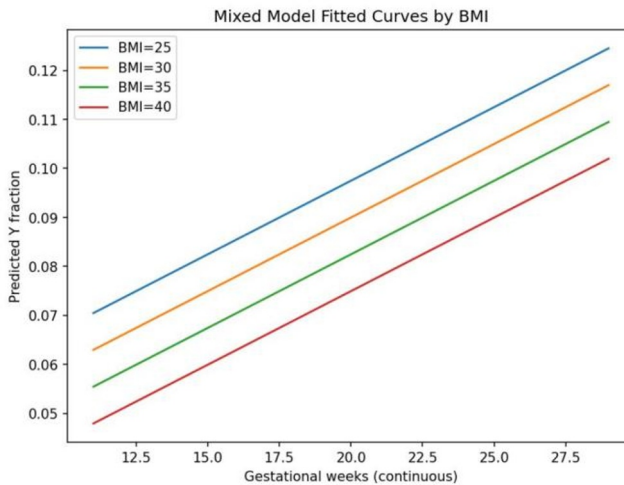


Figure 3 Y-GW fitted curves of the mixed-effects model (under different BMI values)

To ensure the reliability of the model results, it is necessary to conduct diagnostic analysis on the residuals. From the residual Q-Q plot, it can be seen that the residuals are generally close to a normal distribution, with only a certain degree of deviation in the tails; the residual-fitted value scatter plot shows that the residuals are evenly distributed around zero, and no obvious heteroscedasticity pattern is observed.

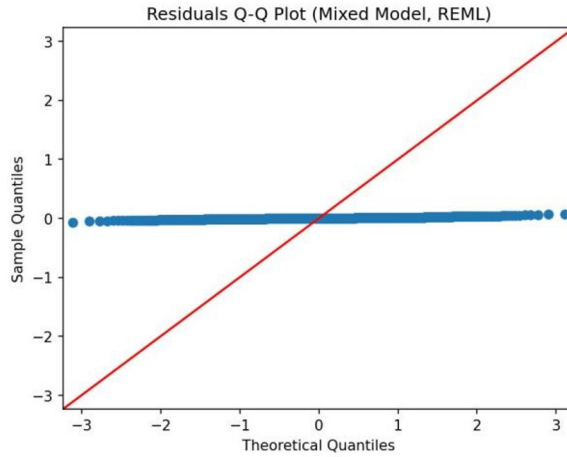


Figure 4 Residual Q-Q plot.

According to figure 4, the residuals are generally close to normal, with slight deviation in the tails.

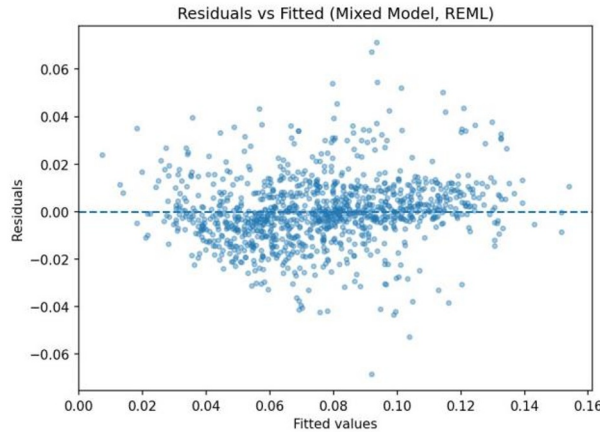


Figure 5 Residual-fitted value scatter plot.

According to figure 5, the residuals are evenly distributed around zero, with no obvious heteroscedasticity.

2.5 Robustness Analysis

To verify the robustness of the conclusions, the Y concentration is logarithmically transformed and the model is re-established. The results show that the direction and significance of the effects of gestational age and BMI remain consistent, the ICC and R^2 are basically unchanged, and the normality of the residuals is improved. Figure 6 shows the fitted curve of the logY model, which is consistent with the trend of the original model, indicating that the conclusions are robust and reliable.

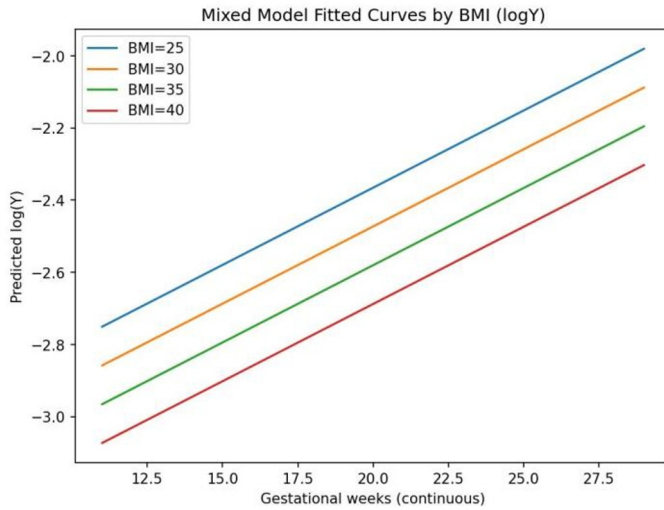


Figure 6 Fitted curves of the mixed-effects model after logarithmic transformation

3. Conclusions

Through mixed-effects modeling of factors influencing fetal cell-free DNA concentration, this study successfully quantified the effects of gestational age and BMI on Y chromosome concentration. The findings confirm that gestational age exerts a significant positive influence, while BMI exerts a significant negative influence. Given that over one-third of pregnant women underwent repeat testing, this study employed a linear mixed-effects model to account for individual variation. Significant inter-individual variability was observed, validating the necessity of this modeling approach and the robustness of the conclusions. These findings are highly consistent with previous research and possess strong statistical and biological plausibility.

However, this study has several limitations. First, some models still assume linear relationships, potentially underestimating the ability to capture abrupt or strongly nonlinear patterns in Y concentration changes over gestational age. Second, the analyzed samples predominantly involved high-BMI pregnant women, limiting data representativeness. Extending conclusions to low-BMI or other special populations may increase prediction errors by 10–15%, necessitating cautious extrapolation.

For future research, survival analysis models like KM/Cox/AFT could be adopted to more robustly address right-censored bias in time-to-achievement estimates. Additionally, incorporating monotonic splines or piecewise regression into curve estimation could enhance the interpretability of the $\mu\text{g}(t)$ curve. Most critically, replicating the entire workflow across multiple batches of data from different centers is essential to achieve multicenter external validation, ensuring the stability and generalizability of the detection strategy.

Authors contribution

All the authors, including Jinghe Che, Ziheng Zhang and Siyu Zhou, contributed equally to this work.

References

1. Schuurman P V L ,Koning D J H ,Meier E , et al. Clinical and economic impact of genome-wide non-invasive prenatal testing (NIPT) as a first-tier screening method compared to targeted NIPT and first-trimester combined testing: A modeling study.[J].*PLoS medicine*,2025,22(11):e1004790.
2. Perrot A ,Smart B H ,Klaiman N T , et al. Decision-making for termination of pregnancy following non-invasive prenatal testing: a qualitative exploration of french, english and German healthcare professionals' perceptions and concerns.[J].*Reproductive health*,2025,22(1):216.
3. Galeva S ,Stoilov B ,Uchikova E . Challenges and clinical implications of discordant non-invasive prenatal testing results: insights from two case studies.[J].*Folia medica*,2025,67(5).
4. Resta C ,Xiong R ,Sturrock S , et al. Non-invasive prenatal testing for the diagnosis of sickle cell disease in high-risk pregnancies: A systematic review and statistical summary of the current literature.[J].*European journal of obstetrics, gynecology, and reproductive biology*,2025,316114799.
5. Smart B H ,Johnston M ,Sands T M , et al. The Function of Non-invasive Prenatal Testing (NIPT) Request Forms in the Australian Context[J].*Health Care Analysis*,2025,(prepublish):1-25.
6. Owen A . "Do you want to know or not?" How prenatal providers manage clinical uncertainty related to chromosomal risk and noninvasive prenatal testing.[J].*Health (London, England : 1997)*,2025,13634593251377111.
7. Zhou W ,Liu F ,Li S , et al. Novel Algorithm for Monogenic Noninvasive Prenatal Testing With Highly Similar Parental Pathogenic Haplotypes: A Representative Case of Congenital Adrenal Hyperplasia Pedigree[J].*Human Mutation*,2025,2025(1):9990873-9990873.
8. Adusumalli R ,Banala R R . Advancements in prenatal diagnostics and the effects of EU regulatory frameworks, including the IVDR and MDR: A systematic review[J].*Egyptian Journal of Medical Human Genetics*,2025,26(1):164-164.
9. Yamamoto K ,Suzumori N ,Miura K , et al. Clinical Implications of Low Cell-Free DNA Fetal Fraction in Non-Invasive Prenatal Testing: A Retrospective Cohort Study of 40,716 Pregnancies.[J].*Prenatal diagnosis*,2025.
10. Cohen M S ,Chen M ,Sun L . Editorial: Advancements in prenatal diagnosis: from noninvasive prenatal tests to novel fetal imaging[J].*Frontiers in Medicine*,2025,121682161-1682161.