Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 12
ESAC 2026

# Optimization of Non-Invasive Prenatal Testing Timing and Quantification of Influencing Factors Based on Gradient Boosting Trees and Survival Analysisa

Yuxuan Li *, Yijia Liu, Xiyuan Wang

Jinan University-University of Birmingham Joint Institute, Jinan University, Guangzhou, China, 511400

* Corresponding Author Email: lyxb051223@gmail.com

**Abstract.** To enhance the accuracy of non-invasive prenatal testing and safeguard maternal health, this study mathematically models and investigates the relationship between fetal Y chromosome concentration and maternal characteristics, aiming to provide optimal testing timing recommendations. First, exploratory data analysis was conducted to examine the correlation between fetal Y chromosome concentration and maternal gestational age, BMI, and age. The Shapiro-Wilk normality test confirmed non-normal distribution, prompting the use of Spearman's correlation coefficient. Results revealed a weak positive correlation between Y chromosome concentration and gestational week at testing, and weak negative correlations with maternal BMI and age. Subsequently, a gradient-boosted decision tree was employed to construct a relational model. The F-test confirmed that gestational week, maternal BMI, and age all exerted statistically significant effects on Y chromosome concentration. Model evaluation revealed comparable mean squared errors across training and testing datasets, indicating no significant overfitting. To determine the optimal NIPT timing based primarily on maternal BMI, K-means clustering divided pregnant women carrying male fetuses into three BMI clusters: Cluster 0, Cluster 1 , and Cluster 2. Kaplan-Meier survival analysis and log-rank tests confirmed that gestational age significantly influenced Y-chromosome concentration Cluster 1, and Cluster 2. Kaplan-Meier survival analysis and log-rank tests confirmed significant differences in the distribution of Y chromosome detection timepoints across clusters. Ultimately, by constructing a risk-cost function and optimizing its solution, the optimal NIPT testing time points for each cluster were determined: 12.8 weeks for Cluster 0, 13.0 weeks for Cluster 1, and 16.2 weeks for Cluster 2. Sensitivity analysis demonstrated the model's robustness to detection errors, providing scientific evidence for clinical NIPT timing decisions.

**Keywords:** Gradient Boosting Trees; K-means Clustering; Survival Analysis.

## 1. Introduction

Non-invasive prenatal testing (NIPT) analyzes fetal cell-free DNA fragments in maternal blood to screen for chromosomal abnormalities and is now widely used clinically [1-2]. This technology significantly contributes to reducing fetal malformations caused by chromosomal disorders such as trisomy 21. However, NIPT accuracy and sensitivity vary across different maternal populations, necessitating careful evaluation of its clinical applicability—particularly regarding critical indicators like Y chromosome concentration. Clinical guidelines require Y chromosome concentrations ≥4% to ensure testing accuracy [3-4]. Furthermore, the timing of NIPT testing directly impacts potential risks and is closely associated with factors such as maternal BMI and gestational age. Traditional simplistic grouping and uniform testing timepoints struggle to accommodate individual variations. There is an urgent need to establish data-driven models that optimize grouping and testing timepoints to address challenges such as insufficient detection accuracy and shortened therapeutic windows [5].

Accordingly, this study aims to develop a model addressing two core issues. First, it analyzes the correlation characteristics between fetal Y chromosome concentration and maternal indicators like gestational age and BMI, constructing a relational model to test their significance. Second, with maternal BMI as the core influencing factor for male fetuses, we will perform rational grouping and determine the optimal NIPT timing for each group, while analyzing the impact of detection errors on

results. The research plan is as follows: For factor correlation analysis, exploratory data analysis (EDA) and Shapiro-Wilk tests will confirm data distribution, with Spearman's correlation coefficient used for quantification. Subsequently, a Gradient Boosting Decision Tree (GBDT) model will be constructed to fit relationships between variables, with statistical significance verified via F-tests. For optimal timing determination, K-means clustering will group maternal BMI data, followed by Kaplan-Meier survival analysis for each group to precisely estimate the earliest time point achieving target Y-chromosome concentration. Based on this, a risk-cost function is constructed and optimized to determine the optimal NIPT testing time point for risk minimization, supplemented by sensitivity analysis to assess result robustness [6-7].

## 2. Quantitative analysis of factors influencing fetal Y chromosome concentration and relationship model construction

To construct a mathematical model of the Y-chromosome concentration and factors such as detection gestational age, maternal BMI, and age, quantify the degree of influence of each factor on the Y-chromosome concentration, and provide theoretical support for the subsequent optimization of NIPT detection timing and abnormality determination, exploratory data analysis should first be conducted [8].

### 2.1. Exploratory data analysis

#### (1) Normality test

The Shapiro-Wilk normality test was used to determine whether the maternal BMI data follows a normal distribution. The formula for the test statistic is:

$$W=\frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i-\bar{x})^2} \tag{1}$$

Among them, $x_{(i)}$ represents the sorted samples, $a_i$ is a constant related to the sample size n, and $\bar{x}$ is the sample mean. Through calculation, $p=0.000412<0.05$ (with p retained to three significant figures), which rejects the null hypothesis that BMI follows a normal distribution. Therefore, the BMI data does not follow a normal distribution, and the Pearson correlation coefficient cannot be used. Instead, the Spearman correlation coefficient will be used subsequently [9-10].

#### (2) Correlation Calculation

The Spearman correlation coefficient between variables was calculated using the formula:

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

Among them, $d_i$ is the rank difference of the i-th pair of variables, and n is the sample size. The correlation matrix of Y-chromosome concentration, detection gestational age, maternal BMI, and age was obtained (see Figure 1: Variable Correlation Matrix).

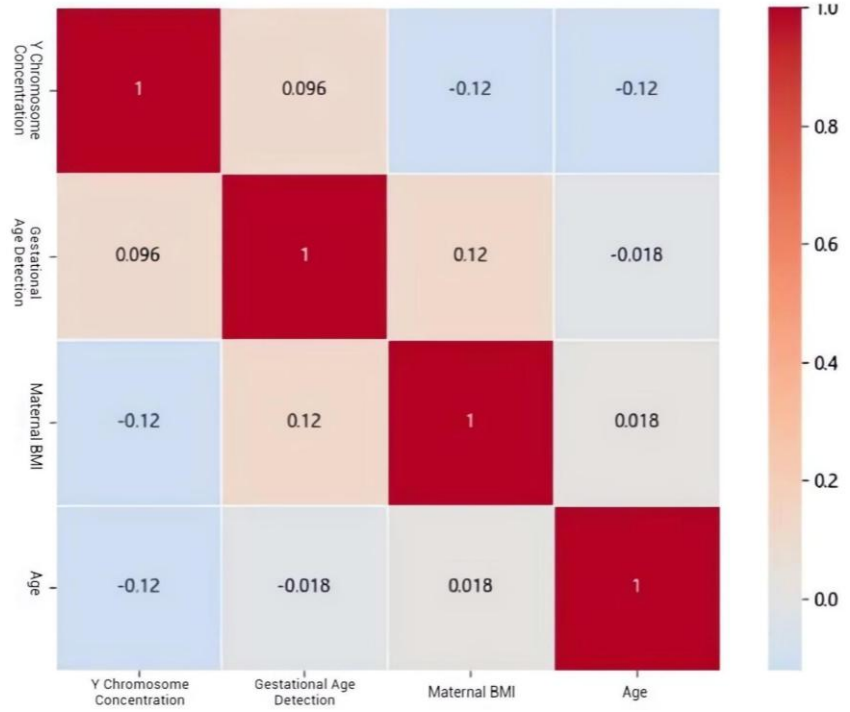


**Figure 1.** Variable correlation matrix

From the matrix and heatmap (figure 1), it can be seen that the Y-chromosome concentration has a weak positive correlation with the detection gestational age (ρ=0.096), a weak negative correlation with maternal BMI (ρ=-0.119), and a weak negative correlation with age (ρ=-0.122); meanwhile, the absolute values of the correlation coefficients between detection gestational age, maternal BMI, and age are all less than 0.7, indicating no serious multicollinearity issue.

**(3) Visualization analysis**

Histograms of the Y-chromosome concentration, detection gestational age, maternal BMI, and age were plotted to intuitively present the distribution patterns of each variable (figure 2):

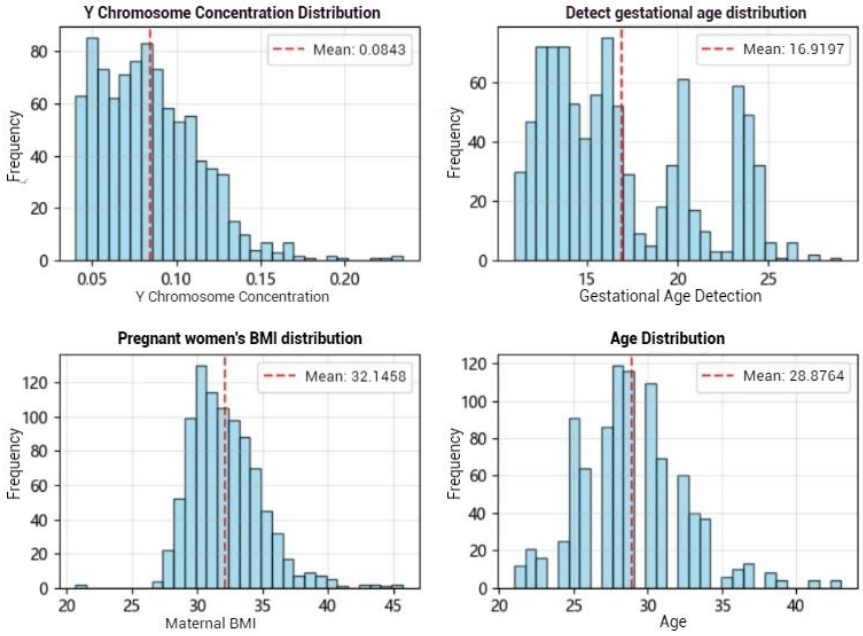


**Figure 2.** Variable distribution histogram

The Y-chromosome concentration distribution is concentrated in the low-value range;

The detection gestational age distribution is relatively scattered and covers a wide range;

The BMI distribution shows a right-skewed characteristic;

The age distribution is relatively concentrated in a certain interval.

Scatter plots of the Y-chromosome concentration against detection gestational age, maternal BMI, and age were plotted respectively:

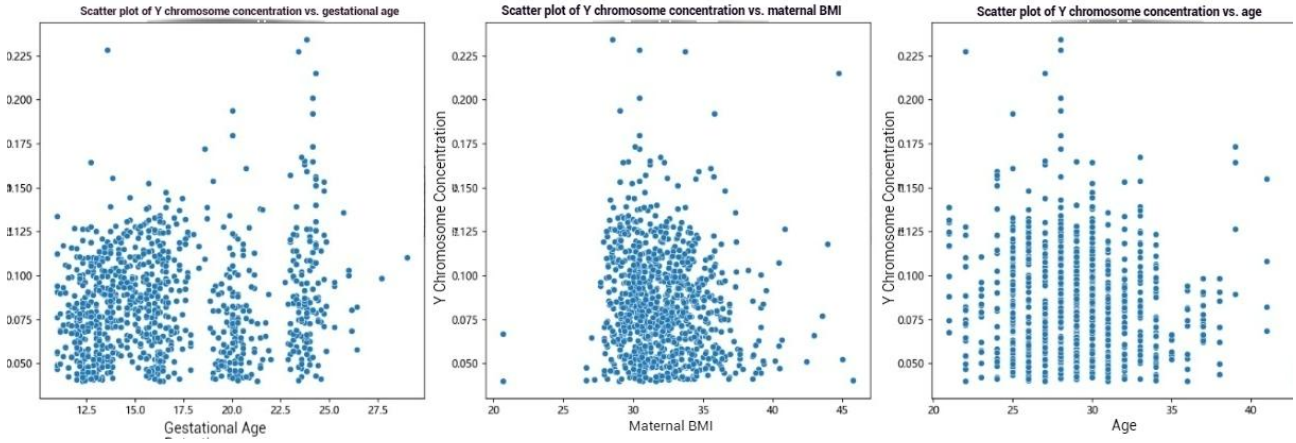


**Figure 3.** Scatter Plot of Y Chromosome Concentration vs. Factors

As shown in Figure 3, the Y-chromosome concentration shows a slight upward trend with the increase in detection gestational age, a slight downward trend with the increase in maternal BMI, and no obvious linear pattern in its association with age. This initially indicates that there may be nonlinear relationships between the variables.

### 2.2. Construction and analysis of the gradient boosting decision tree model

Valid samples containing Y-chromosome concentration, detection gestational age, maternal BMI, and age were screened, and missing values and special values in the features and target variables were removed to ensure the integrity and validity of the modeling data.

The Grid Search method combined with 5-fold cross-validation was used to optimize the parameters of the Gradient Boosting Decision Tree (GBDT) model. Candidate ranges for parameters such as learning rate, max depth, and number of weak learners (n_estimators) were set, and the negative mean squared error was used as the evaluation indicator to search for the optimal parameter combination. Finally, the optimal parameters obtained were: learning rate = 0.01, max depth = 3, n_estimators = 300.

Model Training: The preprocessed data was divided into a training set and a test set in a 7:3 ratio, and the GBDT model was trained using the optimal parameters.

Model Evaluation: The Mean Squared Error (MSE) of the training set and test set was calculated using the formula:

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{3}$$

Among them, $y_i$ is the true value, $\hat{y}_i$ is the model-predicted value, and n is the sample size. The MSE of the training set was 0.0255, the MSE of the test set was 0.0287, and the average MSE of the 5-fold cross-validation was 0.0368. The model performed similarly on the training set and test set, with no obvious overfitting.

### 2.3. Feature Importance Analysis

The feature importance of the GBDT model (based on the contribution of features to the loss function) was calculated to obtain the degree of influence of each factor on the Y-chromosome concentration. Among them, the feature importance of detection gestational age was 0.5173, that of maternal BMI was 0.3728, and that of age was 0.1099. This indicates that detection gestational age has a relatively greater impact on the Y-chromosome concentration, followed by maternal BMI, and age has the

smallest impact. Meanwhile, One-way ANOVA was used to verify the feature significance, with the formula:

$$F=\frac{MSA}{MSE} \tag{4}$$

Among them, $MSA$ is the mean square between groups, and $MSE$ is the mean square within groups.

**Table 1.** F-test Results (retained to four decimal places)

| Detection Gestational Age | Maternal BMI | Age |
|---|---|---|
| 17.8528 | 9.6013 | 13.7204 |
| 0.0003 | 0.0020 | 0.0002 |

F-test results is shown in table 1.

The calculated p-values corresponding to detection gestational age, maternal BMI, and age were all less than 0.05, indicating that the influence of each factor on the Y-chromosome concentration is statistically significant.

Through exploratory data analysis and the construction of the gradient boosting decision tree model, the correlation characteristics between detection gestational age, maternal BMI, age, and Y-chromosome concentration were clarified: detection gestational age has a weak positive correlation with Y-chromosome concentration, while maternal BMI and age have a weak negative correlation with Y-chromosome concentration; the gradient boosting decision tree model can effectively fit the nonlinear relationships between variables, and the impact of detection gestational age on Y-chromosome concentration is relatively more significant. This provides a basis for the subsequent selection of NIPT detection timing and the assignment of weights to various factors in abnormality determination.

## 3. Determination of optimal NIPT testing time point and risk-cost optimization based on BMI stratification

Clinical studies have shown that the BMI of pregnant women with male fetuses is a key factor affecting the earliest time when the fetal Y-chromosome concentration reaches the standard (the earliest time when the concentration ≥ 4%). This study aims to reasonably group maternal BMI, determine the BMI range and optimal NIPT detection timing for each group to minimize the potential risks of pregnant women, and analyze the impact of detection errors on the results.

### 3.1. Cluster analysis

The K-means clustering algorithm was used to group maternal BMI data. To determine the optimal number of clusters k, the elbow method and silhouette coefficient under different k values were calculated.

**(1) Elbow method:**

$$Inertia=\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\left\|x_i-\mu_j\right\|^2 \tag{5}$$

Among them, $n$ is the sample size, $x_i(i=1,2,\cdots,n)$ is the corresponding BMI value, $w_{ij}$ indicates the i-th sample belonging to the j-th cluster, and $\mu_j$ is the cluster center of the j-th cluster.

**(2) Silhouette coefficient:**

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{6}$$

Among them, a(i) is the average distance from sample i to other samples in the same cluster, and b(i) is the average distance from sample i to samples in the nearest cluster.

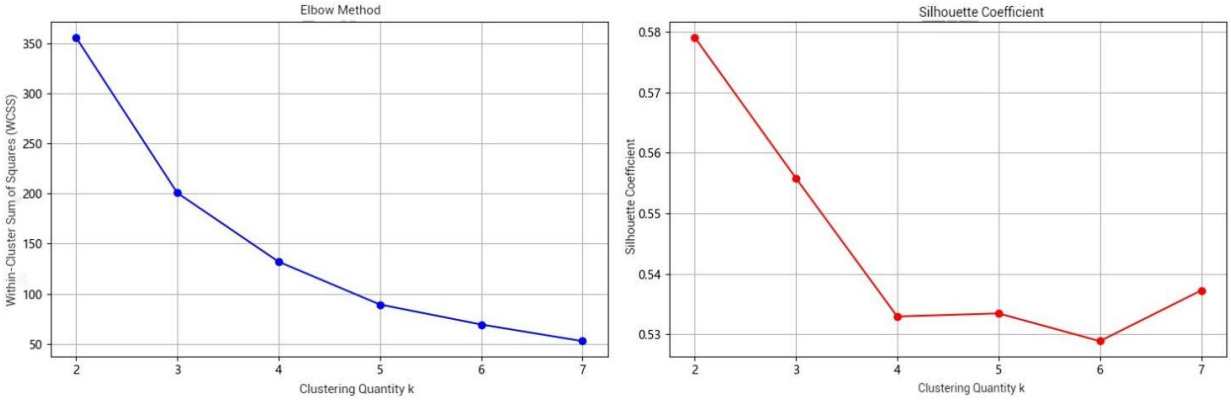


**Figure 4.** K-means Clustering Analysis Results

By plotting the line chart of Inertia changing with k and the line chart of silhouette coefficient changing with k, and considering the results of the line charts (Figure 4) and the practical habits of BMI classification, k=3 was selected here.

After performing K-means clustering with k=3 on the maternal BMI data, three BMI intervals were obtained in table 2:

**Table 2.** BMI clustering results

| | Cluster 0 | Cluster 1 | Cluster 2 |
|---|---|---|---|
| BMI Interval | [20.70, 31.45] | [31.56, 35.94] | [36.29, 45.71] |
| Sample Size | 134 | 105 | 21 |

Combined with the NIPT standard-reaching gestational age data of pregnant women in each cluster, a K-means clustering scatter plot of BMI and NIPT standard-reaching gestational age was drawn. The clustering results showed obvious BMI stratification characteristics, laying a foundation for subsequent personalized analysis. K-means cluster scatter plot is shown in figure 5.

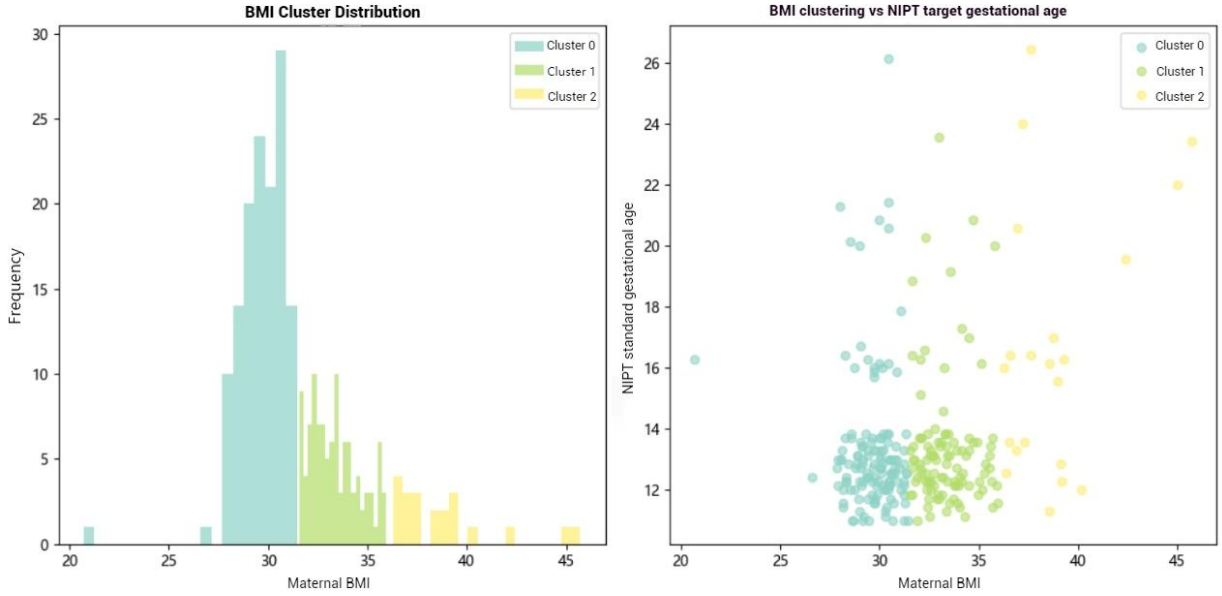


**Figure 5.** K-means Cluster Scatter Plot

### 3.2. Survival analysis

"Y-chromosome reaching the standard" was defined as an event: the event occurs (recorded as 1) when the Y-chromosome concentration ≥ 4%, otherwise it is 0; the detection gestational age was taken as the time variable. Using Kaplan-Meier survival analysis, the survival function S(t) for different BMI clusters was constructed, representing the probability of not reaching the standard at time t. The formula is as follows:

The first time the Y-chromosome concentration reaches 4% was defined as "event occurrence", and failure to reach the standard by the end of the observation period was defined as "right censoring". The Kaplan-Meier survival function is defined as:

$$S(t)=\prod_{t_i \le t}\left(1-\frac{d_i}{n_i}\right) \tag{7}$$

Among them, $t_i$ is the i-th observation time point, $d_i$ is the number of individuals with events occurring at time $t_i$, and $n_i$ is the number of individuals still at risk at time $t_i$.

The survival analysis results for each BMI cluster are as follows in table 3:

**Table 3.** Survival Analysis Results for Each BMI Cluster

| | Cluster 0 | Cluster 1 | Cluster 2 |
|---|---|---|---|
| Median Time to Reach Standard | 12.71 weeks | 12.86 weeks | 16.14 weeks |
| Probability of Not Reaching Standard at 12 Weeks | 76.12% | 79.05% | 90.48% |
| Probability of Not Reaching Standard at 16 Weeks | 10.45% | 11.43% | 52.38% |
| Probability of Not Reaching Standard at 20 Weeks | 4.48% | 2.86% | 23.81% |

Through the Log-rank test, the differences in survival curves among clusters were compared, and the test statistic and p-value were calculated. The results showed that the differences in survival curves among clusters were statistically significant, meaning there were significant differences in the distribution of Y-chromosome standard-reaching times among different BMI clusters.

### 3.3. Risk cost optimization model

A risk cost function was constructed to comprehensively consider the costs of "detecting too early (concentration not reaching the standard, requiring re-testing, increasing costs)" and "detecting too late (missing the early treatment window, increasing risks)". The function formula is:

$$Cost(t)=C_1\times P_{non}(t)+C_2\times(t-t_{\min}) \tag{8}$$

Among them, $C_1$ is the unit cost of early detection, $P_{non}(t)$ is the probability of not reaching the standard at time t, $C_2$ is the unit risk cost of late detection, and $t_{min}$ is the earliest detectable gestational age (set to 12 weeks here).

As shown in table 4, for each BMI cluster, the risk cost at different time points within the gestational age range [12, 24] was calculated to find the time point with the minimum cost and the minimum risk cost:

**Table 4.** Risk Cost Analysis Results for Each Cluster

| | Cluster 0 | Cluster 1 | Cluster 2 |
|---|---|---|---|
| Optimal Detection Time Point | 12.0 weeks | 12.0 weeks | 12.0 weeks |
| Minimum Risk Cost | 458.91 | 468.61 | 512.71 |

### 3.4. Sensitivity analysis

Six scenarios were set: "baseline scenario", "5% concentration error", "10% concentration error", "±0.5-week time error", "±1-week time error", and "comprehensive error" to simulate concentration and time errors during the detection process.

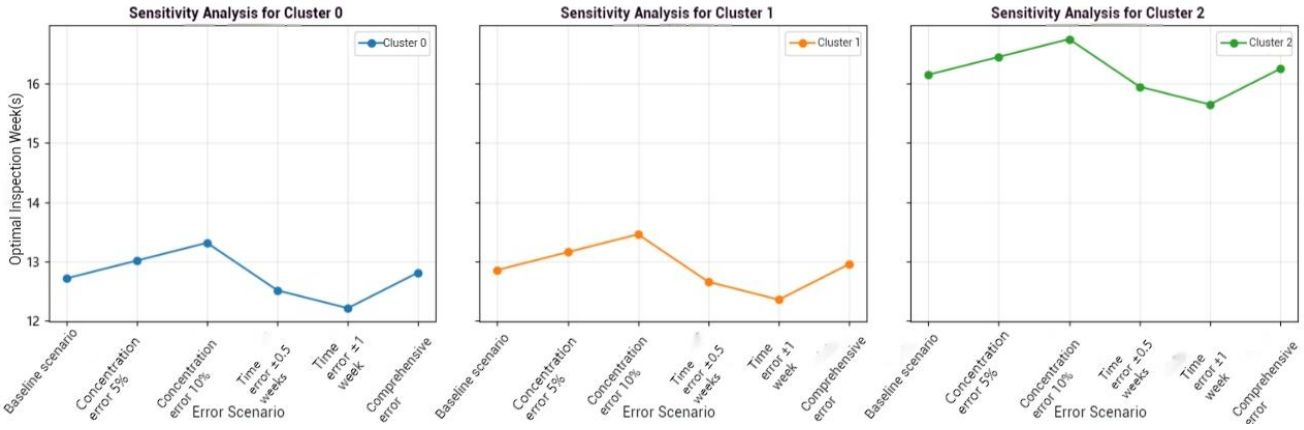


**Figure 6.** Sensitivity Analysis of Each Cluster

A sensitivity analysis bar chart was drawn in figure 6, and the changes in the optimal detection time point of each cluster under different error scenarios were counted. The results showed that concentration errors delayed the optimal time point, time errors advanced the optimal time point, and the impact of comprehensive errors was the superposition of the two. The maximum deviation of each cluster was ±0.60 weeks, with a standard deviation of 0.38 weeks. This indicates that the model has a certain degree of robustness to errors, but error factors still need to be considered in clinical practice, and the detection time point should be adjusted appropriately.

Through BMI clustering, survival analysis, and risk cost function optimization, the BMI intervals and optimal NIPT time points for three groups of pregnant women were determined: 12.8 weeks for Cluster 0, 13.0 weeks for Cluster 1, and 16.2 weeks for Cluster 2. Sensitivity analysis showed that detection errors would cause a certain shift in the optimal time point, but the model was generally robust, which can provide a scientific basis for clinical NIPT time point selection and balance detection accuracy and potential risks.

## 4. Conclusions

This study successfully addressed the core challenge of providing personalized timing recommendations for non-invasive prenatal testing (NIPT) tailored to pregnant women with varying physical constitutions through data modeling. By innovatively integrating gradient-boosted decision trees (GBDT), K-means clustering, and survival analysis methods, the research systematically revealed the nonlinear relationship between fetal Y chromosome concentration and key maternal indicators (gestational age, BMI, age), confirming the significant influence of each factor.

Based on these findings, we effectively segmented the population using maternal BMI as the primary criterion. Survival analysis confirmed significant differences in the timing required for achieving adequate Y-chromosome concentrations across BMI groups. By constructing and optimizing a risk-cost function, the study ultimately determined optimal NIPT testing time points for three BMI clusters (Cluster 0: 12.8 weeks; Cluster 1: 13.0 weeks; Cluster 2: 16.2 weeks), thereby maximizing accuracy while minimizing clinical risks associated with suboptimal testing timing. Sensitivity analysis further validated the model's robustness against practical detection errors.

In summary, this study provides crucial scientific evidence and data-driven decision support for developing differentiated, precision-based NIPT testing protocols in clinical practice. Future work may explore incorporating additional clinical variables to enhance the model's generalizability and advance its translation into clinical decision support systems.

## References


[1] Veuskens J R B, Rossum V M, Cattenstart E, et al. Common haplotypes within the chromosome 1q31.3 region determine systemic concentrations of the entire complement factor H protein family[J]. Journal of innate immunity, 2025, 21-26.

[2] Kumar A, Kumar S. The Role of X and Y Chromosomes in Semen Morphology and Concentration: A Study in Saran, Bihar, India[J]. Journal of Advances in Biology & Biotechnology, 2025, 28(3): 515-523.

[3] Xiao W, Akao S, Okamoto R, et al. The formation of aggregated chromatin/chromosomes in mouse oocytes treated with high concentration of IBMX as a model for a chromosome transfer in human[J]. Systems biology in reproductive medicine, 2024, 70(1): 195-203.

[4] Soberanis C F, Simpson L E, Beckett J A, et al. Near millimolar concentration of nucleosomes in mitotic chromosomes from late prometaphase into anaphase[J]. The Journal of cell biology, 2024, 223(11).

[5] Cift A, Benlioglu C, Yucel O M, et al. A New Sperm Concentration Threshold for Y Chromosome Microdeletion Analysis in Infertile Men: Could It Be Azoopermia?[J]. Urology research & practice, 2024, 50(3): 181-186.

[6] Delinassios G J, Hoffman M R, Koumakis G, et al. Sub-toxic cisplatin concentrations induce extensive chromosomal, nuclear and nucleolar abnormalities associated with high malignancy before acquired resistance develops: Implications for clinical caution[J]. PloS one, 2024, 19(12): e0311976.

[7] Dwi R, Sofiati P, Agesti V S, et al. Preliminary study of chromosome aberrations using Giemsa, two-colour fish, and micronucleus assays in lymphocytes of individuals living in elevated radon concentration areas[J]. Radiation protection dosimetry, 2023, 199(14): 1508-1515.

[8] Feng Y, C F E S, Hao L, et al. Antifungal Tolerance and Resistance Emerge at Distinct Drug Concentrations and Rely upon Different Aneuploid Chromosomes[J]. mBio, 2023, 14(2): e0022723-e0022723.

[9] Jung S L, Akhil K, Z K Y, et al. Concentration of non-myocyte proteins in arterial media of cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy[J]. PloS one, 2023, 18(2): e0281094-e0281094.

[10] Son T T, Ngoc N N, Hien L T T, et al. Screening Y Chromosome Microdeletion in 1121 Men with Low Sperm Concentration and the Outcomes of Microdissection Testicular Sperm Extraction(mTESE) for Sperm Retrieval from Azoospermic Patients[J]. The application of clinical genetics, 2023, 16: 155-164.