A Study on Factor Analysis and Timing Optimization for NIPT in High-BMI Pregnant Women Based on Linear Mixed-Effects Models and Dynamic Programming

Yuhang Zhao, Zhiqiang Zhang *, Ruohan Zhao

School of Computer Science, Qufu Normal University, Qufu, China, 273165

* Corresponding Author Email: 13181717139@163.com

Abstract. Non-invasive prenatal testing (NIPT) is a pivotal prenatal screening tool, yet its accuracy for male fetuses depends on achieving a sufficient fetal Y chromosome concentration (fYCC), which is influenced by individual factors like maternal BMI and gestational age (GA). This study aims to resolve the clinical conflict between early detection and reliable results by optimizing the NIPT timing for high-BMI pregnant women. First, a linear mixed-effects model (LMM) was employed to identify key factors affecting fYCC, revealing significant positive correlations with GA, X chromosome concentration, and number of blood draws, and a negative correlation with BMI. Building on these findings, a personalized BMI grouping strategy was developed using a hierarchical risk mapping function optimized via a dynamic programming (DP) algorithm. This approach defined five optimal BMI intervals and determined the recommended testing time as the 90th percentile GA for fYCC attainment within each group. Error analysis demonstrated the scheme's high robustness with a detection error (σ) of $\leqslant$1%. The proposed model fusion framework provides a data-driven solution for personalized NIPT scheduling, effectively minimizing risk for high-BMI individuals.

Keywords: Non-Invasive Prenatal Testing, Linear Mixed-Effects Model, Dynamic Programming.

1. Introduction

Non-invasive prenatal testing (NIPT) has emerged as a rapidly advancing core prenatal diagnostic technology [1,2]. By analyzing cell-free fetal DNA in maternal peripheral blood, NIPT accurately assesses fetal risk for trisomy 21, 18, and 13. With its safety, accuracy, and convenience, NIPT has become an essential clinical screening tool [3]. However, NIPT accuracy relies on strict conditions: for male fetuses, Y chromosome concentration must exceed 4%, and testing is only feasible between 10 and 25 weeks of gestation. Concurrently, the timing of detecting fetal abnormalities is closely linked to risk—detection before 12 weeks carries the lowest risk, while after 28 weeks, the risk becomes extremely high. Thus, clinical practice must balance the conflicting priorities of "early detection" and "reliable results."

Reviews of prior research in this field indicate that a core challenge in current NIPT strategies lies in the neglect of individual variability. Traditional fixed BMI grouping and uniform testing timepoints fail to adequately account for individual variations such as maternal BMI and gestational age [4,5]. While prior studies have confirmed significant correlations between fetal Y chromosome concentration and both gestational age and BMI [6,7], modeling often faces methodological limitations. For instance, most studies employ traditional linear regression models, which struggle to effectively handle boundary issues when Y chromosome concentration is treated as a ratio variable (e.g., concentrations near 0% or 100%). Furthermore, these models lack adequate statistical control for the inherent structural complexity of repeated measurements from the same pregnant woman (i.e., intra-subject correlation) [8,9]. Furthermore, in risk quantification, prior studies have largely remained at the level of qualitative descriptions of "early, mid, and late" risk periods or employed simplified grouping comparison methods. There has been a lack of a quantitative, optimizable decision framework capable of integrating "concentration attainment time" with "gestational week-related risk" [10]. This has resulted in existing clinical guidelines still falling short in providing personalized guidance for high-risk populations, particularly pregnant women with high BMI.

Addressing these research gaps, this study's marginal contributions are primarily reflected in the following three aspects: First, in factor analysis, this study employs multivariate association analysis based on linear mixed-effects models (LMM), incorporating Smithson-Verkuilen correction and logit transformation to handle ratio data. This approach robustly identifies key factors influencing fetal Y chromosome concentration attainment (such as gestational age, BMI, and X chromosome concentration). Second, in strategy optimization, this study innovatively combines hierarchical risk mapping functions with dynamic programming (DP) algorithms to construct a personalized BMI grouping model targeting risk minimization, overcoming the limitations of traditional fixed grouping. Finally, for protocol validation, error propagation analysis incorporating detection error (σ) was employed to systematically assess the robustness of the optimized strategy, providing quantitative evidence for clinical reliability. Through this multi-model fusion approach, this study aims to deliver data-driven, personalized NIPT timing optimization solutions addressing the core contradiction of "delayed target attainment → increased risk" in high-BMI pregnant women.

2. Linear Mixed-Effects Model Analysis of Factors Influencing Fetal Y Chromosome Concentration

2.1. Problem Analysis

This part is to explore the correlation characteristics between the Y-chromosome concentration of male fetuses and maternal indicators, and to construct a reliable relationship model. First, invalid data such as female fetus samples and outliers should be eliminated through data preprocessing and composite screening (setting a minimum effect threshold p, combining Spearman rank correlation and Benjamini–Hochberg (FDR) correction) to accurately identify the target variable (male fetal Y-chromosome concentration, $V\in(0,1)$), main variables (gestational age, BMI), and candidate covariates (X-chromosome concentration, number of blood draws for testing), so as to avoid model bias and dimensional disaster caused by irrelevant variables. Since the Y-chromosome concentration is proportional data within the interval (0,1), direct modeling is likely to violate the normality assumption. Therefore, the Smithson–Verkuilen correction is used to eliminate boundary bias, and then the Logit transformation is performed to meet the premise of the linear model. Subsequently, a linear mixed-effects model is constructed: first, a baseline model including gestational age, BMI, and their interaction terms is set, and then covariates are added to form a full model. The likelihood ratio test (LRT) verifies that the full model is superior. Finally, the direction of effects is clarified through correlation visualization [7].

2.2. Model Establishment

In this problem, we need to establish a mathematical model to calculate the relationship between the Y-chromosome concentration of male fetuses and factors such as gestational age and BMI. Since multiple factors need to be analyzed for correlation and their significance needs to be tested, and the Y-chromosome concentration of male fetuses is a proportional variable with a natural constraint interval [0,1], the traditional linear regression model cannot handle the characteristics of this bounded response variable. Therefore, we use the Spearman correlation analysis method and Logit transformation for modeling, and finally test through LRT.

First, data preprocessing is carried out, and then the Spearman rank correlation coefficient between each remaining variable and the Y-chromosome concentration is calculated using the formula:

$$\rho_{S,j}=1-\frac{6\sum_{i=1}^{n_j} d_{j,i}^2}{n_j\left(n_j^2-1\right)},\ d_{j,i}=R_{X_j,i}-R_{Y,i} \tag{1}$$

Then, multiple comparison correction (FDR) is performed to obtain the q-value. To prevent variables with excessively low correlation from entering the subsequent model, we define a minimum effect threshold [8]:

$$|\rho| \geq \rho_{min} \tag{2}$$

Filtering criteria: $|\rho| \geq 0.25$ and $q < 0.05$ (adjustable)

Table 1. Top 5 Correlation Coefficients Obtained by the Spearman Method

Rank	Variable	Spearman ρ
0	X-chromosome Concentration	0.474446
1	Number of Blood Draws for Testing	0.322997
2	Z-value of Chromosome 18	0.178163
3	Weight	0.166970
4	Proportion Aligned to the Reference Genome	0.159432

Top 5 Correlation Coefficients Obtained by the Spearman Method is shown in table 1. After initial screening, the relevant variables obtained are X-chromosome concentration and number of blood draws for testing. The initial screening shows that multiple variables have a moderate level of rank correlation with the Y-chromosome concentration; the core biological correlations are concentrated in gestational age (positive) and BMI (mostly showing a negative trend).

Meanwhile, to handle the proportion within (0,1) and improve variance stability, the data should first be truncated and then transformed using the Logit transformation formula:

$$V_{ij}^{*}=\text{clip}\left(V_{ij};\varepsilon,1-\varepsilon\right) \tag{3}$$

$$Y_{ij}=\text{logit}\left(V_{ij}^{*}\right)=\log\frac{V_{ij}^{*}}{1-V_{ij}^{*}} \tag{4}$$

Then, a baseline model (Baseline) for BMI and gestational age is established, expressed in a linear form [9-10]:

$$Y_{ij}=B_0+B_1J_{ij}+B_2K_i+B_3\left(J_{ij}K_i\right)+b_i+\varepsilon_{ij},\ \varepsilon_{ij}\sim N(0,\sigma^2) \tag{5}$$

Then, two covariates obtained from the Spearman initial screening—X-chromosome concentration and number of blood draws—are added to the baseline model to construct a full model (Full):

$$Y_{ij}=B_0+B_1J_{ij}+B_2K_i+B_3\left(J_{ij}K_i\right)+\gamma_1 Xchr_{ij}+\gamma_2 D_{ij}+b_i+\varepsilon_{ij} \tag{6}$$

Then, the likelihood ratio test (LRT) is used for testing and comparison:

$$\chi^2=2(l_{Full}-l_{Base}) \tag{7}$$

2.3. Model Solution

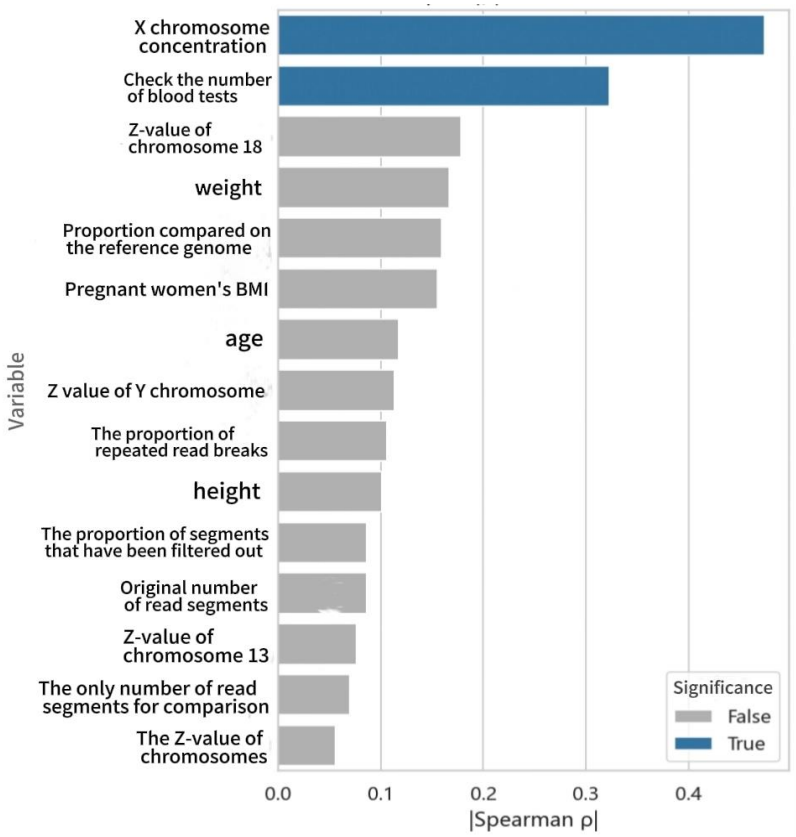

Figure 1. Bar chart of the top 15 |p| values

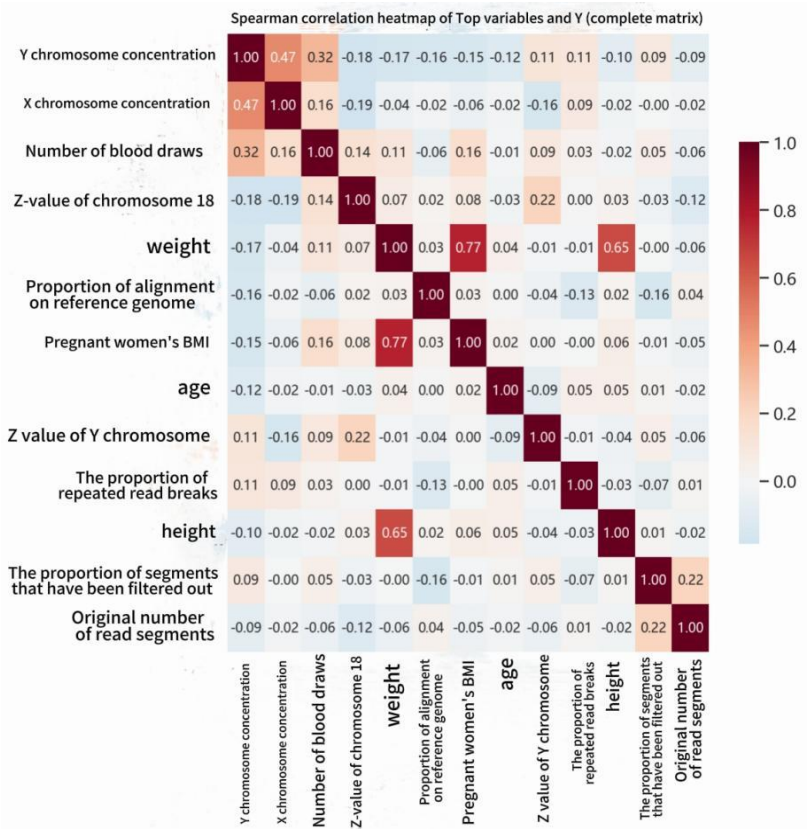

Figure 2. Spearman correlation heatmap between the top 15 variables and Y

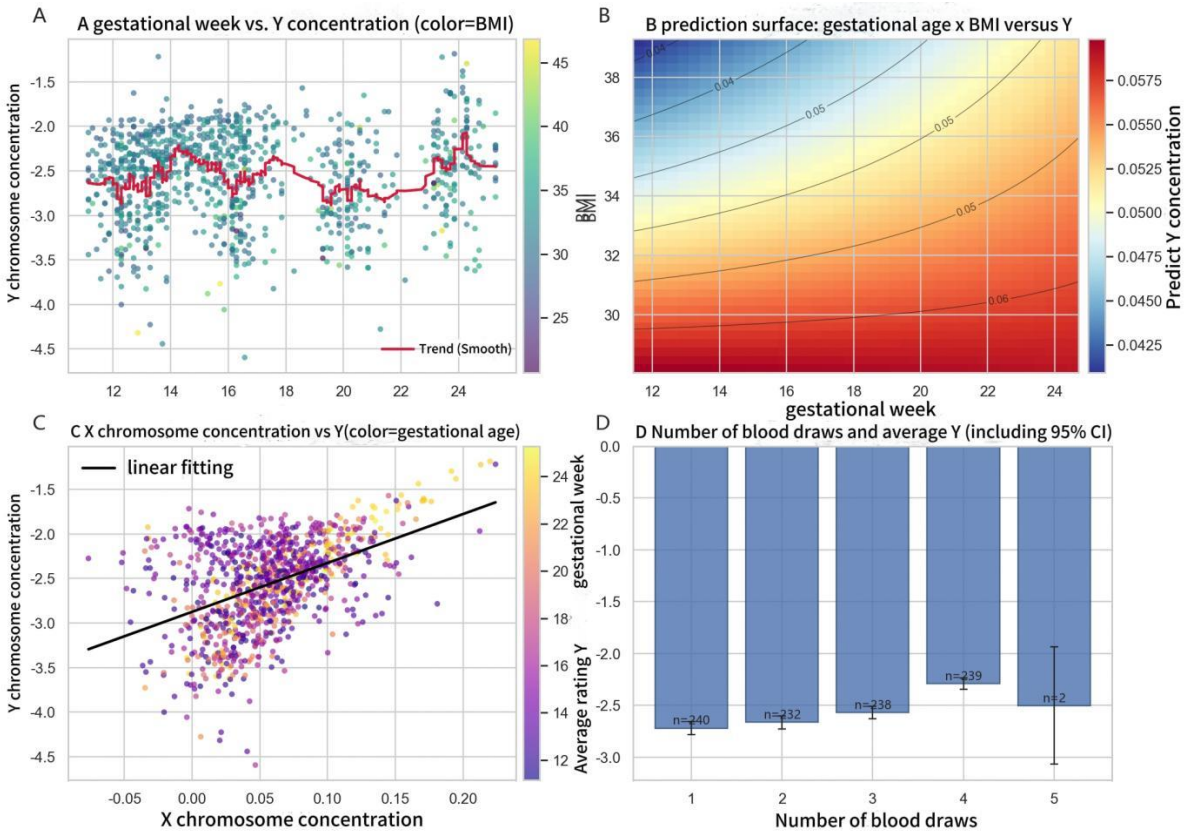

Figure 3. Composite diagram of fetal Y chromosome concentration and key indicators

From the visualization results in Figure 1 and Figure 2, it can be seen that the variables selected through screening are X-chromosome concentration and number of blood draws for testing.

From the results in Figure 3, it can be concluded that gestational age, X-chromosome concentration, and number of blood draws are positively correlated with the Y-chromosome concentration, while BMI is negatively correlated with the Y-chromosome concentration.

3. BMI Grouping and NIPT Optimal Timing Optimization Based on Risk Minimization and Dynamic Programming

3.1. Problem Analysis

In the analysis, first, the gestational age at which a pregnant woman's Y-chromosome concentration first reaches the standard is extracted, and the "accuracy feasible region" is defined to avoid sacrificing accuracy. Then, after transforming and sorting BMI, constraints of "sample size per group ≥12 and minimum BMI range ≥2.5" are set to balance the statistical validity and clinical operability of grouping. Subsequently, the qualitative risks of low (≤12 weeks), high (13-27 weeks), and extremely high (≥28 weeks) are converted into a stepwise cost function, and a comprehensive cost that can be optimized is formed through min-max normalization. The dynamic programming algorithm is used to find the global optimal solution, and finally, the BMI is divided into 5 intervals. The 90th percentile of the standard-reaching gestational age is selected as the optimal time point, taking into account both the standard-reaching coverage rate of 90% of pregnant women and the early stage of testing. Finally, detection errors are simulated with σ=0.5%, 1%, and 2%, and it is verified that the scheme is robust when $\sigma \leq 1\%$, while σ=2% will lead to a significant increase in the risk of the high-BMI group. Overall, following the logic of "defining the feasible region → designing constraints → quantifying risks → global optimization → selecting time points → verifying errors", the goal of risk minimization is achieved.

3.2. Model Establishment

In this problem, with the goal of "minimizing risk", the gestational age at which a pregnant woman's Y-chromosome concentration first reaches the standard (≥4%) is extracted. Then, BMI is screened and sorted for grouping, and the optimal NIPT time point for each group is calculated. Therefore, it is necessary to establish a hierarchical risk function, then calculate the optimal risk solution, use the dynamic programming algorithm for backtracking to solve, and finally introduce σ to detect and analyze the impact of errors on the results.

First, the samples of pregnant women with male fetuses are sorted by BMI and divided into $K=5$ continuous intervals. With the goal of "minimizing risk", the gestational age at which a pregnant woman's Y-chromosome concentration first reaches the standard (≥4%) is extracted.

Set the recommended gestational age (q-th percentile) within each group. For any interval $S= [1, r]$:

$$W_S=Q_q\left(\{G_i^*:i\in S\}\right),\ q=0.90 \tag{8}$$

And the interval $S= [1, r]$ satisfies:

Sample size: $r-l+1$ (sample size per group) ≥ 12

Minimum BMI range ≥ 2

Then, a risk mapping function is constructed based on the content of the article:

$$R(S)=\begin{cases} r_1, & W_S\leq 12 \\ r_2, & 12<W_S<28 \\ r_3, & W_S\geq 28 \end{cases}\quad (r_1<r_2<r_3) \tag{9}$$

Then, min-max normalization is used to define the comprehensive cost (risk priority). First, in the set of all feasible intervals I:

$$R_{min}=\min_{S\in I} R(S),\ R_{max}=\max_{S\in I} R(S),\ W_{min}=\min_{S\in I} W_S,\ W_{max}=\max_{S\in I} W_S \tag{10}$$

Then normalization is performed:

$$\tilde{R}(S)=\frac{R(S)-R_{min}}{R_{max}-R_{min}},\ \tilde{W}(S)=\frac{W_S-W_{min}}{W_{max}-W_{min}} \tag{11}$$

Subsequently, the dynamic programming algorithm is used to divide the BMI into 5 intervals and calculate the minimum cost. Let $dp[k,j]$ represent the minimum total cost of dividing the first j samples into k segments:

$$dp(k,j)=\min_{i\in[k-1,j-1]}\{dp(k-1,i)+C([i+1,j])\} \tag{12}$$

Initialization: $dp(0,0)=0$, and the rest are $+\infty$, backtracking gives $0=r_0<r_1<\cdots<r_5=n$.

The 90th percentile of the standard-reaching gestational age in each group is obtained as the optimal time point:

Then, σ is introduced to quantify the detection error, and the error is visualized to analyze the impact of the detection error. Finally, the impact of the detection error on the results is obtained: when $\sigma\leq 1\%$, the BMI grouping is highly consistent with the recommended NIPT gestational age; when $\sigma=2\%$, the boundary of the high-BMI group and the recommended time point will shift backward, and the risk level will increase significantly.

3.3. Model Solution

Table 2. Calculation of the Optimal NIPT Time Point for Different BMI Values

Group	Index Range	BMI_min	BMI_max	BMI_Range	90% Recommended Gestational Age
1	0-21	20.703125	28.604765	7.901640	16.685714
2	22-111	28.612450	30.629353	2.016903	16.142857
3	112-173	30.646185	32.695042	2.048857	15.785714
4	174-223	32.716022	34.720883	2.004861	14.200000
5	224-259	34.722222	44.982699	10.260477	20.714286

From Table 2, the optimal BMI groupings are obtained as [20.70, 28.14), [28.17, 30.18), [30.18, 32.24), [32.31, 34.33), and [34.48, 46.87], with corresponding time points of 21.28, 16.85, 16.42, 14.00, and 20.85 weeks.

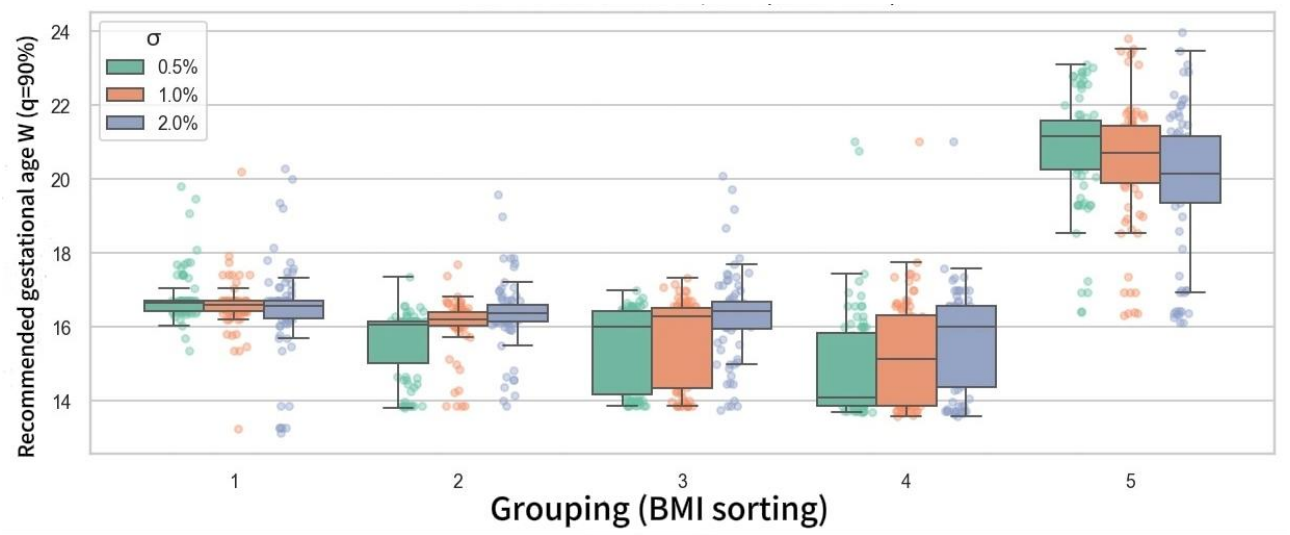

Figure 4. Recommended gestational week distribution across different groups

It can be seen from Figure 4 that as σ increases, the height of the box and the spread range of the whiskers increase, so the uncertainty increases. Finally, the impact of detection error on the results is obtained: when σ≤1%, the BMI grouping and the recommended NIPT gestational age are highly robust; when σ=2%, the boundary of the high-BMI group and the recommended time point will shift backward, and the risk level will increase significantly.

4. Conclusions

This study addresses the optimization of NIPT strategies for high-BMI pregnant women by employing a multi-model fusion approach to identify factors influencing fetal Y chromosome concentration and optimize testing timing. Regarding factor analysis, Spearman rank correlation screening combined with Smithson–Verkuilen correction and logit transformation of Y concentration successfully established a linear mixed-effects model (LMM). Key findings revealed significant positive effects of gestational age, X chromosome concentration, and number of blood draws, while BMI exhibited a significant negative effect. This provides data-driven evidence for subsequent BMI-based individualized strategy stratification. For timing optimization, the study adopted a risk minimization approach, constructing a hierarchical risk function and using dynamic programming to identify five optimal BMI intervals (e.g., [20.7, 28.6), [28.6, 30.6), etc.). Concurrently, the optimal

detection timing was determined as the gestational week corresponding to the 90th percentile for each group. Finally, error propagation analysis validated the high robustness of this optimization scheme when detection error $\sigma \leq 1\%$, effectively supporting its clinical application and resolving the core contradiction faced by high-BMI pregnant women: "delayed target attainment → increased risk.".

This study has several limitations. First, the generalizability of the constructed model requires further validation. Future plans include incorporating larger, multicenter prospective cohort data from diverse regions, ethnicities, and medical centers to assess the model's generalization capability. Second, the current risk function is primarily constructed based on theoretical analysis. Future work will involve collaborating with clinical experts to calibrate and optimize risk weights using real-world clinical outcome data (e.g., induction rates, maternal and fetal complications), making it more aligned with actual clinical decision-making scenarios. Additionally, while this study primarily focuses on macro-level factors like BMI and gestational age, future research will explore integrating multi-dimensional biomarkers (e.g., maternal plasma DNA fragmentomics profiles) to develop more precise predictive models. Finally, we will develop a user-friendly clinical decision support system prototype, transforming the optimization strategies proposed in this study into a practical tool to assist clinicians in designing personalized NIPT testing protocols.

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] Çift A, Benlioğlu C, Yücel Ö 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,2023,16155-164.