Research on Individualized Decision-Making for Timing of Non-Invasive Prenatal Testing (NIPT)Based on Chance-Constrained Programming and Multifactorial Probabilistic Modeling

Jiayi Wu[1], Jiaqi Zhang[1], Lu Li[1], Jiangang Zhang[2], Mingming Gong[2]
[1]School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, China
[2]iFlytek Co., Ltd., Hefei, Anhui, China

Abstract: To address the issues of test failure and elevated risk caused by insufficient fetal cell-free DNA proportions in non-invasive prenatal testing(NIPT)for pregnant women with high body mass index(BMI),this study designed a decision-making framework that integrates multifactorial probabilistic modeling and stochastic optimization to generate individualized precision testing protocols. First, multiple linear regression analysis of male fetus Y chromosome concentrations identified X chromosome concentration as the strongest predictive factor. Subsequently, a gestational age-dependent probability model for meeting target fetal DNA fraction thresholds was established, incorporating chance-constrained programming: optimizing optimal testing time points for high-BMI pregnant women by minimizing comprehensive risks under the constraint that achievement probability exceeds predefined confidence levels. Pregnant women were categorized into six risk groups, with recommended testing time windows ranging progressively from 16.0 weeks for low-risk groups to 20.0 weeks for high-risk groups. Finally, extending to abnormality detection in female fetuses, a two-stage methodology combining rule-based prioritization with logistic regression assistance was established through sample balancing and classifier comparison, improving diagnostic accuracy to 90.11%.This systematic approach provides a complete clinical solution for NIPT implementation, while its methodological framework demonstrates generalizability to similar decision-making scenarios involving uncertainty management.

Keywords: Non-Invasive Prenatal Testing; Chance-Constrained Programming; Probabilistic Modeling; Personalized Decision Making; Optimization of Testing Timing; Abnormality Determination

1. Introduction

Non-Invasive Prenatal Testing(NIPT)is one of the most widely applied technologies in current prenatal screening, playing a significant role in preventing birth defects and ensuring maternal and infant safety[1][1].

However, the detection accuracy of this technology largely depends on the proportion of fetal cell-free DNA(cfDNA)in maternal peripheral blood. When the cfDNA proportion falls below the critical threshold of 4%,the reliability of test results decreases significantly, leading to increased risks of false negatives and false positives[2].This issue is particularly pronounced in populations with high body mass index(BMI).Previous studies have demonstrated that elevated BMI substantially reduces fetal cfDNA concentration, causing delayed attainment of detection thresholds[3].Current clinical practices still commonly employ fixed gestational age or coarse grouping strategies for testing, which overlook inter-individual physiological variations. This approach increases the likelihood of test failure or delayed risk notification in pregnant women with high BMI, thereby narrowing the time window for subsequent diagnostic interventions.

Additionally, for female fetus samples lacking Y chromosome signals as reference markers, cfDNA proportion estimation typically relies on indirect indicators such as Z-scores and GC content[4]. Traditional threshold-based determination methods exhibit insufficient stability when confronting sample imbalance or sequencing biases, potentially resulting in missed diagnoses or misjudgments that challenge screening accuracy.

To address these challenges, existing research has primarily focused on statistical analyses of single factors or empirical grouping strategies,

lacking systematic modeling frameworks capable of integrating multi-source information, quantifying uncertainties, and providing individualized recommendations[5]. Consequently, current protocols face limitations in terms of accuracy, robustness, and clinical interpretability.

To overcome these limitations, this study proposes an individualized decision-making framework integrating chance-constrained programming with multi-factor probabilistic modeling. First, multiple linear regression analysis was conducted to identify key factors influencing Y chromosome concentration in male fetuses. Based on this, a probability model for achieving sufficient cfDNA proportion was established, marking the first application of chance-constrained programming methodology in optimizing NIPT timing. The model employs target confidence levels for fetal fraction adequacy as constraints while minimizing comprehensive maternal risks as objectives, thereby determining individualized optimal testing timepoints. Finally, the probabilistic modeling concept was extended to abnormality determination tasks for female fetuses, developing a two-stage classification approach combining rule-based criteria with model predictions[6].

2. Research Methods

The core objective of this study is to establish a data-driven mathematical model for guiding individualized decision-making in NIPT testing. The overall research framework is illustrated in Figure 1.

The core algorithm of this study primarily consists of three major modules.

The data preprocessing and feature engineering module is responsible for transforming raw multi-source data into a high-quality, information-rich feature matrix, laying the foundation for subsequent modeling. For male fetuses, the probabilistic modeling module first constructs a multiple linear regression model to analyze the mechanism of Y concentration effects, followed by establishing compliance probability models(e.g., logistic regression, random forest)to predict gestational-week-dependent FF compliance probabilities. For female fetuses, a two-stage abnormality determination model centered on logistic regression is developed: the first stage employs hard rules for rapid screening, while the second stage performs probabilistic discrimination for gray-zone samples.

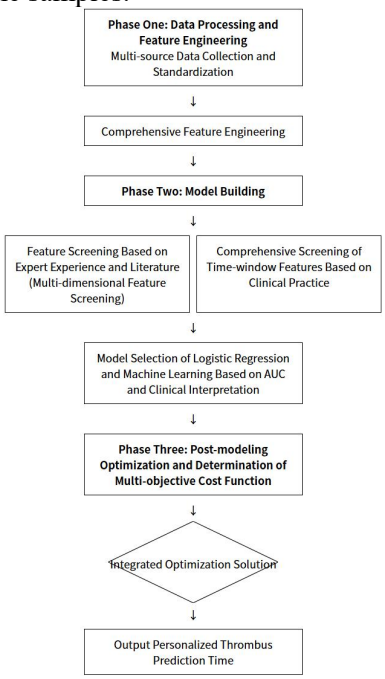

Figure 1. Flowchart of the Individualized Decision-making Framework for NIPT

The chance-constrained optimization decision module represents the innovative core of this paper. This study defines a comprehensive cost function for each BMI group and solves for the optimal detection time point that minimizes costs under specified chance constraints. This module translates probabilistic predictions into actionable clinical decisions. These components are interconnected to form an integrated analytical workflow spanning from data cleaning to recommended detection timing.

2.1 Data Preprocessing and Multidimensional Feature Engineering

2.1.1 Dataset and preprocessing

Prior to formal modeling, this study conducted unification and standardization of the original data. To address inconsistencies in gestational age recording formats(e.g.,"15w+3"),these were uniformly converted into numerical values(15.43 weeks)to facilitate subsequent calculations and analyses. The dataset used in this study originated from authentic NIPT clinical testing data collected from regional hospitals, comprising two subsets. The male fetus dataset contained 1082 testing records of pregnant women carrying male fetuses, with a total of 31 original variables. The target variable was defined as continuous Y chromosome concentration. As shown in Table 1.

Table 1. Partial Examples of the Male Fetus Dataset

Serial number	Maternal BMI.	Original read count	Alignment rate on the reference genome	Proportion of duplicate reads	Number of uniquely aligned reads	GC content	Z-score for chromosome 13
1	28.125	5040534	0.8067259	0.0276035	3845411	0.3992619	0.782096634
2	28.515625	3198810	0.8063927	0.02827083	2457402	0.3932988	0.692855699
3	28.515625	3848846	0.8038578	0.03259621	2926292	0.3998897	0.888701998

The female fetus dataset comprises 605 examination records from 147 pregnant women with female fetuses, including 67 abnormal cases(chromosomal aneuploidy)identified as positive samples. The anomaly rate was calculated as 11.07%, indicating a significant class imbalance issue. The target variable is a binary anomaly label generated based on the AB column(chromosomal aneuploidy),as shown in Table 2.

Table 2. Partial Example of Data for Female Fetuses

Serial number	Maternal BMI.	Original read count	Alignment rate on the reference genome	Proportion of duplicate reads	Number of uniquely aligned reads	GC content	Z-score for chromosome 13
1	31.24523701	6531559	0.8070552	0.02861943	4993952	0.4025352	0.893295439
2	31.24523701	5222588	0.8069975	0.0266811	3997118	0.4001708	1.355154825
3	32.38835543	5637820	0.8015402	0.02967534	4277620	0.4073859	1.256408067

The data preprocessing pipeline in this study was rigorous and targeted, primarily consisting of the following steps. First, for format standardization, gestational age information was converted into numerical format(e.g.,"15w+3"transformed to 15.43 weeks)to facilitate subsequent calculations and statistical analyses. Second, missing values were imputed using median imputation for numerical features and mode imputation for categorical variables, thereby enhancing model robustness against outliers. Subsequently, numerical encoding was applied to certain categorical variables, such as IVF pregnancy types(spontaneous conception=0,artificial insemination=1,in vitro fertilization=2)and fetal health status(no=0,yes=1),to ensure feature consistency and enable proper recognition and utilization by the model. In the feature engineering phase, interaction terms(e.g., BMI×gestational age)and polynomial terms(e.g., gestational age^2)were constructed. Additionally, a comprehensive quality control score was generated based on multiple sequencing QC metrics, including alignment rate, duplication rate, filtering rate, and GC content deviation, to more thoroughly explore latent patterns within the data. Finally, during data partitioning and standardization, the dataset was divided into training and testing sets at a 7:3 ratio using stratified sampling to maintain distribution consistency across different categories. All continuous features underwent Z-score normalization prior to model training to eliminate scale differences and improve model stability.

2.1.2 Comprehensive feature engineering

To enhance the model's expressive capability, this study constructed three categories of derived variables based on original features. Interaction features were designed to capture potential synergistic effects between variables, such as BMI×gestational age and X chromosome concentration×gestational age; nonlinear features incorporated quadratic terms of variables including gestational age, BMI, and maternal age to capture nonlinear relationships; quality scoring features integrated multiple sequencing technical metrics(e.g., mapping rate, duplication rate, filtering rate, GC content deviation)through standardization and weighted averaging to generate a "quality control risk score" comprehensively reflecting sample sequencing quality.

To achieve systematic data understanding and ensure representative and interpretable model inputs, this study conducted rigorous feature analysis. For Y chromosome concentration in male fetus samples, Pearson and Spearman correlation coefficients were calculated between 25 initial features and Y chromosome concentration. Results demonstrated that X chromosome concentration exhibited the most significant positive correlation with Y chromosome concentration(Pearson $r=0.519$,$p<0.001$),showing substantially stronger influence than traditionally emphasized factors like gestational age($r=0.127$)and BMI($r=-0.151$).Additionally, blood draw frequency and

certain sequencing parameters(e.g., mapping rate, filtering rate)also displayed significant correlations, indicating combined effects of technical and biological factors on Y chromosome concentration. Through variance inflation factor(VIF)analysis, features with excessive multicollinearity were removed, ultimately retaining 15 significant variables for subsequent regression modeling.

Female fetus abnormality determination feature importance analysis: Using a random forest model, this study quantified each feature's contribution to female fetal abnormality detection. The top five features and their importance scores were: X chromosome concentration(0.126),chromosome 13 GC content(0.125),chromosome 18 GC content(0.104),maternal BMI(0.088),and chromosome 21 GC content(0.078).These results indicate that GC content-related features and X chromosome concentration constitute the most critical predictors for female fetal abnormality detection, collectively accounting for over 50%of total importance. This suggests potential impacts of GC bias on Z-score accuracy and highlights the fundamental role of fetal DNA fraction(indirectly reflected through X chromosome concentration).

2.2 Model Construction

2.2.1 Model for the association of male fetal Y chromosome concentration

For male fetal samples, this study employed multivariate linear regression to analyze factors influencing Y chromosome concentration. The model was constructed with Y chromosome concentration as the dependent variable, while independent variables included physiological characteristics, technical parameters, and chromosome-related indicators.

Feature selection was performed using a combination of "dual correlation testing" and recursive feature elimination(RFE),ultimately identifying 15 significant variables. The model was validated through normality tests, significance tests, and heteroscedasticity tests, demonstrating both statistical significance and robustness.

2.2.2 Probability model of fetal DNA Meeting target levels and chance-constrained optimization

(1)Probability modeling of reaching the standard

Define a "target achieved" event as Y chromosome concentration≥4%,and fit the probability function of target achievement using logistic regression and random forest algorithms:

$$\mathrm{p(t,x)} = \mathrm{P(Y} \geq 4\% \mathrm{|t,x)} \quad (1)$$

where represents the gestational age and denotes individual characteristics. Considering physiological principles, the model imposes a monotonically increasing constraint on gestational age to ensure that the probability of meeting the standard does not decrease as gestational age increases.

(2)Chance-constrained optimization decision

Based on this, chance-constrained programming is introduced to determine the optimal detection time point. A comprehensive cost function is defined as follows:

$$C(t) = w_1[1 - p(t,x)] + w_2 TimeRisk(t) + w_3 Cost(t) \quad (2)$$

Among them, the three terms represent the risks of not meeting standards, delayed testing, and economic costs, respectively, with corresponding weights. The optimization objective is, subject to constraints. Here, represents the confidence level (e.g.,0.90 for low-risk groups),indicating that the probability of meeting standards must exceed this threshold for the majority of pregnant women within each group.

Using numerical optimization methods, optimal testing time points are determined for different BMI groups within the gestational age range [10,25]weeks.

2.2.3 Two-stage mixed model for female fetus abnormality determination

To address the issue of lacking a Y chromosome reference for female fetuses, this study designed a two-stage determination framework combining "rule-based prioritization and model-assisted classification".

In Stage 1(rule-based screening),hard criteria were established based on clinical experience: samples were directly flagged as high-risk when the absolute Z-scores of chromosomes 13,18,21,or X exceeded 3;if sequencing quality metrics (e.g., alignment rate, read depth)fell below predefined thresholds, the result was classified as "quality control uncertain".

In Stage 2(model-assisted determination), machine learning models were applied to refine classification for "gray zone" samples that passed the initial rule-based screening. Among seven evaluated classifiers, logistic regression was identified as the optimal model based on its area under the curve (AUC)and recall

performance. This model incorporated over 20 features including Z-scores, GC content, and X chromosome concentration to output an abnormality probability.

Final determinations were made by applying a probability threshold, thereby maintaining high sensitivity for detecting true abnormalities while effectively controlling the false positive rate.

3. Analysis of Results

This study introduces the chance-constrained programming approach into decision-making for NIPT testing timing, establishing a decision framework that integrates probabilistic prediction with optimization. Under specified confidence levels, this method calculates the optimal testing time point that minimizes comprehensive risk.

Results demonstrate that the model effectively characterizes relationships between Y chromosome concentration, X chromosome concentration, and other indicators(R^2=0.471),thereby providing differentiated optimal testing times for pregnant women at different BMI levels(e.g., approximately 16 weeks for low-risk groups and 20 weeks for high-risk groups)along with corresponding confidence intervals.

Additionally, the female fetus abnormality detection model achieves a recall rate of approximately 70%in abnormal samples, significantly enhancing screening sensitivity.

3.1 Experimental Results and Analysis of Male Fetuses

The final constructed multiple linear regression model for Y chromosome concentration in male fetuses demonstrated an adjusted R-squared value of 0.471,with an F-statistic of 65.18($p<0.001$),indicating overall model significance at the extreme level. Among all predictors, X chromosome concentration exhibited the highest standardized regression coefficient(0.401),identifying it as the strongest influencing factor. This model effectively elucidated key driving factors for Y chromosome concentration, establishing a theoretical foundation for determining optimal timing of individualized testing[7][7].

To provide clinically actionable stratified strategies, this study developed a testing protocol based on BMI stratification and Chance-Constrained Programming, with specific details presented in Table 3.

Table 3. Opportunity-Constrained Programming-Based BMI Grouping and Optimal Detection Time Point Recommendation

BMI Interval (kg/m^2)	Risk Level	Optimal Time Point (weeks)	95% Confidence Interval for Time Point (weeks)	Stratified Confidence Level (τ)
18.5-24.9	Low risk	16.0	[15.6,16.4]	0.90
25.0-32.9	Medium risk	17.0	[16.2,17.8]	0.80
≥33.0	High risk	20.0	[18.6,21.4]	0.50

The optimization results indicated that the recommended timing for testing in low-risk pregnant women is at 16.0 weeks of gestation, while high-risk pregnant women are advised to delay testing until 20.0 weeks.

The uncertainty in these timing recommendations was further quantified through Monte Carlo simulations, which demonstrated smaller fluctuation ranges for the low-risk group(95%CI:±0.4 weeks)[8] [8]compared to relatively larger variation intervals observed in the high-risk cohort(95%CI:±1.4 weeks).

These findings provide an evidence base for establishing flexible clinical appointment windows. This approach offers clear clinical guidance and can effectively inform clinical practice.

3.2 Experimental Results and Analysis of Female Fetuses

In the classification of abnormal female fetuses, this study compared seven classifiers and ultimately selected the logistic regression model as the optimal approach, as shown in Table 4[9][10].[9,10]

Table 4. Key Assessment Results

Model	Accuracy	Test-AUC	CV-AUC	Exception class recall
Decision Tree	0.7747	0.5208	0.5864±0.0549	0.30
Random Forest	0.9011	0.7253	0.7370±0.0982	0.20
SVM	0.8516	0.8068	0.7112±0.0665	0.55
Logistic Regression	0.7967	0.8123	0.7498±0.0710	0.70

XGBoost	0.8846	0.6901	0.7047±0.0716	0.20
XGBoost+SMOTE	0.8956	0.7336	0.9868±0.0081	-
Random Forest+ADASYN	0.8297	0.7460	0.9596±0.0113	-

As shown in Table 4, the random forest model achieved the highest prediction accuracy, reaching 90.11%.The test set AUC of the logistic regression model reached 0.8123, with an abnormal recall rate as high as 70%.This result indicates that under conditions of extreme sample imbalance, the model can effectively identify true abnormal cases, which is particularly critical for clinical screening where "preventing missed diagnoses" is the primary objective.

The design of the two-stage framework not only ensures a high recall rate but also guarantees interpretability through rule-based screening in the first stage.

To address the issues of Y chromosome reference absence in female fetuses and sample imbalance, this study developed a two-stage approach combining rule-based determination with model prediction. The final logistic regression model achieved approximately 70%recall rate among abnormal samples, thereby enhancing the sensitivity of abnormality screening for female fetuses.

4. Conclusion

This study addresses the issue of low fetal cell-free DNA proportion in non-invasive prenatal testing(NIPT)among individuals with high body mass index(BMI).We developed and validated an individualized decision-making methodology integrating chance-constrained programming with multifactorial modeling, transforming clinical challenges into computable models. This approach achieved comprehensive optimization from data analysis to optimal timing recommendation for testing.

Experimental validation demonstrated that gestational age and BMI exert significant influences on Y chromosome concentration, with X chromosome concentration identified as the most critical predictive indicator. Additionally, technical factors including sequencing parameters showed substantial contributions to the outcomes, providing specific references for subsequent quality control protocols.

Overall, this research established an integrated analytical framework encompassing "probabilistic modeling-optimization decision-making-clinical determination", which not only enhanced the precision and reliability of NIPT but also demonstrated excellent versatility and scalability of the methodology. The framework can be extended to other low-concentration cell-free nucleic acid detection scenarios such as non-invasive tumor early screening and neonatal genetic disease screening, offering systematic modeling strategies and algorithmic pathways for addressing similar uncertainty decision-making problems.

To further improve model performance and applicability, future work will focus on: first, incorporating diverse clinical data types to continuously refine the feature system, thereby enhancing model robustness and generalization capability; second, advancing deep integration with hospital information systems to develop real-time clinical decision support tools, facilitating practical implementation of research findings. In the long term, this framework is expected to expand to broader screening applications, systematically validating its cross-domain applicability and effectiveness.

References

[1] Xue, Y., Zhao, G. D., Qiao, L. W., et al. Research progress on fetal cell-free DNA enrichment technology in maternal peripheral blood for NIPT. Chinese Journal of Birth Health&Heredity,2023,31(05):1087-1090.DOI:10.13404/j.cnki.cjbhh.2023.05.029.

[2] Ju,J., Li,J., Liu,S., et al. Estimation of cell-free fetal DNA fraction from maternal plasma based on linkage disequilibrium information. npj Genomic Medicine,2021,6:85.

[3] Guo,Z., Zhang,B., Yang,D., Wang,L., et al. Multidimensional roles of cfDNA fragmentomics in preeclampsia: from placental hypoxia and TLR9 inflammation to clinical risk stratification. Frontiers in Medicine,2025,12:1539651.

[4] Liscovitch-Brauer, N., Mesika, R., Rabinowitz, T., et al. Machine learning-enhanced noninvasive prenatal testing of monogenic disorders. Prenatal Diagnosis, 2024,44(9):1024-1032.DOI:10.1002/pd.6570.

[5] Zhang, J., Wu, Y., Chen, S., et al.

Prospective prenatal cell-free DNA screening for genetic conditions of heterogenous etiologies. Nature Medicine, 2024, 30(2): 470-479. DOI:10.1038/s41591-023-02774-x.

[6] Liu, W. Q., Yang, J. X., Zhang, J., et al. Technical standards consensus for high-throughput sequencing of fetal cell-free DNA in maternal peripheral blood plasma for pathogenic copy number variation screening. Chinese Journal of Medical Genetics, 2021, 38(7): 613-619.

[7] Xu, C., Li, J., Chen, S., et al. Genetic deconvolution of fetal and maternal cell-free DNA in maternal plasma enables next-generation non-invasive prenatal screening. Cell Discovery, 2022,8(1): 109. DOI:10.1038/s41421-022-00457-4.

[8] Pender, J., Ko, Y. M., & Xu, J. The number of overlapping customers in Erlang-A queues: an asymptotic approach. Probability in the Engineering and Informational Sciences,2025:1-26.

[9] Lu, Y., Chen, Y., Ding, S., et al. Performance analysis of non-invasive prenatal testing for trisomy 13,18,and 21:A large-scale retrospective study(2018–2021).Heliyon,2024,10(13).

[10] Chen, T., & Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 785-794.

