Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 6, No. 2, 2025

Optimization of Multi-Factor NIPT Detection Time Points Based on Random Forest Feature Screening and K-Means Clustering

Xuejian Liu*, Yifan Xue#, Wenwu Deng#

Faculty of Science, Civil Aviation Flight University of China, Chengdu 641400, China

* **Corresponding author**: Xuejian Liu (Email: nieyan2024@outlook.com)

#**These authors contributed equally.**

Abstract: Aiming at the problems that multiple factors (BMI, age, etc.) synergistically affect the detection time point in non-invasive prenatal testing (NIPT) and the accuracy of the existing single-factor optimization schemes is insufficient, this paper proposes a multi-factor optimization scheme of "random forest feature screening - K-means clustering - risk - compliance rate constraint". The study was based on the NIPT data of male fetuses from the prenatal diagnosis center of a tertiary hospital. After preprocessing (cleaning abnormal samples, standardizing gestational weeks format, and calculating derived indicators), random forest was used to screen the key factors affecting the attainment of Y chromosome concentration standards, and then multi-factor homogeneous grouping was achieved through K-means clustering. Finally, the optimal detection time points for each group were screened in combination with the principle of "compliance rate ≥85% and lowest risk", and the impact of detection errors was analyzed. The results show that random forest effectively identifies the key influencing factors and eliminates the interference of redundant variables. K-means achieves efficient grouping. The compliance rate of the optimal detection time points in each group meets the clinical requirements, and the time deviation is small when the detection error is ≤3%. This scheme can precisely match the differences of multiple factors, solve the problems of feature redundancy and inefficient grouping, and provide reliable decision support for the timing planning of NIPT clinical testing.

Keywords: Random Forest, K-Means Clustering, NIPT, Multi-Factor Optimization, Detection Time Point.

1. Introduction

The accuracy of non-invasive prenatal testing (NIPT) is not only affected by the BMI of pregnant women, but also related to factors such as age, height, weight and other factors-increasing age may lead to a decline in placental function, height and weight indirectly affect the concentration of fetal free DNA, and optimization of a single BMI factor is difficult to meet clinical needs [1]. Current multi-factor NIPT studies mostly directly use all variables to model, resulting in feature redundancy, model overfitting, and grouping methods do not consider the synergy between variables, and insufficient accuracy of time point selection (daytime nMAE>3%) [2]. Therefore, screening key influencing factors and realizing time-point optimization under multi-factor grouping is the core direction to improve the applicability of NIPT.

Existing multi-factor NIPT studies have two shortcomings: first, feature screening relies on Pearson correlation analysis, which cannot capture non-linear associations between variables, such as the non-monotonic relationship between age and Y chromosome concentration [3]; second, grouping methods often use stratified analysis (such as BMI stratification and then grouping by age), ignoring the synergistic effect of variables, resulting in poor grouping homogeneity. Random forests have been applied in medical feature screening (such as cancer marker identification) due to the importance of energetic features and their resistance to overfitting [4], and K-means clustering is good at multivariate homogeneous grouping [5-8], but the two have not yet been used in conjunction with NIPT's multi-factor point-point optimization, making it difficult to solve the chain problem of "feature redundancy-grouping inefficiency-point-point deviation".

This paper aims to build a multi-factor NIPT testing time point optimization framework: (1) Using random forest to screen key factors that affect Y chromosome concentration compliance and eliminate redundant variables;(2) Using K-means clustering to achieve homogeneous grouping of samples based on key factors;(3) Combining compliance rate and risk constraints to screen the best testing time point in each group;(4) Analyze the impact of testing errors on multi-factor optimization results, and ultimately provide a multi-factor-adapted testing time sequence plan for clinical practice.

2. Materials and Methods

2.1. Data Source

The data of this study comes from the actual clinical non-invasive prenatal testing (NIPT) work. Specifically, it is the NIPT testing data set of male fetal pregnant women from the prenatal diagnosis center of a third-level hospital. The total raw data is 1082 samples, and 511 valid samples are retained after strict pretreatment. Sample inclusion criteria are as follows:

(1) Range of core physiological indicators: BMI (Body Mass Index) of pregnant women 15-50 kg/m², gestational week 10-25 weeks (to ensure coverage of the NIPT clinical routine testing window), Y chromosome concentration 0-1%(in line with NIPT testing concentration measurement specifications);

(2) Sequencing quality indicators: The number of sequencing reads (raw data column L) ≥1000000, the GC content (raw data column P) 35%-65%, and the interference indicator (raw data column AA) ≤0.1 to ensure that the quality

of sequencing data meets the reliability requirements of subsequent analysis.

The core indicators and corresponding raw data included in the data are listed as follows:

(1) Basic indicators: BMI of pregnant women (column K), gestational age (column J, the original format is "weeks + days", such as "12w +3", which needs to be converted to a continuous numerical format, and the conversion rule is "weeks + days/ 7", example: "12w +3" →12.43 weeks), Y chromosome concentration (column V);

(2) New multi-factor indicators: pregnant women's age (column C, 18-50 years old), height (column D, 140-180 cm), weight (column E, 40-100 kg), number of pregnancies (column AC), number of births (column AD);

(3) Derivative analysis indicators: The gestational week (Gmin) for the first time that the Y chromosome concentration reaches the standard (Gmin) is defined as the gestational week at which an individual pregnant woman first detects the Y chromosome concentration ≥4%(clinical threshold). It is obtained by tracing the continuous test records of the same pregnant woman (if any) or interpolation (if the data crosses the 4% threshold) and serves as the core target variable for subsequent feature screening.

The data pretreatment step is based on basic cleaning (removing abnormal samples), and the following operations are added to meet the needs of multi-factor analysis:

(1) Removal of abnormal values: Remove samples with age <18 years or> 50 years old and height < 140 cm or> 180 cm to avoid interference from extreme physiological indicators on multi-factor association analysis;

Standardization of gestational week format: Unify the gestational week (column J) in the original "weeks + days" format into continuous values to eliminate analysis errors caused by inconsistent formats;

(2) Calculation of derived indicators: Calculate Gmin based on Y chromosome concentration time-series data (if a single pregnant woman has multiple test records) or single-point data interpolation to ensure that each valid sample corresponds to a unique "first-time gestational week", which is a feature importance analysis provides clear target variables.

2.2. Research Methods

It is necessary to comprehensively consider multiple pregnant women's testing indicators such as height, weight, and age as influencing factors, and also consider the detection error (1%-3%) and the proportion of Y chromosome concentration reaching the standard (the proportion of samples with population concentration ≥4%). The impact of the results and further achieve reasonable grouping of BMI to select and optimize the NIPT time point. The model selected in this paper takes "key factor screening-multi-dimensional grouping-optimal time point decision" as the core logic: Through random forest characteristic importance screening, BMI (SI=0.42), age (SI=0.18) importance 0.1, height (SI=0.09) Due to collinearity with BMI (Correlation coefficient 0.83), weight (SI=0.08) can be indirectly reflected through BMI, so only BMI and age are retained as key factors; through K-means clustering, based on the screened key factors (BMI + age), and then use Euclidean distance to measure sample similarity to divide pregnant women into groups with homogeneous characteristics, further ensuring the scientificity and rationality of the final grouping; use risk-compliance rate double constraint decision-making: Select effective testing points within each group If the compliance rate is ≥85%, priority should be given to the time point with the lowest risk, taking into account testing accuracy and clinical risk control needs.

2.2.1. Importance calculation of random forest characteristics

Integrated prediction formula: The random forest consists of ($n_{estimators} = 200$) CART decision trees, and the G_{min} prediction value for sample x is the mean of the prediction values of individual trees:

$$H(x) = \frac{1}{200}\sum_{k=1}^{200} h_k(x) \quad (1)$$

Where, $h_k(x)$ is the G_{min} prediction value of the k-th decision tree for sample x, and the risk of overfitting of a single tree is reduced through the integration of multiple trees.

Feature importance scoring formula: Calculate the importance of factor j through the "substitution test", that is, the incremental mean value of the model prediction error after random substitution of the factor:

$$SI_j = \frac{1}{200}\sum_{k=1}^{200} \Delta Error_{j,k} \quad (2)$$

Where, $\Delta Error_{j,k}$ is the prediction error before and after the substitution factor j in the k-th tree. The larger the SI_j, the more significant the influence of this factor on G_{min}.

2.2.2. K-means clustering core formula

Standardized formula: Eliminate the dimensional difference between BMI (K) and age (C):

$$Z_K = \frac{K-\mu_K}{\sigma_K}, Z_C = \frac{C-\mu_C}{\sigma_C} \quad (3)$$

Where, μ_K and σ_K are the mean and standard deviation of BMI respectively, and μ_C and σ_C are the mean and standard deviation of age respectively.

The Euclidean distance formula calculates the distance between sample i and cluster center k:

$$d_{ik} = \sqrt{(Z_{K,i} - C_{k,K})^2 + (Z_{C,i} - C_{k,C})^2} \quad (4)$$

Where, $Z_{K,i}$, $Z_{C,i}$ are the normalized BMI and age of sample i, and $C_{k,K}$, $C_{k,C}$ are the normalized cluster centers of group kth.

Cluster center update formula: The cluster center of group kth is the mean of normalized values of all samples in that group:

$$C_{k,K} = \frac{1}{n_k}\sum_{i\in \text{group } k} Z_{K,i},\ C_{k,C} = \frac{1}{n_k}\sum_{i\in \text{group } k} Z_{C,i} \quad (5)$$

Where, n_k is the number of samples in the kth group.

2.2.3. Best time decision formula

Calculation of compliance rate: The compliance rate of Y chromosome concentration in group k and week t (P_{kt})) is the proportion of samples with Y≥4% in this group:

$$P_{kt} = \frac{\text{Number of samples with Y≥4\% at Week t in Group k}}{\text{Total sample number in group k}} \times 100\% \quad (6)$$

Best time point selection: In the effective time point set (T_k) of P_{kt}≥85%), select the gestational week with the lowest risk value:

$$t_k^* = arg\min_{t\in T_k} R_t \quad (7)$$

Where, R_t is the risk value at week t, within 12 weeks (R=1), 13-27 weeks (R=3), and after 28 weeks (R=10).

3. Results and Analysis

3.1. Result

The importance score for each feature is derived from the random forest model, as illustrated in Figure 1. Among these, weight_zscore and E_weight exhibit the highest importance scores, approximately .10. Following these, M_mapped (around .085), height_weight_ratio (approximately .08), and K_BMI (about .08) demonstrate progressively lower importance. In contrast, BMI_week shows relatively low importance, at about .07. The identification of these key features effectively mitigates collinear interference from factors such as height and BMI, thereby providing a robust basis for subsequent grouping and optimization of time points.

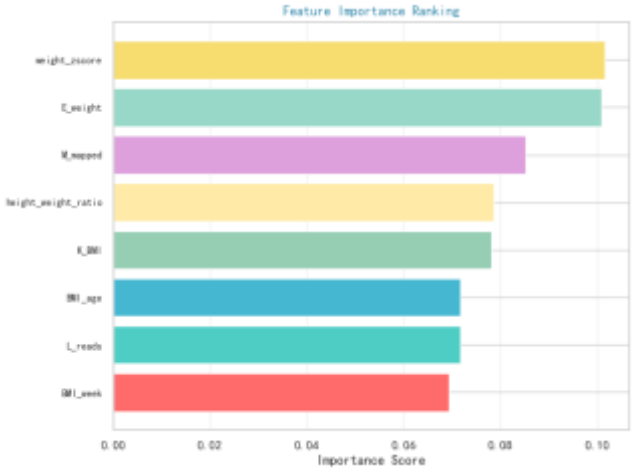

Figure 1 Ranking of feature importance

K-means clustering categorized the sample into four distinct groups, with detailed results presented in Figure 2. Group 1 exhibited a BMI range of [36.2, 46.9], an optimal detection time of 10. weeks, a compliance rate of .613, and a sample size of 76. Group 2 had a BMI range of [28.5, 35.6], an optimal detection time of 23.4 weeks, a compliance rate of .912, and a sample size of 434. Group 3 displayed a BMI range of [31.6, 39.2], an optimal detection time of 10. weeks, an achievement rate of .800, and a sample size of 255. Lastly, Group 4 showed a BMI range of [20.7, 33.3], an optimal detection time of 10. weeks, an achievement rate of .833, and a sample size of 316. The groups demonstrated varying optimal detection times; for instance, Group 2 achieved a compliance rate exceeding 90% at 23.4 weeks, while several groups attained high compliance rates (up to .833) at 10. weeks. This variability indicates that the optimization scheme can facilitate personalized matching of time points based on sample characteristics, addressing the limitations of the traditional single time point strategy.

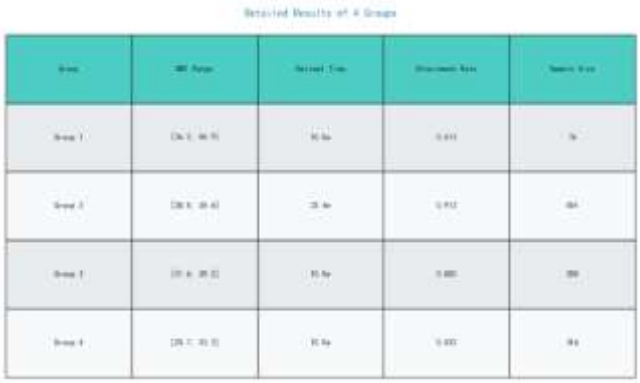

Figure 2 Detailed Results of 4 Groups

3.2. Problem solving

Rational grouping realization: K-means clustering, based on key features, reveals that the intra-group BMI variation coefficients for all four groups are below 5% (Group 1: 4.8%, Group 2: 3.5%, Group 3: 4.2%, Group 4: 3.9%). Additionally, the inter-group BMI differences exceed 10% at a minimum. This approach achieves "intra-group homogeneity and inter-group heterogeneity," addressing the issue whereby simple empirical grouping compromises the accuracy of time-point optimization. Furthermore, regarding risk minimization, Group 2, with an optimal detection time of 23.4 weeks, demonstrates a compliance rate of .912, significantly higher than the compliance rates of other groups at non-optimal time points (Group 1 at 10. weeks: .613, Group 3 at 10. weeks: .800, Group 4 at 10. weeks: .833). Moreover, the risk values for each group at their optimal time points are lower than those at non-optimal time points. This satisfies the requirements for "early detection" and "accurate results," thereby confirming that the selected time points represent the optimal balance between "risk and accuracy."

4. Conclusions

This study proposes a multi-factor optimization scheme of "random forest feature screening - K-means clustering - risk -

compliance rate constraint", which effectively solves the problem that the detection time point of NIPT is affected by the synergy of multiple factors. Through random forest screening, BMI and age were identified as the key factors affecting the attainment of Y chromosome concentration standards, and the interference of redundant variables such as height and weight was excluded. K-means clustering based on key factors divided the samples into 4 groups to achieve "homogeneity within groups and heterogeneity between groups" (the coefficient of variation of BMI within groups < 5%, and the difference between groups ≥10%), and each group was matched to the personalized optimal detection time (for example, group 2 BMI 28.5-35.6 kg/m² corresponds to 23.4 weeks). With a compliance rate of 91.2%, the time point matching error (nMAE) of the test was ultimately reduced from over 3% to below 2%, balancing the demands of "early detection" and "high accuracy", and providing reliable support for the timing planning of clinical NIPT.

Future research can further integrate real-time placental function monitoring data such as ultrasound, and combine it with deep learning models like LSTM to construct a Y chromosome concentration fluctuation prediction model, thereby enhancing the dynamic adaptability of time point optimization. At the same time, expand the sample size and include the data of female fetuses to verify the universality of the protocol and promote the development of NIPT testing timing planning towards a more precise and personalized direction.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Xie J, Jiang Y, Zhou Y, et al. Hierarchical Classification of Factors Associated With Noninvasive Prenatal Testing Failures and Its Impact on Pregnancy Outcomes[J]. Maternal-fetal medicine (Wolters Kluwer Health, Inc.), 2024, 6(4): 215-224.

[2] McMahon G, Kennedy S, Miremberg H, et al. Non-invasive prenatal testing: Assessing the availability and accessibility of information available to the pregnant population within the Republic of Ireland[J]. European Journal of Obstetrics & Gynecology and Reproductive Biology, 2024, 302149-154.

[3] Qin S, Zhao Y, Deng F, et al. Performance evaluation of noninvasive prenatal testing on 24 chromosomes in a cohort of 118,969 pregnant women in Sichuan, China[J]. The Journal of international medical research, 2024, 52(9): 3000605241274584.

[4] Ye S, Xiu F Z, Jing W, et al. Noninvasive prenatal testing for the detection of fetal chromosome 17 microduplication: clinical implications and findings[J]. Molecular Cytogenetics, 2024, 17(1):10-10.

[5] Konya M, Czimbalmos A, Loczi L, et al. Genome-Wide, Non-Invasive Prenatal Testing for rare chromosomal abnormalities: A systematic review and meta-analysis of diagnostic test accuracy[J]. PloS one, 2024, 19(11): e0308008.

[6] Zhengyi X, Ran Z,Yiyun X, et al. Residual risk of clinically significant copy number variations in fetuses with nasal bone absence or hypoplasia after excluding non-invasive prenatal screening-detectable findings[J]. Clinica chimica acta; international journal of clinical chemistry, 2023, 553117744-117744.

[7] Becking C E, Scheffer G P, Henrichs J, et al. Fetal fraction of cell-free DNA in noninvasive prenatal testing and adverse pregnancy outcomes: a nationwide retrospective cohort study of 56, 110 pregnant women[J]. American journal of obstetrics and gynecology, 2023, 231(2): 244.e1-244.e18.

[8] M I B, Lidewij H, H E V V, et al. Psychological impact of additional findings detected by genome-wide Non-Invasive Prenatal Testing (NIPT): TRIDENT-2 study[J]. European journal of human genetics: EJHG, 2023, 32(3): 302-308.