A multivariate modeling method for the prediction of low fetal fraction before noninvasive prenatal testing

Liang Hu1, Yuanyuan Pei1, Xiaojin Luo1, Lijuan Wen1, Hui Xiao2, Jinxing Liu1, Liping Wu3, Gaochi Li1 and Fengxiang Wei1,4

1Central Laboratory, Longgang District Maternity & Child Healthcare Hospital of Shenzhen City, Shenzhen, China
2Department of Pathology, Shenzhen Longgang People’s Hospital, Shenzhen, China
3Prenatal Diagnosis Center, Longgang District Maternity & Child Healthcare Hospital of Shenzhen City, Shenzhen, China
4Zunyi Medical University, Zunyi, Guizhou, China

Abstract
Objectives: To investigate factors associated with fetal fraction and to develop a new predictive method for low fetal fraction before noninvasive prenatal testing.

Methods: The study was a retrospective cohort analysis based on the results of noninvasive prenatal testing, complete blood count, thyroxin test, and Down’s syndrome screening during the first or second trimester in 14,043 pregnant women. Random forests algorithm was applied to predict the low fetal fraction status (fetal fraction < 4%) through individual information and laboratory records. The performance of the model was evaluated and compared to predictions using maternal weight.

Results: Of 14,043 cases, maternal weight, red blood cell, hemoglobin, and free T3 were significantly negatively correlated with fetal fraction while gestation age, free T4, pregnancy-associated plasma protein-A, alpha-fetoprotein, unconjugated estriol, and β-human chorionic gonadotropin were significantly positively correlated with fetal fraction. Compared to predictions using maternal weight as an isolated parameter, the model had a higher area under the curve of receiver operating characteristic and overall accuracy.

Corresponding author:
Fengxiang Wei, Central Laboratory, Longgang District Maternity & Child Healthcare Hospital of Shenzhen City, No.6 Ailong Road, Longgang district, Shenzhen 518172, China.
Email: haowei727499@126.com

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (https://us.sagepub.com/en-us/nam/open-access-at-sage).
Conclusions: The comprehensive predictive method based on combined multiple factors was more effective than a single-factor model in low fetal fraction status prediction. This method can provide more pretest quality control for noninvasive prenatal testing.

Keywords
Fetal fraction, noninvasive prenatal testing, prediction model, prenatal screening, random forests

Introduction
Noninvasive prenatal testing (NIPT) is a fetal chromosomal aneuploidy screening method mainly based on next-generation sequencing (NGS). Multiple clinical cohorts have validated NIPT as highly sensitive and specific for patients at increased risk of T13, T18, and T21 aneuploidies.1–3 It also has potential application value in the prenatal screening of copy number variables (CNV) and single-nucleotide variables (SNV) by target sequencing and extending the sequencing depth.4,5 Recently, NIPT is a widespread first-tier prenatal screening method used for high-risk pregnancies with aneuploidy.6

Cell-free DNA (cfDNA) in maternal plasma derived from both mother and fetus is the sequencing target of NIPT. Fetal fraction (FF) is the proportion of fetal origin in cell-free DNA and is considered critical to NIPT results.7,8 In the NIPT workflow, FF is based on the NGS results and calculated using a concentration of chromosome Y or algorithms of generalized linear regression, such as seqFF.9 It has been confirmed that FF <3.5%–4% often results in no-call or false results, and NIPT is not recommended for pregnant women with low FF.10,11 It is important to determine FF levels to ensure the quality of NIPT results.

FF comes out after the sequencing of NIPT, so it is necessary to perform a pretest evaluation to screen out the pregnant women who may have low FF to avoid invalid sampling or to adopt FF enrichment. Some factors, such as maternal weight,12,13 are associated with low FF, which can be used as a basis for prediction. FF is also speculated to be associated with chromosomal aneuploidy and the observation of Down’s syndrome screening which partially relates to placental development and volume.14

However, due to insufficient knowledge about the complex mechanism of generation and degradation of cell-free DNA, correlations of multiple known factors are not significant enough for FF prediction. By any isolated variable, it is difficult to accurately predict whether FF is too low for NIPT.15 But if these factors are integrated and analyzed comprehensively to predict low FF status, can the accuracy of the prediction model be improved?

The aim of this study is to use a multivariable and random forest algorithm16 to predict FF is too low before NIPT sampling. It can provide more accurate FF information for clinics without increasing healthcare costs.

Methods
Population selection
This study was a retrospective analysis involving a total of 14,043 pregnant women who received prenatal care in the Shenzhen Longgang Maternal and Child Healthcare
Hospital. Inclusion criteria consisted of women with singleton pregnancies who underwent NIPT between 12 and 24 weeks of gestation, first or second-trimester serum screening, and underwent at least one of the following clinical laboratory tests: complete blood count (CBC) test and thyroxin test. Each participant in our study was informed that there is no additional need to draw blood for this study and written consented to authorize the anonymous use of their laboratory and clinic information for nonprofit scientific studies. All data were stored in a local server without unauthorized access, and the sequencing data of NIPT was only used for FF estimation. The study was approved by the ethics committee of the Shenzhen Longgang Maternal and Child Healthcare Hospital before implementation.

NIPT and FF calculation

For each participant, a 5 ml maternal whole blood sample was collected using EDTA-K$_2$ tubes (BD, UK). Cell-free maternal plasma was separated and purified by centrifugation at 1600 $\times$ g (10 min, 4°C) for whole blood and 13,000 $\times$ g (10 min, 4°C) for plasma sequentially. The specimens for NIPT were stored at −80°C for no more than 3 days before NIPT. DNA extraction and library construction were performed for each sample following the manufacturer’s instructions of the NIFTY chromosomal abnormality test kit (BGI, Wuhan, China). Library qualification was determined using Qubit 3.0 (Thermo, USA). Then, a total of 48 libraries were pooled into one mixed library, which was single-end sequenced on the BGI SEQ500 platform (BGI, Wuhan, China) with 35-read length and 9.7 M average reads per sample. First and second-trimester serum screening, CBC, and thyroxin test were quantified within 1 month before FF, and stored at 4°C for no more than 24 h.

After reads alignment to the reference genome, hg19 (bwa-0.7.11), duplication removal of polymerase chain reaction (PCR) (SAMtools-1.9), read counting for 30 kb bins and GC bias correction based on locally weighted regression (LOWESS). A three-step FF measurement workflow was implemented as follows. First, FF of all male fetal samples was calculated using the Y-specific method as a training set. Second, a SeqFF model was trained and fitted on the training set. Finally, the SeqFF model was applied to all samples including male and female fetal samples as the final value of FF for the following analysis. All downstream analysis was based on FF measured with the SeqFF model. Standalone chromosome Y fraction reference and SeqFF model were established in our lab to fit the BGI sequencing platform. In this study, a cut-off of relatively low FF status was defined as FF lower than 4%. The percentage of low FF samples was recorded as FF$_{low}$$\%$. While relative low FF status is a logical variable, the existence of the status was defined as 1 and its nonexistence as 0.

All individuals with FF lower than 4% were informed and invited for resampling and reexamination of NIPT using the same workflow.

Association between FF and clinical variables

To construct a machine learning model predicting the FF, we collected clinical and individual information (Table 1) for each participant. The “Not Available” (NA) value was accepted to enhance the robustness of the model. Maternal weight during the first or
second trimester was included separately instead of the body mass index (BMI) because patient height data were not accessible.

For each variable, the Pearson correlation coefficient with FF and area under the curve (AUC) of the receiver operating characteristic (ROC) curve for predicting FF status was calculated to explore its impact on FF. To investigate the risk factors for FF status, the adjusted odds ratios (OR) and 95% confidence intervals (CIs) for each variable were calculated and adjusted by maternal age and gestation age. In addition, all specimens were divided into higher and lower groups for each variable; then, the Mann–Whitney U test was performed to assess the significance of FF in both groups.

### Random forest model and k-fold cross-validation

Random forest is an ensemble machine learning algorithm commonly used in various prediction scenarios and appropriate for data with high dimension and collinearity.\(^{16}\) Data pretreatment was performed before fitting the prediction model. First, clinical

| Characteristics | Overall | NFS | LFS | RLFS |
|-----------------|---------|-----|-----|------|
| Observations [N (%)] | 14,043 | 13,486 | 557 | 307 |
| Maternal age at delivery (years) | 29 (27–32) | 29 (27–32) | 30 (27–32) | 30 (28–32) |
| ≥35 [N (%)] | 1975 (14.06) | 1878 (13.93) | 97 (17.41) | 50 (16.29) |
| <35 [N (%)] | 12,068 (85.94) | 11,608 (86.07) | 460 (82.59) | 257 (83.71) |
| Gestation age at lab test (days) | | | | |
| First Down's syndrome screening | 88 (86–91) | 89 (86–91) | 88 (87–90) | 89 (88–90) |
| Second Down's syndrome screening | 118 (115–122) | 118 (115–122) | 118 (115–122) | 118 (115–122) |
| NIPT | 115 (98–123) | 115 (98–123) | 115 (101–119) | 136 (122–142) |
| Thyroxine test | 115 (75–124) | 115 (75–125) | 114 (75–121) | 115 (76–122) |
| Complete blood count | 113 (93–121) | 113 (93–121) | 112 (99–117) | 113 (101–117) |
| Maternal weight at Down's syndrome screening (kg) | | | | |
| First trimester | 53.0 (48.0–58.5) | 52.9 (48.0–58.0) | 57.8 (51.6–64.9)* | 59.0 (53.2–65.5)* |
| Second trimester | 54.0 (49.4–59.8) | 54.0 (49.1–59.2) | 59.6 (53.5–66.2)* | 61.0 (55.6–66.7)* |
| Gravidity [N (%)] | | | | |
| <3 | 8206 (58.43) | 7880 (58.43) | 326 (58.53) | 177 (57.65) |
| ≥3 | 5837 (41.57) | 5606 (41.57) | 231 (41.47) | 130 (42.35) |
| Parity [N (%)] | | | | |
| 0 | 6517 (46.41) | 6245 (46.31) | 272 (48.83) | 146 (47.56) |
| 1 | 6845 (48.74) | 6594 (48.90) | 251 (45.06) | 145 (47.23) |
| >1 | 681 (4.85) | 657 (4.87) | 24 (4.31) | 16 (5.21) |

*:\(^{p} < 0.05\), compared to normal FF status group.

FF: fetal fraction; LFS: low FF status group; NFS: normal FF status group; RLFS: low FF status in resampling of specimens in LFS group.
laboratory test results and information were normalized using the z-score method after replacing outliers with NA values. Second, NA values were imputed using the k-nearest neighbors (kNN) method. Finally, the dataset was split randomly and equally into 10 folds for 10 iterations. For each iteration, 7 folds (70% specimens) were selected as the training set and the remaining 3 folds (30% specimens) as the test set.\textsuperscript{21}

We established a supervised regression model using the R package randomForest 4.6-14, which implements the random forests algorithm introduced by Breiman et al.\textsuperscript{22} All clinical test results and training set information were input as independent variables and relative low FF status as the response variable. Then, the fitted model was validated using a test set to predict the relative low FF status with an output range of [0,1]. Samples with predicted values lower than 4% were marked as “predictive low FF status”.

ROC and precision-recall (PR) curves were selected for the performance evaluation. To compare this multivariate modeling predictive method with a single-variable predictive method, we also defined a prediction method based on maternal weight during the first and second trimesters: samples with a maternal weight higher than $1 - \text{FF}_{\text{low}}\%$ were marked as “predictive low FF status”. We compared the effectiveness of the predictive model in classifying low and normal FF status against maternal weight in the first and second trimesters. For specimens with no call results due to low FF, the model was also tested in the prediction of the repeated low FF status in resampling.

**Statistical analysis**

Statistical analysis was performed using R-4.1.0 (https://www.r-project.org/). Non-Gaussian distribution data are expressed by median and interquartile spacing. NA value imputation was performed using the R package DMwR (0.4.1).\textsuperscript{23} Kolmogorov–Smirnov test was applied for determining whether continuous variables could be fitted with a Gaussian distribution. Student’s $t$ and Mann–Whitney U tests were performed to compare differences between Gaussian and non-Gaussian distributed continuous variables, respectively. Chi-square and Fisher’s exact tests were used for categorical variables, and the DeLong test was used for comparing the performance of two ROC curves. A probability value ($p$-value) lower than 0.05 was considered statistically significant.

**Results**

**Study population profiles**

In this dataset, NIPT sequencing data and laboratory information from 14,043 women were collected. For each variable, a sample size of accessible data are shown in Table 1. The cut-off of low FF status was set to 4% and the percentage ($\text{FF}_{\text{low}}\%$) was 3.7%. Characteristics and profiles of NIPT participants with normal (FF > 4%) or low (FF ≤ 4%) FF statuses are described in Table 1. Among the 557 participants with low FF status, 542 of them (97.31%) had resampling, and the median of the interval between 2 blood sampling was 21 days, of which 235(44.88%) had normal FF and 307(55.12%) still had low FF. Compared to the normal FF status group (NFS), the maternal weight at first and second Down’s syndrome screening was higher in the low FF status
group (LFS) and resampling low FF status group (RLFS), while no significant difference was observed in maternal age, gestation age, gravidity, and parity between each group.

The percentages and distribution of each lab measurement and clinical information in this dataset are presented in Table 2. In this dataset, all distributions of each variant were not Gaussian distribution ($p < 0.05$).

**Associations between laboratory measurements and FF**

To reveal the relationships between laboratory measurements and FF, Pearson correlation coefficients were calculated for FF values. In addition, AUC of ROC and adjusted OR were obtained for low or normal FF status (Table 2). The difference of each measurement in the low and normal FF status groups is shown in Figure 1(a). There was no significant linear correlation between all the serum markers and FF ($r < 0.2$). However, some variables were predictive and suggestive for FF status. In the low FF status group, red blood cell (RBC), hemoglobin (HGB), and free T3 were significantly higher while free T4, pregnancy-associated plasma protein-A (PAPP-A), alpha-fetoprotein (AFP), conjugated estriol (uE3), and $\beta$-human chorionic gonadotropin ($\beta$-hCG) were significantly lower. AUCs and adjusted ORs showed that RBC, HGB, and free T3 were relative risk factors of low FF status, whereas PAPP-A, AFP, uE3, and $\beta$-hCG were protective factors (Figure 2(a)). Remarkably, higher maternal weight was significantly associated with low FF status. In contrast, a significant correlation between thyroid-stimulating hormone (TSH) with FF status was not observed. In this cohort, gestational age has little effect on FF between 11 and 24 weeks. Gestation age showed a weakly positive correlation with FF ($r = 0.2184$) but was not an obvious predictor of FF status (AUC = 0.5318, Table 2).

**NA value imputation**

The kNN algorithm was applied in NA value imputation to make the dataset complete. The median of each variant before and after NA value imputation is shown in Table 2. For each variable, no significant difference was observed before and after imputation ($p > 0.05$; Mann–Whitney U test). Density plots showed that the distribution density curves before and after imputation were similar, except for RBC and HGB, indicating that there was no significant impact on the data distribution characteristics induced by NA value imputation (Figure 1(b)). The distribution transformation in RBC and HGB might be caused by a high proportion of NA values.

**Performance evaluation of predictive model**

According to the correlation analysis before, TSH, maternal age, gestation age, gravidity, and parity were excluded for model training. Maternal weight, RBC, HGB, free T3, free T4, PAPP-A, AFP, uE3, and $\beta$-hCG in the first and second trimesters were input into the predictive model as dependent variables. After 10 iterations of model training and validation, an average of the ROC and PR curves for the predicted FF status was generated (Figure 2(b)). Compared to maternal weight, the AUC of the random forest model for predicting low FF the first time of sampling (AUC = 0.7146) and resampling (AUC =
Table 2. NA imputation and association with FF and clinical variables.

| Variables                              | Clinical data available [N (%)] | NA numbers [N (%)] | Before NA imputation | After NA imputation | r    | AUC               | OR               |
|----------------------------------------|---------------------------------|-------------------|----------------------|---------------------|------|-------------------|-------------------|
| **First Down’s syndrome screening**    |                                 |                   |                      |                     |      |                   |                  |
| β-HCG (ng/ml)                          | 11,381 (81.04)                  | 2662 (18.96)      | 57.12 (36.13–90.88)  | 57.65 (38.64–88.90) | 0.1372 | 0.5837 (0.5676–0.5998) | 0.7485 (0.693–0.8063) |
| PAPP-A (mIU/l)                         | 11,381 (81.04)                  | 2662 (18.96)      | 3834.83 (2472.23–5681.68) | 4036.46 (2738.46–5583.26) | 0.1259 | 0.6032 (0.5877–0.6186) | 0.6848 (0.6345–0.7375) |
| **Second Down’s syndrome screening**   |                                 |                   |                      |                     |      |                   |                  |
| β-HCG (ng/ml)                          | 10,734 (76.44)                  | 3309 (23.56)      | 15.70 (10.07–25.52)  | 16.00 (10.82–25.99) | 0.1234 | 0.5903 (0.5746–0.6061) | 0.7321 (0.6664–0.8002) |
| AFP (IU/ml)                            | 10,734 (76.44)                  | 3309 (23.56)      | 37.29 (29.51–47.02)  | 37.99 (31.10–45.73) | 0.1403 | 0.6052 (0.5897–0.6208) | 0.7136 (0.6549–0.7756) |
| uE3 (nmol/l)                           | 10,417 (74.18)                  | 3626 (25.82)      | 5.22 (4.23–6.49)     | 5.27 (4.50–6.21)     | 0.0678 | 0.5544 (0.5386–0.5702) | 0.8501 (0.7902–0.9128) |
| **Thyroxin test**                      |                                 |                   |                      |                     |      |                   |                  |
| TSH (mIU/l)                            | 12,256 (87.27)                  | 1787 (12.73)      | 1.27 (0.83–1.81)     | 1.31 (0.89–1.76)     | –     | 0.5067 (0.4909–0.5225) | 1.0091 (0.9493–1.0673) |
| Free T3 (pmol/l)                       | 12,240 (87.16)                  | 1803 (12.84)      | 4.29 (3.96–4.64)     | 4.33 (4.01–4.62)     | 0.0259 | 0.5923 (0.5768–0.6078) | 1.2076 (1.146–1.277) |
| Free T4 (pmol/l)                       | 12,238 (87.15)                  | 1805 (12.85)      | 11.56 (10.73–12.51)  | 11.51 (10.79–12.38)  | 0.1101 | 0.5338 (0.5178–0.5498) | 0.9106 (0.8514–0.972) |
| **Complete blood count**               |                                 |                   |                      |                     |      |                   |                  |
| RBC (10⁹)                              | 5011 (35.68)                    | 9032 (64.32)      | 3.79 (3.55–4.07)     | 3.79 (3.67–3.94)     | –     | 0.5991 (0.5833–0.6148) | 1.3433 (1.2338–1.4613) |
| HGB (g/l)                              | 5011 (35.68)                    | 9032 (64.32)      | 114 (108–121)        | 113.52 (110.03–117.00) | 0.1023 | 0.5911 (0.5752–0.6071) | 1.3612 (1.2379–1.4982) |
| Maternal age (years)                   | 14,043 (100)                    | 0 (0.00)          | 29 (27–32)           | 29 (27–32)           | –     | 0.4822 (0.4662–0.4981) | 1.0342 (0.9777–1.0939) |
| Gestation age (days)                   | 14,043 (100)                    | 0 (0.00)          | 115 (98–123)         | 115 (98–123)         | 0.0561 | 0.5204 (0.5056–0.5353) | 0.8682 (0.8154–0.923) |

(Continued)
Table 2. (continued)

| Variables                                                                 | Clinical data available [N (%)] | NA numbers [N (%)] | Before NA imputation | After NA imputation | r     | AUC                  | OR                  |
|--------------------------------------------------------------------------|---------------------------------|-------------------|----------------------|---------------------|-------|----------------------|---------------------|
| Maternal weight at Down’s syndrome screening (kg)                         |                                 |                   |                      |                     |       |                      |                     |
| First trimester                                                          | 11,831 (81.04)                  | 2662 (18.96)      | 53.0 (48.0–58.5)     | 53.0 (48.5–58.3)    | –     | 0.6629 (0.6475–0.6783)| 1.7177 (1.6241–1.8169)|
| Second trimester                                                         | 10,734 (76.44)                  | 3309 (23.56)      | 54.0 (49.4–59.8)     | 54.0 (49.7–59.6)    | 0.2451| 0.6676 (0.6523–0.6829)| 1.7565 (1.6569–1.8626)|

*p < 0.05, compared to normal FF status group.

AFP: alpha-fetoprotein; AUC: area under the curve; FF: fetal fraction; HGB: hemoglobin; NA: not available; OR: odds ratio; PAPP-A: pregnancy-associated plasma protein-A; RBC: red blood cell; TSH: thyroid-stimulating hormone; T3: triiodothyronine; T4: thyroxine; uE3: unconjugated estriol; β-hCG: β-human chorionic gonadotropin.
0.7534) was significantly higher ($p < 0.001$), indicating that the model was more effective in predicting low FF status than maternal weight in the first (AUC = 0.6222) and second (AUC = 0.6307) trimesters. The AUC was higher in the prediction of low FF in the RLFS group than in the prediction of low FF for the first time of sampling ($p < 0.001$). The PR curve also showed that the random forest model performed better than maternal weight. FF was significantly lower in the predicted low FF status group than in the predicted normal FF status group (Figure 2(c)). The overall accuracy of the predictive model was 0.8678 in the LFS group and 0.8872 in the RLFS group, higher than the overall accuracy of the prediction using maternal weight in the first (0.8104) and second (0.8167) trimesters.
Discussion

The amount of fetal DNA in the maternal plasma sample is affected by multiple factors which pose a challenge in the prediction of fetal DNA fraction before NIPT. To predict the existence of FF status, we collected 14 laboratory test results and the individual’s information and fitted a regression model based on the random forests algorithm. In
our cohort, we confirmed that the model was more effective in predicting FF status than using isolated maternal weight.

FF is a varied and complex biological indicator, which is often influenced by individual differences and laboratory factors. It has been reported that overweight pregnant women are at higher risk of test failure due to low FF. This might be because of a dilutional effect and higher release of maternal cfDNA from adipose cells into the systemic circulation.24 Gestation age is a relevant factor of FF after 20 weeks of gestation, but not in the early second trimester.13 In our study, we found the predictive model to not be significant. Some studies reported that FF increases with maternal serum levels of free β-hCG and PAPP-A, which reflect placental volume and development.25–27 Besides, fetal numbers, fetal aneuploidies, physical activity, maternal smoking, and sample transport or laboratory workflow have been associated with lower FF.28–31 In our cohort, free T4, and serum markers of the Down’s syndrome screen, such as PAPP-A, AFP, uE3, and β-hCG, in the first and second trimesters were all positively correlated with FF, in agreement with previous studies.25 Notably, this is a novel report of the negative correlation between RBC, HGB, and free T3 with FF.

Although it is difficult to reliably predict the value of FF without quantitative experiments, we hypothesized that qualitative prediction of FF status is practicable. Independent variables with constant cut-off values, such as maternal weight or gestation age, have been used to evaluate low FF status and minimize unreliable reports. In this study, we confirmed that prediction based on combined clinical and laboratory variables using machine learning algorithms is more effective. Theoretically, these variables are affected by some conjoint factors such as maternal weight and placental status, so collinearity could have existed in the data. However, in our cohort, collinearity was not significant by testing correlations of all variables in pairwise comparison. For positive predictions, sequencing shorter cfDNA fragments or cfDNA enrichment can be used to improve the FF of NIPT.32 Furthermore, there were many missing values in the raw clinical dataset for which direct filtering of missing values would have resulted in insufficient specimens for subsequent analyses. Incomplete data are common in clinical and cannot be avoided. Therefore, we adopted the imputation of missing values to reduce their impact and increase the robustness of the model.

There are some limitations to our study. First, it was a single-center study and the results should be validated in a multicenter cohort study because of diversities in sample size or laboratory methodology. NIPT is usually applied in the first trimester. But in our cohort, the sample size of the first trimester was insufficient for fitting the random forests model. Enlargement of the sample size is necessary for any following studies. Second, although the performance of the predictive model was promoted compared to maternal weight-based prediction, the model performance can still be improved to make it more clinically applicable. Furthermore, thyroid hormones tests are not common and regular for every individual on a worldwide scale. It is standard obstetrical care in local clinical service and is covered by local medical insurance. In this study, each variable had little impact on the prediction effect, and the prediction effect is generated by superposition and accumulation of multivariables. Therefore, when applied in other healthcare systems, the accessible variables should be used to replace thyroid hormones.
In this study, we report a new method, based on multivariate clinical and laboratory data, for predicting FF status before NIPT sampling. We confirmed that the prediction model was more effective than prediction using maternal weight as an independent variable. This study displayed the application of machine learning in prenatal screening and would provide more reference for making clinical choices for NIPT.

Acknowledgments
The authors would like to thank Song Xu and Zhuyan Huang from Shenzhen Longgang Maternal and Child Health Hospital for data collection; Hsi-Yuan Huang from Chinese University of Hong Kong (Shenzhen) for statistical guidance; and all the women who consented to participate in this study. This work was supported by grants from the Shenzhen Science and Technology Innovation Commission (grant number: JCYJ20180305125647151), the Shenzhen Longgang Science and Technology Innovation Commission (grant number: LGKCZSYS2018000010), and Shenzhen Longgang District Science and Technology Innovation Commission (grant numbers: LGKCYLWS2020111 and LGKCYLWS2020149).

Declaration of conflicting interests
The authors declare that there are no conflicts of interest that could be perceived as prejudicial to the impartiality of the reported research.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Shenzhen Longgang District Science and Technology Innovation Commission (grant numbers: LGKCYLWS2020111, LGKCYLWS2020149, LGKCZSYS2018000010, JCYJ20180305125647151).

ORCID iD
Fengxiang Wei https://orcid.org/0000-0001-7559-2504

References
1. Cheung SW, Patel A and Leung TY. Accurate description of DNA-based noninvasive prenatal screening. N Engl J Med 2015; 372: 1675–1677.
2. Noh JJ, Ryu HM, Oh S-Y, et al. A two-year experience of non-invasive prenatal testing (NIPT) at an urban tertiary medical center in South Korea. Taiwan J Obstet Gynecol 2019; 58: 545–551.
3. Sago H, Sekizawa A and Japan NIPT consortium. Nationwide demonstration project of next-generation sequencing of cell-free DNA in maternal plasma in Japan: 1-year experience. Prenat Diagn 2015; 35: 331–336.
4. Liang D, Cram DS, Tan H, et al. Clinical utility of noninvasive prenatal screening for expanded chromosome disease syndromes. Genet Med 2019; 21: 1998–2006.
5. Lv W, Li Z, Wei X, et al. Noninvasive fetal genotyping in pregnancies at risk for PKU using a comprehensive quantitative cSMART assay for PAH gene mutations: a clinical feasibility study. BJOG 2019; 126: 1466–1474.
6. Palomaki GE, Kloza EM, O’Brien BM, et al. The clinical utility of DNA-based screening for fetal aneuploidy by primary obstetrical care providers in the general pregnancy population. *Genet Med* 2017; 19: 778–786.

7. Balslev-Harder M, Richter SR, Kjaergaard S, et al. Correlation between Z score, fetal fraction, and sequencing reads in non-invasive prenatal testing. *Prenat Diagn* 2017; 37: 943–945.

8. Hui I and Bianchi DW. Fetal fraction and noninvasive prenatal testing: what clinicians need to know. *Prenat Diagn* 2020 Jan; 40: 155–163.

9. Kim SK, Hannum G, Geis J, et al. Determination of fetal DNA fraction from the plasma of pregnant women using sequence read counts. *Prenat Diagn* 2015; 35: 810–815.

10. Hestand MS, Bessem M, van Rijn P, et al. Fetal fraction evaluation in non-invasive prenatal screening (NIPS). *Eur J Hum Genet* 2019; 27: 198–202.

11. Gregg AR, Skotko BG, Benkendorf JL, et al. Noninvasive prenatal screening for fetal aneuploidy, 2016 update: a position statement of the American college of medical genetics and genomics. *Genet Med* 2016; 18: 1056–1065.

12. Kinnings SL, Geis JA, Almasri E, et al. Factors affecting levels of circulating cell-free fetal DNA in maternal plasma and their implications for noninvasive prenatal testing. *Prenat Diagn* 2015; 35: 816–822.

13. Wang E, Batey A, Struble C, et al. Gestational age and maternal weight effects on fetal cell-free DNA in maternal plasma. *Prenat Diagn* 2013; 33: 662–666.

14. Wataganara T, Metzenbauer M, Peter I, et al. Placental volume, as measured by 3-dimensional sonography and levels of maternal plasma cell-free fetal DNA. *Am J Obstet Gynecol* 2005; 193: 496–500.

15. Tamminga S, van Maarle M, Henneman L, et al. Maternal plasma DNA and RNA sequencing for prenatal testing. *Adv Clin Chem* 2016; 74: 63–102.

16. Liaw A and Wiener M. Classification and regression by randomForest. *R News* 2002; 2: 18–22.

17. Jiang F, Ren J, Chen F, et al. Noninvasive fetal trisomy (NIFTY) test: an advanced noninvasive prenatal diagnosis methodology for fetal autosomal and sex chromosomal aneuploidies. *BMC Med Genomics* 2012; 5: 57.

18. Li H and Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 2009; 25: 1754–1760.

19. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/Map format and SAMtools. *Bioinformatics* 2009; 25: 2078–2079.

20. Hudecova I, Sahota D, Heung MMS, et al. Maternal plasma fetal DNA fractions in pregnancies with low and high risks for fetal chromosomal aneuploidies. *PLoS ONE* 2014; 9: e88484.

21. Jung Y and Hu J. A K-fold averaging cross-validation procedure. *J Nonparametr Stat* 2015; 27: 167–179.

22. Breiman L. Random forests. *Machine Learning* 2001; 45(1): 5–32.

23. Torgo L. *Data mining with R, learning with case studies*. Chapman and Hall/CRC, 2010.

24. Haghiac M, Vora NL, Basu S, et al. Increased death of adipose cells, a path to release cell-free DNA into systemic circulation of obese women. *Obesity (Silver Spring)* 2012; 20: 2213–2219.

25. Ashoor G, Syngelaki A, Poon LCY, et al. Fetal fraction in maternal plasma cell-free DNA at 11–13 weeks’ gestation: relation to maternal and fetal characteristics. *Ultrasound Obstet Gynecol* 2013; 41: 26–32.

26. Hou Y, Yang J, et al. Factors affecting cell-free DNA fetal fraction: statistical analysis of 13,661 maternal plasmas for non-invasive prenatal screening. *Hum Genomics* 2019; 13: 62.

27. Scott FP, Menezes M, Palma-Dias R, et al. Factors affecting cell-free DNA fetal fraction and the consequences for test accuracy. *J Matern Fetal Neonatal Med* 2018; 31: 1865–1872.

28. Rava RP, Srinivasan A, Sehnert AJ, et al. Circulating fetal cell-free DNA fractions differ in autosomal aneuploidies and monosomy X. *Clin Chem* 2014; 60: 243–250.
29. Suzumori N, Ebara T, Yamada T, et al. Fetal cell-free DNA fraction in maternal plasma is affected by fetal trisomy. *J Hum Genet* 2016; 61: 647–652.

30. Schlütter JM, Hatt L, Bach C, et al. The cell-free fetal DNA fraction in maternal blood decreases after physical activity. *Prenat Diagn* 2014; 34: 341–344.

31. Qiao L, Mao J, Liu M, et al. Experimental factors are associated with fetal fraction in size selection noninvasive prenatal testing. *Am J Transl Res* 2019; 11: 6370–6381.

32. Qiao L, Yu B, Liang Y, et al. Sequencing shorter cfDNA fragments improves the fetal DNA fraction in noninvasive prenatal testing. *Am J Obstet Gynecol* 2019; 221: e345–3345.e11.