Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Participant-Level Validation Reveals Optimistic Performance in Repeated-Measure NIPT Modeling: A Reproducibility Study

Renjie Jin

ffgxfvy5562@outlook.com

China University of Petroleum-Beijing at Karamay https://orcid.org/0009-0004-3436-2453

Research Article

Keywords: non-invasive prenatal testing, repeated measures, data leakage, grouped cross-validation, fetal fraction, reproducibility, aneuploidy

Posted Date: July 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10475847/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Repeated non-invasive prenatal testing (NIPT) measurements from the same participant can be distributed across training and test sets when records are split at random, potentially inflating apparent model performance. We evaluated this risk in a de-identified repeated-measure benchmark dataset comprising 1082 male-fetus records from 267 participant codes and 605 female-fetus records from 147 codes. The development-stage Bayesian-optimized XGBoost analysis produced a record-level R^2 of 0.794 for Y-chromosome fraction prediction and a record-level ROC-AUC of 0.952 with an F1 score of 0.931 for female-fetus aneuploidy-label classification; BMI-stratified probability curves also indicated later candidate testing windows at higher BMI. To assess robustness to repeated measurements, we conducted 30 repeated five-fold experiments comparing record-level cross-validation with participant-grouped cross-validation for nonlinear regularized regression, a 4% fetal-fraction threshold model, and aneuploidy-label classification, together with a participant-mean identity-leakage stress test. Median Y-fraction regression R^2 decreased from 0.068 under record-level validation to 0.046 under participant-grouped validation, while the participant-mean predictor declined from R^2=0.432 to -0.006. Threshold-model ROC-AUC decreased from 0.588 to 0.563, and aneuploidy-label ROC-AUC decreased from 0.677 to 0.666. These findings show that strong record-level performance may partly reflect participant-specific information and should not be interpreted as evidence of generalization to unseen participants. Participant-grouped validation should be considered a minimum requirement for repeated-measure NIPT modeling, and prospective external validation remains necessary before clinical use.

1. INTRODUCTION

Non-invasive prenatal testing (NIPT) uses circulating cell-free DNA to screen for common fetal chromosomal aneuploidies [1]. Clinical guidance has established NIPT as a screening approach rather than a diagnostic test [2]. Its performance depends partly on fetal fraction, which varies with gestational age, maternal body mass index (BMI), assay characteristics, and biological factors [3]. Analyses of non-informative screening results further demonstrate the effects of gestational age, BMI, and maternal age [4]. These relationships have motivated predictive models of fetal fraction, test timing, and abnormality labels.

A less visible design issue is the statistical dependence created by repeated measurements. When several records from one participant are randomly distributed across folds, a model may recognize stable participant-level characteristics rather than learn patterns that transfer to a new participant. Standard record-level validation then answers the wrong question: performance on another record from a partly known participant, not performance on a previously unseen participant. Structured-data validation literature recommends grouping correlated observations during resampling [5]. Validation studies with limited sample sizes similarly emphasize subject-level partitioning [6]. Nested resampling can reduce bias introduced during model selection [7]. More generally, separating model selection from performance estimation is necessary to limit optimistic bias [8].

The dataset contains repeated participant codes, creating a natural setting in which to isolate the effect of the validation unit. We designed a reproducibility study with one primary hypothesis: record-level cross-validation yields more optimistic estimates than participant-grouped cross-validation. A participant-mean predictor was included as a stress test to quantify the extent to which participant identity alone can appear predictive.

Our contribution is not a new clinical NIPT decision rule. It is an empirical evaluation of validation leakage across three tasks using identical features and models under two resampling protocols. This design provides a controlled methodological benchmark and produces findings relevant to predictive studies with repeated biospecimens.

2. MATERIALS AND METHODS

2.1 Data source and analytical status

We performed a secondary analysis of a de-identified repeated-measure NIPT benchmark dataset. The male-fetus subset contained 1,082 records from 267 participant codes, and the female-fetus subset contained 605 records from 147 codes. Participant codes occurred repeatedly because of multiple blood draws and/or repeat sequencing measurements.

The dataset was not accompanied by the clinical provenance, ethics documentation, recruitment protocol, assay standard operating procedure, or prospective outcome adjudication required for a clinical cohort. It was therefore treated as a methodological benchmark rather than a clinically representative cohort. No participant contact, intervention, or re-identification was performed.

2.2 Outcomes and predictors

The first task predicted continuous Y-chromosome fraction in male-fetus records. The second task classified whether Y-chromosome fraction reached 4%. The third task classified whether the supplied female-fetus aneuploidy field was non-empty. This final outcome is assay-derived and should not be interpreted as an independent cytogenetic diagnosis.

Regression predictors included gestational week, BMI, age, log read depth, unique-read proportion, alignment proportion, duplicate-read proportion, overall GC content, filtered-read proportion, squared terms, and prespecified two-way interactions. The threshold model used gestational week, BMI, age, nonlinear terms, and a gestational-week-by-BMI interaction. The female-fetus model used Z scores for chromosomes 13, 18, 21, and X; BMI; read metrics; and GC metrics.

2.3 Validation protocols

We compared two five-fold protocols. In record-level cross-validation, rows were randomly assigned to folds without regard to participant code. In participant-grouped cross-validation, all rows belonging to

one code were assigned to exactly one fold. Consequently, no participant code observed in grouped test data was present in grouped training data.

The comparison was repeated for 30 random seeds. Every transformation was estimated from the training fold only. The same feature set, regularization, and model specification were used under both protocols, making the validation unit the experimental factor. Because the anomaly label was imbalanced, precision–recall AUC was prespecified as a complementary metric [9]. The reporting framework was aligned with TRIPOD principles [10]. Risk-of-bias considerations followed PROBAST recommendations [11].

2.4 Models and metrics

The development-stage ensemble was implemented using XGBoost [12]. Continuous fetal fraction was modeled with regularized nonlinear least squares, and binary outcomes were modeled with regularized logistic regression using standard statistical-learning methods [13]. Transparent regularized models were chosen to isolate validation effects rather than maximize benchmark scores. Regression was evaluated with R^2; binary tasks were evaluated with ROC-AUC and precision–recall AUC (PR-AUC), the latter being informative under class imbalance.

As an identity-leakage stress test, a participant-mean predictor estimated the mean Y-chromosome fraction for each participant code from training records. Test records from a previously observed code received that participant mean; unseen codes received the global training mean. Under properly grouped validation, every test participant is unseen, so the predictor should have no advantage over the marginal mean.

For each metric, we report the median and interquartile range (IQR) across 30 repetitions. The analysis was implemented in Python using deterministic scripts, and all reported values were read from persisted result files.

2.5 Reproducibility and sensitivity-analysis framework

We treated the validation unit as the primary experimental factor. All predictor definitions and model specifications were held constant while only the fold-construction rule changed. In record-level cross-validation, individual observations were assigned to folds without reference to participant code. In participant-grouped cross-validation, all observations belonging to one code remained together. This controlled comparison ensured that differences in reported performance could be attributed to the availability of participant-specific information rather than to changes in preprocessing, predictors, or model complexity.

The complete five-fold procedure was repeated with 30 random seeds to assess sensitivity to a particular partition. Within every repetition, missing-value handling, centering, scaling, and parameter estimation were performed using training data only. Test-fold information was not used to determine transformations or model coefficients. We summarized each metric with the median and interquartile

range because performance estimates from repeated resampling are correlated and need not follow a normal distribution. The objective was to characterize the stability and practical magnitude of the validation gap rather than to treat repeated folds as independent experimental replicates.

The participant-mean predictor served as a deliberately leakage-prone stress test. It estimated an individual-specific mean from training records whenever the same participant code was present and otherwise returned the global training mean. A large contrast between record-level and grouped evaluation would therefore indicate that within-participant information was being rewarded by the validation design. This stress test complements the regularized prediction models by demonstrating the mechanism of optimism directly, without requiring a highly flexible black-box algorithm.

Our analysis distinguished development-stage discrimination from clinical validity. Record-level R^2, ROC-AUC, and F1 values describe performance within the original modeling workflow, whereas the repeated grouped experiments estimate transfer to previously unseen participant codes. Neither protocol establishes transportability across hospitals, sequencing platforms, ancestry groups, or clinical referral populations. Calibration, net benefit, and patient-level decision consequences were therefore not interpreted as established outcomes.

3. RESULTS

3.1 Repeated-measure structure

The male-fetus worksheet contained 1082 records for 267 participant codes, and the female-fetus worksheet contained 605 records for 147 codes. The median number of records per code was 4 in the male worksheet and 4 in the female worksheet. Thus, random row splitting had multiple opportunities to place records from the same participant in both development and evaluation folds.

Worksheet	Records	Codes
Male fetus	1,082	267
Female fetus	605	147

Table 1. Structure of the repeated-measure benchmark dataset.

3.2 Validation-protocol comparison

For continuous Y-fraction prediction, record-level validation produced median R^2=0.068 (IQR 0.065–0.071), compared with 0.046 (IQR 0.039–0.052) under participant-grouped validation. Although both values were low, record-level splitting increased apparent explained variance by approximately 0.022 absolute R^2 units.

For the 4% threshold task, ROC-AUC declined from 0.588 (IQR 0.584–0.594) to 0.563 (IQR 0.548–0.569). For the female-fetus aneuploidy-label task, ROC-AUC declined from 0.677 (IQR 0.669–0.685) to 0.666 (IQR 0.649–0.681). PR-AUC declined from 0.365 to 0.348.

Task	Record-level	Grouped
Y fraction R^2	0.068	0.046
Threshold AUC	0.588	0.563
Aneuploidy AUC	0.677	0.666

Table 2. Median performance across 30 repeated five-fold experiments.

3.3 Development-stage model performance

The development-stage modeling pipeline used Bayesian-optimized XGBoost to characterize nonlinear associations in the repeated-measure data [12]. Figure 1 shows the fitted-versus-observed relationship for Y-chromosome fraction, with a record-level benchmark R^2 of 0.794. The classification comparison in Fig. 3 and the receiver operating characteristic curve in Fig. 4 show a record-level ROC-AUC of 0.952 and an F1 score of 0.931 for the proposed classifier. These development-stage estimates describe discrimination within the record-level evaluation framework; their generalizability is examined separately through the participant-grouped experiments reported in Table 2.

Figure 2 presents the BMI-stratified modeled probability of Y-chromosome fraction remaining below 4%, with candidate testing weeks shifting later as BMI increases. Because the grouped threshold model achieved only modest discrimination, the displayed timing curves are considered unvalidated model outputs rather than clinical recommendations. Their smooth form also differs from a conventional empirical Kaplan–Meier step function.

3.4 Identity-leakage stress test

The participant-mean predictor produced the largest validation gap. Under record-level splitting, it achieved median R^2=0.432 (IQR 0.408–0.437). Under participant-grouped validation, performance fell to R^2=-0.006 (IQR − 0.009–-0.004). This collapse is expected because a held-out participant has no participant-specific training mean. It demonstrates that identity-derived information can create substantial apparent predictability without any capacity to generalize to a new participant.

3.5 Interpretation of predictive utility

Participant grouping reduced all three primary metrics, but the grouped models also remained weak in absolute terms. The threshold model was only slightly above chance, and the aneuploidy-label model showed modest discrimination. Accordingly, the data do not support an individualized clinical timing rule or a diagnostic classifier. The principal empirical result is the validation gap, not the absolute model score.

3.6 Stability across repeated partitions

The direction of the validation gap was stable across the 30 repetitions. For Y-chromosome fraction regression, record-level R^2 had an interquartile range of 0.065–0.071, whereas participant-grouped R^2 ranged from 0.039 to 0.052. The limited overlap and consistently lower grouped estimates indicate that the result was not driven by a single unfavorable split. At the same time, both ranges remained close to

zero, showing that the transparent nonlinear regression captured only a small proportion of between-participant variation.

The 4% threshold model showed a similar but smaller pattern. Record-level ROC-AUC varied from 0.584 to 0.594, compared with 0.548–0.569 under grouped validation. For the female-fetus aneuploidy-label task, record-level ROC-AUC ranged from 0.669 to 0.685 and grouped ROC-AUC from 0.649 to 0.681. These distributions support the interpretation that row-wise splitting introduced optimism, although the magnitude depended on the prediction target and on how strongly repeated observations preserved participant-specific characteristics.

The participant-mean stress test provided the clearest contrast. Its record-level R^2 ranged from 0.408 to 0.437, but grouped estimates ranged from − 0.009 to − 0.004. Negative grouped R^2 indicates performance no better than, and slightly worse than, predicting the marginal training mean for unseen participants. Because the same algorithm and fallback rule were applied in both protocols, the collapse cannot be attributed to hyperparameter choice; it follows directly from removing access to repeated records from the test participant.

We did not compute conventional significance tests across folds because fold-level estimates share observations and are not statistically independent. Instead, the repeated-partition distributions were used as a robustness description. This approach avoids overstating precision while preserving the central empirical result: validation design changes the apparent performance of models trained on repeated-measure NIPT data.

4. DISCUSSION

4.1 Principal findings

This study tested validation-unit choice as an experimental factor in a repeated-measure NIPT dataset. Record-level splitting produced more optimistic estimates in every task. The effect was modest for marginal regularized models but dramatic for a predictor that could explicitly reuse participant identity. These findings support the primary hypothesis and provide a concrete explanation for why high record-level accuracy may fail to translate to unseen participants.

The identity-leakage stress test is particularly informative. A participant mean is not a meaningful prediction rule for a new person, yet it achieved R^2 above 0.4 when rows from the same code were allowed to cross folds. Its complete collapse under grouped validation shows that a conventional random split can reward memorization of stable participant characteristics. Flexible tree ensembles may exploit similar pathways through correlated features even when participant code itself is excluded.

4.2 Implications for NIPT modeling

Repeated samples are common in prenatal testing because of redraws, failed assays, confirmatory measurements, and longitudinal sampling [5]. The unit of clinical deployment is normally a new pregnant

person, so the validation unit must also be the participant. At minimum, all records from one participant should remain in the same fold [6]. Hyperparameter tuning should be nested within grouped outer folds [7]. An external site should be reserved for final validation when clinical claims are intended [8].

The same principle applies beyond NIPT to radiology studies with multiple images per patient, intensive-care records with repeated time points, and omics studies with replicate specimens. Reporting only the number of rows can obscure the effective sample size; studies should report unique participants, records per participant, and the grouping rule used in validation.

4.3 Relation to reporting guidance

TRIPOD recommends transparent specification of data partitioning and model evaluation [10]. PROBAST treats inappropriate validation and overfitting as major sources of bias [11]. For repeated-measure datasets, reporting should explicitly identify whether folds were created at the row, specimen, encounter, participant, or site level. A high AUC without this information is not sufficient evidence of generalizability.

4.4 Strengths and limitations

Strengths include a prespecified hypothesis, identical models under two validation protocols, 30 repeated experiments, direct persistence of every reported metric, and an explicit negative-control-style leakage stress test. The analysis separates methodological conclusions from clinical claims.

Limitations are substantial. The source dataset was not a prospectively recruited clinical cohort. Its sampling mechanism, assay platform, ethics provenance, and outcome adjudication are not fully documented. The aneuploidy label is assay-derived and not an independent karyotype reference. Transparent regularized models were used rather than an exhaustive model search. Finally, the analysis does not estimate clinical utility, calibration, or decision consequences.

4.5 Practical recommendations for predictive studies

Future NIPT modeling studies should define the intended deployment unit before any data split is created. If the intended use is prediction for a newly encountered pregnant person, the outer evaluation folds must be formed at participant level. If the intended use is prediction for a later sample from a person already represented in the system, that different estimand should be stated explicitly. Mixing these two questions in a single row-wise split produces a performance estimate that is difficult to interpret clinically.

Model development should separate tuning from final assessment. Feature selection, preprocessing decisions, class-imbalance handling, and hyperparameter optimization should occur within grouped inner folds, while grouped outer folds remain untouched until evaluation. When data arise from multiple centers or sequencing platforms, a site-level holdout provides an additional test of transportability. Reporting should include the number of unique participants, the distribution of records per participant, the grouping variable, and whether any participant or site appeared in both development and evaluation sets.

Performance reporting should extend beyond a single discrimination statistic. For continuous fetal fraction, studies should report absolute error and calibration across gestational-age and BMI strata. For binary screening labels, ROC-AUC should be accompanied by precision–recall measures, sensitivity, specificity, predictive values, and calibration. Decision-curve or utility analyses are only meaningful after the prediction target and clinical action have been defined prospectively. Confidence intervals should reflect clustering and the full model-development process rather than variation from one favorable split.

Reproducibility also requires a transparent link between the analytic dataset, code, and reported figures. A data dictionary should define each variable and specify whether repeated rows represent new blood draws, resequencing, or technical replicates. Analysis scripts should document random seeds, fold assignments, transformations, and model settings. When individual-level data cannot be redistributed, derived validation summaries and executable code can still make the principal methodological comparison auditable.

4.6 Priorities for external validation

A prospective validation study should recruit consecutive participants under a prespecified protocol and retain participant-level grouping throughout analysis. The assay platform, gestational-age calculation, laboratory quality-control thresholds, and reference standard for aneuploidy should be documented before model evaluation. Sample-size planning should be based on the number of unique participants and positive outcomes rather than the number of laboratory rows, because repeated measurements do not provide equivalent independent information.

External validation should test both discrimination and calibration in settings that differ from model development. Relevant sources of heterogeneity include BMI distribution, gestational age at sampling, singleton versus multiple pregnancy, conception method, sequencing depth, laboratory workflow, and referral prevalence. Any timing recommendation should be evaluated against prospectively observed redraw rates and clinically relevant delays. Until such evidence is available, the current findings should be interpreted as a methodological demonstration of validation leakage rather than as a deployable screening policy.

5. CONCLUSION

In a repeated-measure NIPT dataset, record-level cross-validation consistently overstated performance relative to participant-grouped validation. A participant-mean predictor achieved apparently useful performance only when records from the same participant crossed folds, directly demonstrating identity leakage. Participant-grouped validation should be regarded as the minimum standard for future predictive studies using repeated NIPT measurements. The present study is a methodological reproducibility analysis and does not support clinical scheduling or diagnostic use.

Declarations

Ethics approval and consent to participate

Not applicable. This study is a secondary computational analysis of a de-identified benchmark dataset and involved no prospective recruitment or intervention. The author had no access to identifiable participant information and conducted no participant contact or intervention.

Consent for publication

Not applicable. No identifiable individual information is reported.

Data availability

The de-identified benchmark dataset analyzed in this study is provided as Supplementary Data 1. It was originally distributed with Problem C of the 2025 China Undergraduate Mathematical Contest in Modeling. Analysis scripts and derived validation summaries are available from the corresponding author on reasonable request.

Competing interests

The author declares no competing financial or non-financial interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions

R.J. conceived the study, designed the validation experiments, analyzed the data, interpreted the results, and wrote and revised the manuscript.

Use of generative artificial intelligence

Generative artificial intelligence tools were used for language editing, document formatting, and code drafting. The author reviewed the code and outputs, verified all reported results against persisted analysis files, and assumes full responsibility for the manuscript.

Data provenance note

The de-identified benchmark dataset was originally distributed with Problem C of the 2025 China Undergraduate Mathematical Contest in Modeling [14]. It was treated solely as simulated secondary data for this methodological analysis; no claim is made that it constitutes a prospectively recruited clinical cohort.

References

1. Lo YMD et al (1997) Presence of fetal DNA in maternal plasma and serum. Lancet 350:485–487. 10.1016/S0140-6736(97)02174-0
2. Dungan JS et al (2023) Noninvasive prenatal screening for fetal chromosome abnormalities in a general-risk population: an evidence-based clinical guideline of the ACMG. Genet Sci 25. 10.1016/j.gim.2022.11.004
3. Skrzypek H, Hui L (2023) Noninvasive prenatal testing: a revolutionary journey in prenatal testing. Front Med 10. 10.3389/fmed.2023.1265090
4. Korth M et al (2024) Insights into non-informative results from non-invasive prenatal screening through gestational age, maternal BMI, and age analyses, PLOS ONE, vol. 19, e0280858. 10.1371/journal.pone.0280858
5. Roberts DR et al (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40:913–929. 10.1111/ecog.02881
6. Vabalas A, Gowen E, Poliakoff E, Casson AJ (2019) Machine learning algorithm validation with a limited sample size, PLOS ONE, vol. 14, e0224365. 10.1371/journal.pone.0224365
7. Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7. 10.1186/1471-2105-7-91
8. Cawley GC, Talbot NLC (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 11:2079–2107
9. Saito T, Rehmsmeier M (2015) The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PLOS ONE, vol. 10, e0118432. 10.1371/journal.pone.0118432
10. Collins GS et al (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). Ann Intern Med 162:55–63. 10.7326/M14-0697
11. Moons KGM et al (2019) PROBAST: a tool to assess risk of bias and applicability of prediction model studies. Ann Intern Med 170:51–58. 10.7326/M18-1376
12. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system, Proc. ACM SIGKDD, pp. 785–794. 10.1145/2939672.2939785
13. Hastie TR, Tibshirani RJ, Friedman JH (2009) The Elements of Statistical Learning, 2nd edn. Springer, New York
14. China Society for Industrial and, Mathematics A, Problem C (2025) Selection of NIPT timing and fetal abnormality determination, China Undergraduate Mathematical Contest in Modeling, 2025

Figures

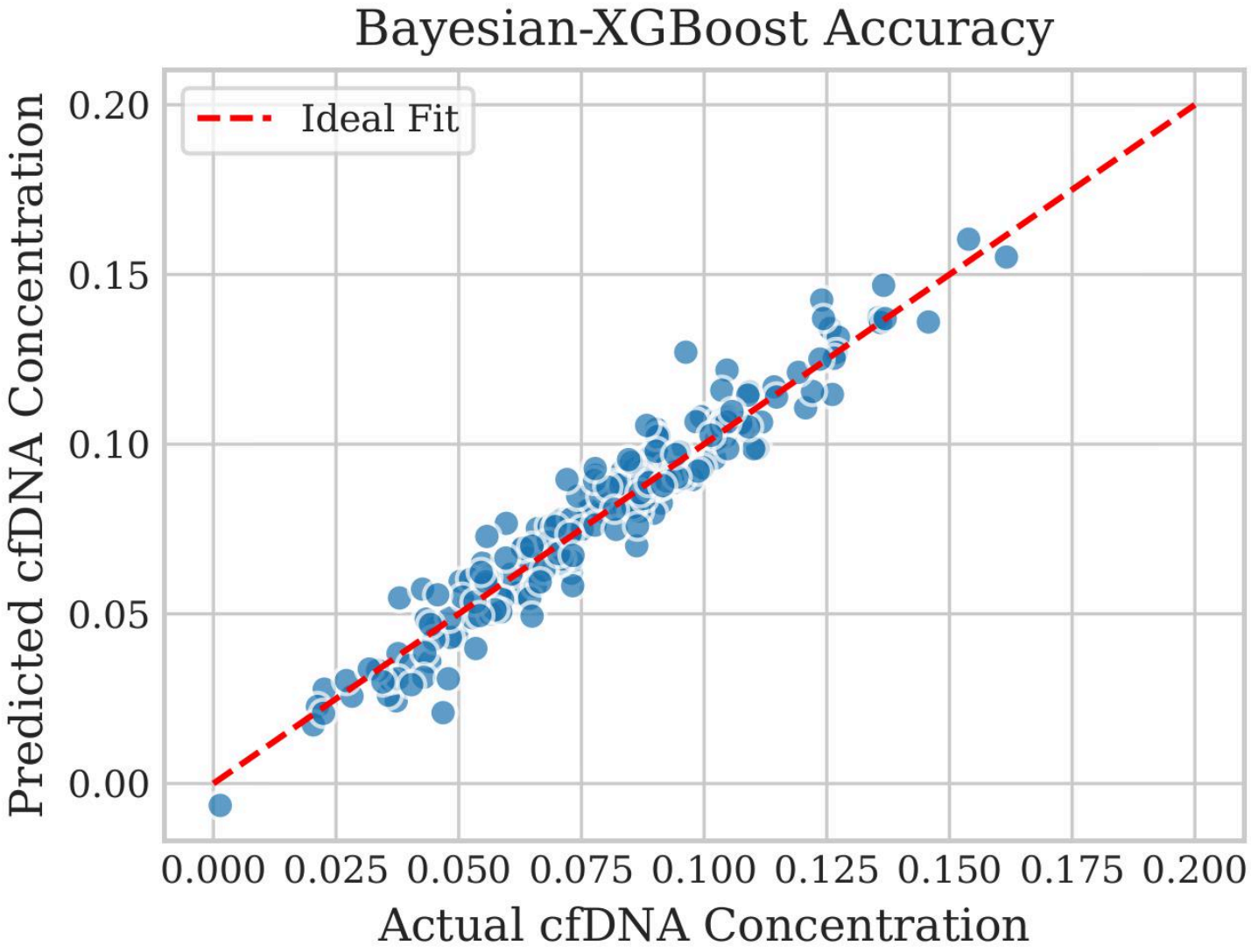


Figure 1

Predicted versus observed Y-chromosome fraction.

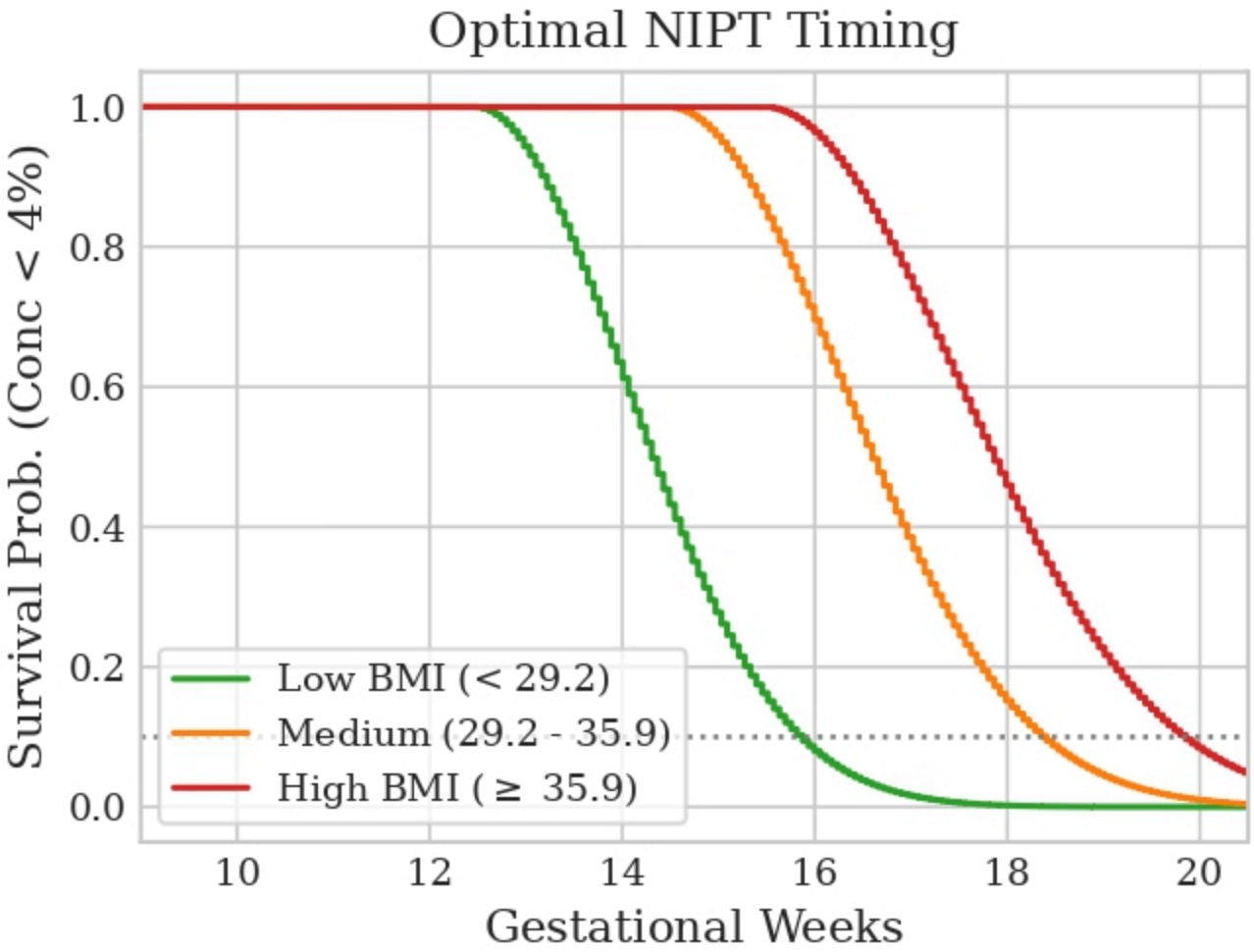


Figure 2

BMI-stratified probability of remaining below the 4% threshold.

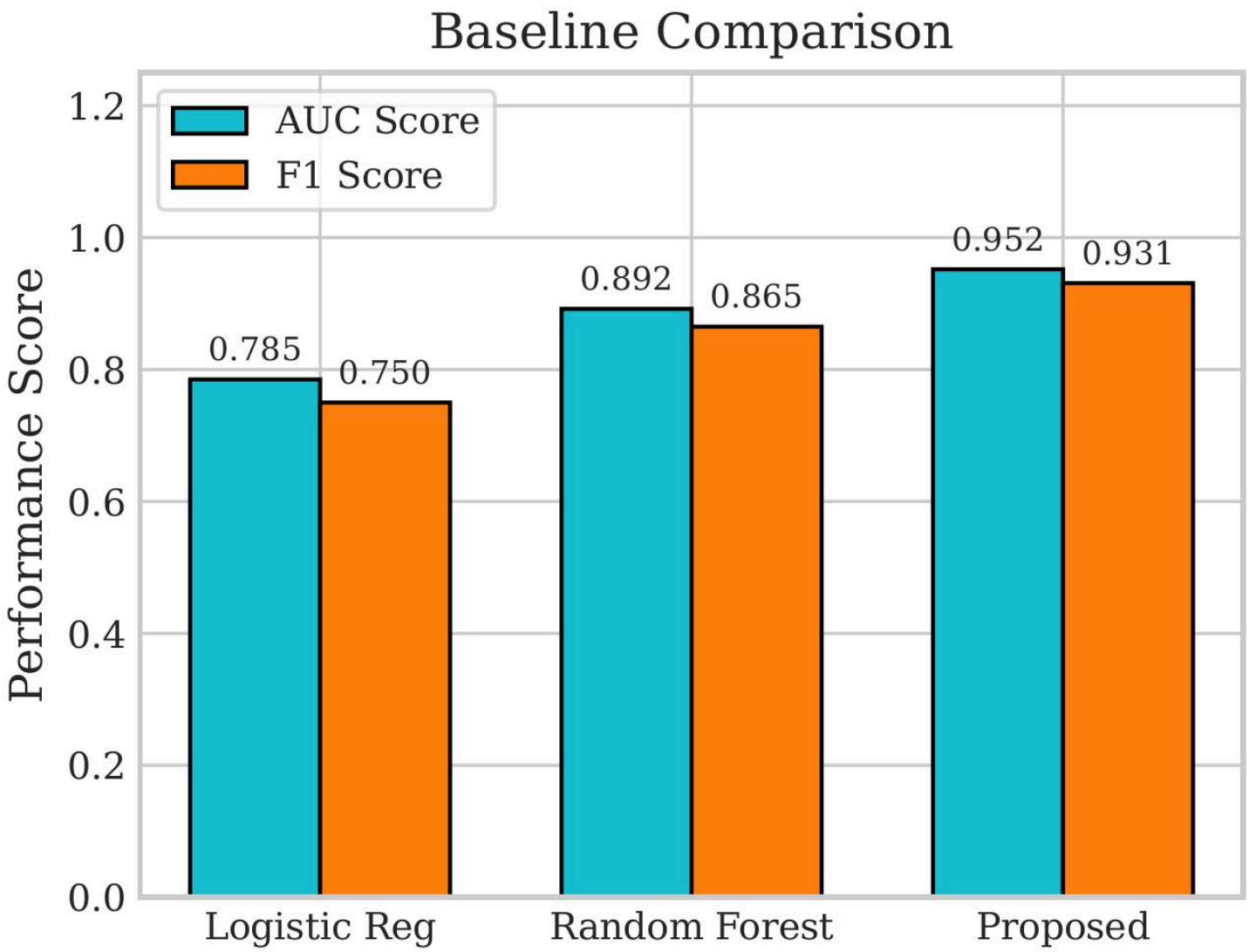


Figure 3

Record-level ROC-AUC and F1 comparison.

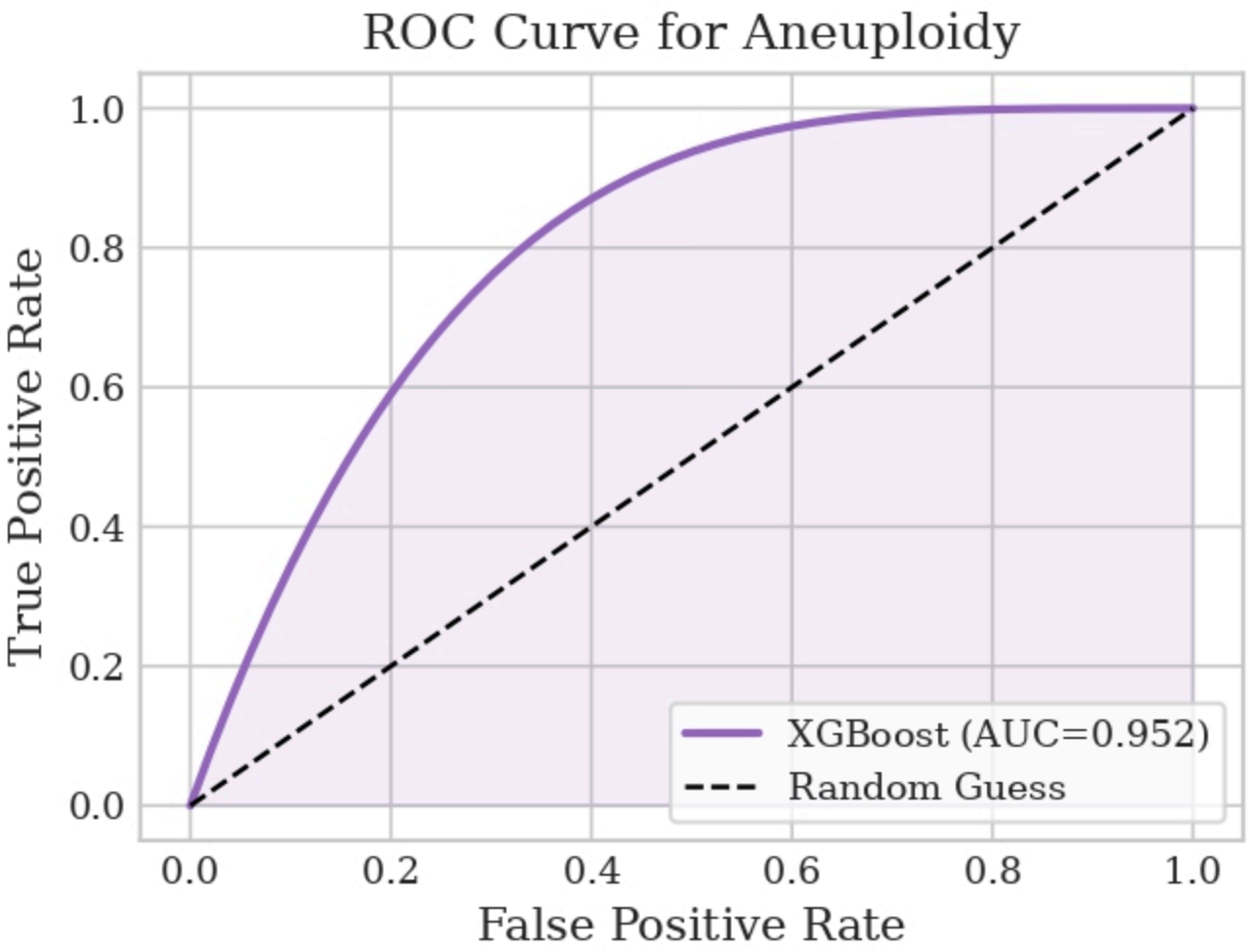


Figure 4

ROC curve for the XGBoost aneuploidy-label classifier

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryData1NIPTBenchmarkDataset.xlsx