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.

Temporal external validation of random forest and XGBoost models for identifying concurrent diabetic retinopathy among hospitalized patients with type 2 diabetes mellitus: a single-centre temporal external validation study

Yiting Zhou
Nanjing Drum Tower Hospital, China Pharmaceutical University
Nan Gao
Beijing University of Chinese Medicine Third Affiliated Hospital
Yafei Luo
Nanjing Drum Tower Hospital, China Pharmaceutical University
Lintong Li
Nanjing Drum Tower Hospital, the Affiliated Hospital of Nanjing University Medical School
Lingjun Sun
Nanjing Drum Tower Hospital, the Affiliated Hospital of Nanjing University Medical School
Cheng Ji
getcct@sina.com

Nanjing Drum Tower Hospital, the Affiliated Hospital of Nanjing University Medical School

Research Article

Keywords: diabetic retinopathy, machine learning, random forest, XGBoost, temporal external validation, model calibration

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11374928/v1

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

Additional Declarations: No competing interests reported.

Manuscript title

Temporal external validation of random forest and XGBoost models for identifying concurrent diabetic retinopathy among hospitalized patients with type 2 diabetes mellitus: a single-centre temporal external validation study

Authors and affiliations

Zhou Yiting[1]†, Gao Nan[3]†, Luo Yafei[1]†, Li Lintong[2], Sun Lingjun[2]*, Ji Cheng[2]*

[1] Department of Pharmacy, Nanjing Drum Tower Hospital, China Pharmaceutical University, Nanjing, Jiangsu 210008, China

[2] Department of Pharmacy, Nanjing Drum Tower Hospital, the Affiliated Hospital of Nanjing University Medical School, Nanjing, Jiangsu 210008, China

[3] Department of Pharmacy, Beijing University of Chinese Medicine Third Affiliated Hospital, Beijing, China

* Corresponding authors: Sun Lingjun, buffet520@126.com; Ji Cheng, getcct@sina.com

† These authors contributed equally to this work.

Abstract

Background: Machine-learning prediction models for diabetic retinopathy are frequently reported with promising internal performance, yet high-quality temporal external validation remains scarce. We aimed to conduct temporal external validation of locked random forest (RF) and XGBoost models for identifying concurrent diabetic retinopathy (DR) in hospitalized patients with type 2 diabetes mellitus (T2DM).

Methods: This single-centre temporal external validation study used a development cohort of 3 874 hospitalized T2DM patients (2018-2025), split into training (n = 2 711) and internal test (n = 1 163) datasets. After finalising all predictors, pre-processing, hyper-parameters and classification thresholds, we validated the fixed models in 670 consecutive T2DM inpatients enrolled June-October 2025. We assessed discrimination (ROC-AUC, PR-AUC), Brier score, calibration metrics, fixed-threshold classification indices and decision-curve-analysis-derived net clinical benefit. Standardised mean differences quantified predictor distribution shifts between cohorts. paired DeLong method compared ROC-AUC within each cohort.

Results: In the internal test set, RF and XGBoost showed comparable discrimination (ROC-AUC 0.880 versus 0.877, P = 0.607). In temporal external validation, both models experienced performance decay; RF achieved significantly higher ROC-AUC than XGBoost (0.807 vs 0.783, P = 0.015). Only TyG-BMI exhibited substantial inter-cohort distribution shift (|standardised mean difference| = 0.419). Both models systematically underestimated DR risk but displayed heterogeneous miscalibration patterns: calibration slope 1.544 for RF and 0.668 for XGBoost. Positive net clinical benefit was observed for both models across low-to-moderate threshold probabilities. Sensitivity markedly decreased for pre-set classification thresholds in validation data.

Conclusions: Under temporal external validation, the locked RF model outperformed XGBoost for identifying concurrent DR in hospitalised T2DM patients. Both models underestimated disease risk and presented divergent miscalibration. Further multi-institutional external validation is required to establish real-world generalisability.

Keywords: diabetic retinopathy; machine learning; random forest; XGBoost; temporal external validation; model calibration

List of abbreviations

Abbreviation	Full name
apoAI	Apolipoprotein AI
apoB	Apolipoprotein B
BMI	Body mass index
CI	Confidence interval
CITL	Calibration-in-the-large
DCA	Decision-curve analysis
DBIL	Direct bilirubin
DR	Diabetic retinopathy
GA	Glycated albumin
GGT	γ-glutamyl transferase
Hb	Haemoglobin
IQR	Interquartile range
MICE	Multiple imputation by chained equations
NPV	Negative predictive value
PPV	Positive predictive value
PR-AUC	Area under the precision-recall curve
RF	Random forest
ROC-AUC	Area under the receiver operating characteristic curve
SMD	Standardised mean difference
T2DM	Type 2 diabetes mellitus
TP	Total protein
TyG-BMI	Triglyceride-glucose body-mass index
UmAlb-24h	24-h urinary microalbumin

Background

Diabetic retinopathy (DR) is a common microvascular complication of type 2 diabetes mellitus (T2DM) and a leading cause of preventable vision loss among working-age adults [1]. Early-stage DR is often asymptomatic, which highlights the necessity of regular fundus ophthalmic screening for timely case detection. Routine laboratory and clinical information collected during hospital admission may serve as supplementary clues to identify patients with concurrent DR and facilitate timely ophthalmology referrals [2].

Ensemble machine-learning algorithms including random forest (RF) and XGBoost are widely adopted for building DR risk-prediction models [3,4]. However, most existing publications only report internal validation results, while rigorous external and temporal external validation is frequently missing [6]. Previous systematic reviews have pointed out that insufficient external validation constitutes a major methodological limitation in this research field [6]. Satisfactory performance observed in model development cannot guarantee comparable discrimination and calibration when models are applied to unseen real-world cohorts; evaluation on fully independent datasets is therefore indispensable for prediction-model research [5].

Patient characteristics, clinical workflows, measurement procedures and treatment patterns change over calendar time, producing dataset shift. Dataset shift can degrade discrimination and calibration performance of previously-trained prediction models, even when internal test-set performance is excellent [7]. Temporal external validation evaluates pre-locked prediction models on prospectively-collected later patient cohorts and tests the temporal transportability of prediction tools.

Following our prior model-development work, the present study aimed to perform temporal external validation for fully-locked RF and XGBoost models built for identifying concurrent DR among hospitalized T2DM patients. No retraining, recalibration or threshold adjustment was performed during validation. We comprehensively evaluated discriminative power, global prediction error, calibration behaviour, fixed-threshold classification performance and clinical net benefit derived from decision-curve analysis, and described predictor distribution drift between development and validation cohorts. This study was reported following the TRIPOD reporting guideline for clinical prediction models.

Methods

Study design and setting

This was a single-centre temporal external validation observational study, conducted at Nanjing Drum Tower Hospital, the Affiliated Hospital of Nanjing University Medical School. The model-development source cohort was derived from hospitalized patients with type 2 diabetes mellitus (T2DM) receiving inpatient care between January 2018 and May 2025. A total of 3 874 eligible patients formed the model-development cohort. Using stratified random splitting according to diabetic retinopathy (DR) status, participants were divided into a training set (n = 2 711) used for model building and hyper-parameter tuning and an independent internal test set (n = 1 163) that served as a fixed performance benchmark for the model-development phase.

After completing model building and finalization, all input predictors, data-preprocessing protocols, hyper-parameters and classification thresholds were permanently locked with no further modification. For temporal external validation, we consecutively enrolled 670 new T2DM inpatients admitted from June 2025 to October 2025. No model refitting, no recalibration and no readjustment of classification thresholds were performed before applying the locked models onto this temporal external validation cohort.

Participants

Diagnosis of T2DM was established according to 1999 World Health Organization criteria [8]. Eligible participants were aged 18 years or older and had unambiguously documented DR diagnostic status from inpatient clinical records and ophthalmic examinations.

Major exclusion criteria were as follows: type 1 diabetes mellitus, gestational diabetes or other specific subtypes of diabetes; non-diabetic retinal diseases that would interfere with fundus evaluation; severe systemic conditions that substantially alter patients' metabolic profiles; unresolvable missing values for key study variables; indeterminable DR diagnostic status. For patients with multiple separate hospital admissions that fulfilled inclusion criteria within the study time windows, only the earliest admission record was retained for analysis to avoid duplicated observations.

The internal test set represented independent samples collected during the original model-development time period. The temporal external validation cohort consisted of entirely new patients enrolled in subsequent calendar time, representing temporally independent real-world samples.

Outcome definition and input predictors

The binary primary study outcome was concurrent (prevalent) DR identified at the index hospitalisation, rather than incident future-onset DR or disease progression over follow-up. DR status was abstracted from inpatient clinical diagnostic records and corresponding ophthalmic examination reports. Diagnosis and severity grading followed the International Clinical Diabetic Retinopathy Disease Severity Scale [9]. Patients showing retinal lesions meeting DR diagnostic criteria in either eye were allocated to the DR positive group; all remaining patients were assigned to the non-DR group.

Twelve fixed input predictors were selected from routine clinical and laboratory measurements collected during the index admission, including: 24-h urinary microalbumin (UmAlb-24h), apolipoprotein B (apoB), apolipoprotein AI (apoAI), direct bilirubin (DBIL), body mass index (BMI), urea, glycated albumin (GA), total protein (TP), triglyceride-glucose body-mass index (TyG-BMI), haemoglobin (Hb), γ-glutamyl transferase (GGT), and fasting C-peptide.

Data pre-processing

Every step of data pre-processing was fully defined during the model-development phase and applied identically to both the internal test set and the temporal external validation cohort.

During processing of the training dataset, we pre-screened candidate predictors according to their missing proportions. Variables with a missing fraction ≥ 30 % in the training set were excluded from the predictor pool. Remaining missing values were handled using multiple imputation by chained equations (MICE) [10]. Predictive mean matching (PMM) was applied for continuous variables, while logistic regression was used for imputation of binary variables. Five separate imputed datasets were generated, with three imputation iterations for each dataset. Following the pre-established analytical pipeline, continuous variables were aggregated by taking mean values across the five imputed datasets; mode values were used for categorical variables when combining imputed datasets.

Outlier handling for continuous variables was performed by Winsorization. The inter-quartile range (IQR) boundaries used for Winsorization were exclusively calculated from training-set data only. Observations below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ were clamped to the respective boundary values. Categorical variables were coded strictly following coding rules defined in the model-development phase.

In the temporal external validation cohort, data completeness for the 12 locked predictors was high before imputation. Only one entry for TyG-BMI was missing (0.15 %), while the other 11 predictors contained zero missing records. This single missing value for TyG-BMI was processed according to the locked MICE pre-processing workflow described above.

Model specifications and classification thresholds

Hyper-parameter tuning for the random forest (RF) model in the development phase combined repeated 10-fold cross-validation (10 folds × 2 repeats) with Bayesian optimisation; ROC-AUC was chosen as the objective metric for model selection. The final locked RF model contained 1 000 decision trees, with hyper-parameters mtry = 4 and min_n = 31.

For the XGBoost model, hyper-parameter optimisation was also performed using 10-fold cross-validation coupled with Bayesian optimisation, also maximising ROC-AUC. The final locked hyper-parameters for XGBoost were: trees = 222, min_n = 7, tree_depth = 4, learn_rate = 0.0519, loss_reduction = 0.00289, sample_size = 0.818. All model fitting was completed within R version 4.5.1 environment using the tidymodels framework.

Classification thresholds for both models were derived from out-of-fold predicted probabilities obtained during cross-validation on the training dataset. Different threshold-selection strategies were used for the two models. For RF, the threshold (0.380) was selected by maximising the Youden index. For XGBoost, a sensitivity-prioritised selection strategy was adopted: candidate thresholds yielding sensitivity ≥ 0.80 were first screened, and among those candidates the threshold delivering the highest possible specificity was selected, giving the final XGBoost threshold of 0.396. The 0.80-sensitivity constraint only served as a screening criterion for threshold selection and exerted no influence over hyper-parameter optimisation.

After finalising predictors, pre-processing rules, hyper-parameters and classification thresholds, both models were directly applied to the internal test set and temporal external validation cohort. No refitting, recalibration or threshold adjustment was executed in the validation procedures. Because sensitivity, specificity and other classification-table metrics are strongly threshold-dependent, inter-model comparisons of discriminative performance relied primarily on threshold-independent ROC-AUC.

Statistical analysis

Model evaluation covered discriminative capacity, global prediction error,

calibration performance, fixed-threshold classification metrics, and clinical net benefit obtained from decision-curve analysis [11].

Discriminative performance was quantified by ROC-AUC with 95 % confidence intervals computed by the DeLong method [12]. PR-AUC calculated as average precision was used as supplementary discrimination metric, with cohort-specific DR prevalence taken as the baseline reference. Within each separate cohort, the paired DeLong method was used for head-to-head comparison of ROC-AUC between RF and XGBoost; AUC differences together with corresponding 95 % CIs were reported. Changes of ROC-AUC, PR-AUC and Brier score between the internal test set and temporal external validation cohort were interpreted descriptively without formal statistical hypothesis testing. The paired DeLong method was only applied for model-versus-model comparison within the identical patient cohort.

Confusion matrices were built based on the locked classification thresholds inherited from training-phase cross-validation. We calculated accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score and false-negative rate. Wilson-score intervals were used to compute 95 % CIs for accuracy, sensitivity, specificity, PPV, NPV and false-negative rate; F1-score was reported as a point estimate only [13].

Global overall prediction error was quantified using the Brier score. Calibration assessment included calibration intercept, calibration slope and calibration-in-the-large (CITL) [14]. Calibration intercept and slope were estimated from logistic recalibration models built upon logit-transformed model predicted probabilities. CITL represented overall probability miscalibration under the constraint that calibration slope is fixed to 1. Under ideal calibration conditions, calibration intercept and CITL approach zero and calibration slope approximates one.

Decision-curve analysis (DCA) was performed to assess model-derived clinical net benefit across varying threshold probabilities; “treat-all” and “treat-none” strategies were set as reference benchmarks [15].

Standardised mean differences (SMDs) were calculated for all 12 input predictors to quantify distributional discrepancies between the full development cohort (n = 3 874) and temporal external validation cohort (n = 670). SMD was also calculated for the difference in DR positive proportion. Absolute SMD values were reported. An $|SMD| \geq 0.10$ was adopted as reference threshold indicating potentially meaningful univariate distribution divergence [16].

All statistical hypothesis tests were two-sided, and $P < 0.05$ was defined as the threshold for statistical significance.No formal power calculation was performed, as this temporal external validation study used all consecutive eligible inpatients available within the pre-specified study time window.

Results

Temporal external validation cohort and distributions of input variables

Participant recruitment and dataset splitting workflows are illustrated in **Figure 1**. A total of 3 874 hospitalized T2DM patients were included within the development cohort, among whom 1 499 were DR-positive and 2 375 were DR-negative. Dataset splitting assigned 2 711 subjects to the training set and 1 163 subjects to the independent internal test set. Subsequent to model finalization, 670 consecutive inpatients hospitalized between June and October 2025 formed the temporal external validation cohort, containing 265 DR-positive patients and 405 DR-negative patients.

Distributions of 12 locked input predictors for development and temporal external validation cohorts are summarized in **Table 1**, while absolute SMD values are visualized in Figure 2. The largest distributional discrepancy appeared for TyG-BMI, with median (P25, P75) of 234.45 (207.18, 274.12) in the development cohort versus 222.49 (197.98, 250.83) in temporal external validation cohort ($|SMD| = 0.419$). Absolute SMD values for the remaining 11 input variables were all below 0.10. DR positive proportions reached 38.69% and 39.55% in development and temporal external validation cohorts respectively ($|SMD| = 0.018$).

Internal test results

Within 1 163 subjects of the internal test set, ROC-AUC values were 0.880 (95% CI, 0.859–0.901) for RF and 0.877 (95% CI, 0.856–0.899) for XGBoost. The paired DeLong method yielded an AUC difference of 0.0024 (95% CI, −0.0069 to 0.0118; $P = 0.607$). PR-AUC values were 0.859 for RF and 0.865 for XGBoost. Corresponding Brier scores stood at 0.1378 and 0.1276.

Under fixed classification thresholds, accuracy, sensitivity and specificity were 0.811, 0.769 and 0.837 for RF; corresponding values for XGBoost were 0.814, 0.747 and 0.857. F1-scores were 0.759 for RF and 0.757 for XGBoost. Full classification performance metrics are listed in **Table 2**.

Discrimination and classification performance in temporal external validation

For the temporal external validation cohort containing 670 patients, ROC-AUC

was 0.807 (95% CI, 0.773 – 0.840) for RF and 0.783 (95% CI, 0.746 – 0.819) for XGBoost (**Figure 3**). The paired DeLong method produced an AUC difference of 0.024 (95% CI, 0.005 – 0.043; P = 0.015). PR-AUC values were 0.781 for RF and 0.766 for XGBoost. Corresponding precision-recall curves are provided in Supplementary **Figure S2**. At respective locked classification thresholds, RF achieved accuracy of 0.751, sensitivity of 0.585 and specificity of 0.859. XGBoost delivered accuracy of 0.733, sensitivity of 0.517 and specificity of 0.874. PPV, NPV and F1-score were 0.731, 0.760 and 0.650 for RF, compared with 0.729, 0.734 and 0.605 for XGBoost. RF generated 110 false-negative cases (false-negative rate = 41.5%), and XGBoost generated 128 false-negative cases (false-negative rate = 48.3%). Complete classification metrics are presented in **Table 3**.

Changes in performance from internal testing to temporal external validation

Descriptive shifts of probability-based performance metrics between internal test set and temporal external validation cohort are summarized in **Figure 4** and **Table 4**. RF ROC-AUC declined from 0.880 to 0.807 (Δ = −0.073). XGBoost ROC-AUC decreased from 0.877 to 0.783 (Δ = −0.095). PR-AUC dropped from 0.859 to 0.781 for RF and from 0.865 to 0.766 for XGBoost. Brier score increased from 0.1378 to 0.1757 for RF and from 0.1276 to 0.1840 for XGBoost.

Calibration and decision curve analysis

Observed DR prevalence within temporal external validation cohort reached 39.55%. Mean model-predicted probabilities were 33.3% for RF and 29.8% for XGBoost. For RF, calibration intercept, calibration slope and CITL were 0.733 (95% CI, 0.468–0.998), 1.544 (95% CI, 1.295–1.794), and 0.328 (95% CI, 0.158–0.498). Corresponding values for XGBoost were 0.338 (95% CI, 0.106–0.570), 0.668 (95% CI, 0.551–0.785), and 0.756 (95% CI, 0.554–0.957) (**Table 5**).

Mean predicted probabilities for both models fell below observed DR prevalence, accompanied by positive CITL values, which indicated systematic underestimation of event probability within temporal validation samples. Calibration slope exceeded 1 for RF and remained below 1 for XGBoost, demonstrating distinct miscalibration patterns across two algorithms. Brier scores were 0.1757 for RF and 0.1840 for XGBoost. Calibration curves are displayed in **Figure S1**.

Results from decision-curve analysis are shown in **Figure S3**. Positive net benefit was documented for both models across most low-to-moderate threshold

probability ranges. Numerically higher net benefit was observed for RF over multiple threshold intervals.

Discussion

Similar performance in internal testing was not fully maintained in temporal external validation

Temporal external validation for RF and XGBoost models was performed under fixed input variables, pre-processing pipelines, hyper-parameters and classification thresholds. Comparable discriminative outputs were generated by the two algorithms on internal-test-set data sourced from model-development work. No statistically significant inter-model gap existed in this dataset. A drop in ROC-AUC could be seen for both models after deployment on temporally separated validation samples. Statistically stronger discriminative capacity was detected in RF compared with XGBoost. The scale of shifts in PR-AUC and Brier-score varied between algorithms. Both prediction tools showed systematic underestimation of predicted probability alongside heterogeneous miscalibration patterns.

These findings demonstrate that promising performance documented in the model-development phase cannot be directly extrapolated to patient cohorts sampled from later time periods. Despite nearly equivalent discrimination for RF and XGBoost in internal hold-out data, relative performance ranking shifted within temporally independent samples. Such observations underline the necessity of temporal independent evaluation for pre-developed clinical prediction models.

Threshold-selection workflows differed for RF and XGBoost. Classification indices including sensitivity and specificity measured at fixed operating points arise from a combination of algorithm properties and matching cut-off setups. For this reason, inter-model comparisons of discriminative capacity draw mainly from ROC-AUC outputs alongside paired DeLong method outcomes.

Changes in input distributions and calibration

Machine-learning models may suffer performance decay when transferred to temporally new patient cohorts. Possible triggers include shifts in patient composition, clinical practice routines, measurement workflows, or changed association patterns linking predictors and study endpoints. These triggers can produce multiple forms of dataset shift [7]. Univariate distributional characteristics of the 12 locked predictors were compared across development and temporal validation cohorts. Substantial

distribution divergence was detected only for TyG-BMI (|SMD| = 0.419). Minor univariate distributional differences existed for the other 11 predictors. DR positive rates displayed high similarity across the two cohorts.

Widespread univariate distributional shift among input predictors was not detected, even though performance deterioration took place. SMD statistics, however, only describe marginal distribution gaps for single variables. This metric fails to capture shifts in the joint distribution of multiple predictors. It also cannot reflect time-dependent adjustments to predictor-outcome correlation structures. Reliance purely on univariate SMD is inadequate for designing future model-updating strategies. Surveillance frameworks ought to track drift in model discrimination and calibration as well.

Calibration outputs uncovered divergent probability-prediction behaviours from the two models within temporal validation datasets. DR prevalence stood at 39.55% in the temporal external validation cohort. Mean predicted probabilities were 33.3 % for RF and 29.8 % for XGBoost. Positive CITL values signal broad underestimation of DR risk probability. RF returned a calibration slope above 1, while the slope for XGBoost fell below 1. Established interpretive frameworks assign specific meaning to these slope values. A calibration slope below 1 signals overly extreme predicted risk values. A slope above 1 reflects insufficient variance among predicted probabilities [14]. For the target temporal cohort, the two locked prediction models exhibited heterogeneous miscalibration instead of a consistent pattern of calibration degradation.

Comparison with previous external validation studies

Logistic regression, RF, XGBoost and other analytical methods see broad application in DR-focused clinical prediction research. Even so, the general quality of model validation still needs enhancement. A 2025 systematic review completed by Huang et al. synthesized outputs from 15 DR prediction-model investigations. All included studies were rated with high overall bias risk, and insufficient external validation represented a major limitation for this research field [6]. A separate systematic review targeted machine-learning models assessing DR progression risk. It summarized 13 published works and likewise noted shortages of authentic longitudinal validation frameworks [17]. Included works covered diverse prediction tasks such as identification of concurrent DR and forecasting future disease progression. Direct cross-study comparison of AUC values cannot be performed

without full consideration of such task heterogeneity.

Published independent-validation results indicate no universal trend for performance alteration of machine-learning models on unseen datasets. Wan et al. constructed DR prediction models based on routine laboratory indicators. An AUC of 0.831 was achieved for the XGBoost model within the development dataset, whereas AUC decreased to 0.709 in an independent external validation sample containing 323 subjects [18]. By contrast, a 2026 multicenter investigation by He et al. reported that XGBoost maintained AUC near 0.86 in external validation data, with development-phase AUC approximating 0.87 [19]. Cumulative evidence demonstrates external-cohort performance is jointly shaped by algorithm attributes, target-population characteristics and underlying data-generation mechanisms. Real-world performance cannot be inferred merely from development-phase metrics or algorithm categories.

A major strength of the present investigation lies in complete locking of input predictors, hyper-parameters, data-processing workflows and classification operating points prior to temporal validation. Direct observation of pre-built-model behaviours on temporally later patient samples becomes feasible under such study setup. Degrees of discrimination decline and elevation of overall prediction error differed across RF and XGBoost. This observation further stresses the necessity of performance assessment on target real-world cohorts instead of simple extrapolation from algorithm comparisons obtained in model-development work.

Multidimensional model evaluation and potential implications for use

ROC-AUC alone is insufficient to mirror full-scale model performance. Evaluation frameworks built for clinical prediction models should combine discrimination, calibration, global prediction error, classification outputs at fixed operating points, as well as clinical decision-making value [11]. To address this requirement, PR-AUC, Brier score, calibration intercept, calibration slope, CITL and DCA outputs were reported. These indices serve to characterise behaviours of the two locked models across multiple dimensions.

When viewed from global-prediction-error perspective, RF yielded a numerically lower Brier score (0.1757) compared against XGBoost (0.1840) in temporal external validation. Even so, general underestimation of DR probability was detected for both models, paired with calibration-slope deviations running in opposite directions. Calibration assessment demands joint interpretation of mean predicted probability,

CITL, calibration slope, Brier score alongside calibration curves. Reliance on one single calibration metric for final judgement should be avoided [14].

DCA outputs detected positive net benefit for both models across most low-to-moderate threshold-probability ranges. RF generated numerically higher net-benefit values within several intervals. Net-benefit metrics quantify trade-offs: gains originating from true-positive detection against costs caused by false-positive classification under given risk thresholds [15]. These results point toward the need for further quantified assessment of potential risk-stratification use cases. Existing evidence, however, cannot verify whether clinical deployment of these models delivers measurable improvements for patient outcomes. For the prediction endpoint set within this research, the models function merely as auxiliary reference information to spot concurrent DR. They cannot substitute standard ophthalmic screening workflows.

Substantial reduction of sensitivity at locked operating points was observed for both models in temporal validation data, with XGBoost sensitivity dropping to 0.517. This suggests classification performance of thresholds optimized in model-development phases may deteriorate after deployment on new patient cohorts, even when overall discrimination remains at acceptable levels. Future cross-institutional validation work should report not only ROC-AUC but also calibration metrics and clinically relevant fixed-threshold classification results. Decisions concerning recalibration or model updating ought to match practical application scenarios.

Limitations

Several methodological limitations should be acknowledged. First, both development and temporal external validation cohorts consisted of hospitalized T2DM patients recruited within a single medical institution. The temporal validation cohort contained 670 participants and covered approximately five-month enrolment duration. Study conclusions mainly reflect temporal-shift performance within one hospital setting. Independent multi-center, multi-region and multi-tier medical-setting validation remains absent. Distinct clinical profiles exist between hospitalized diabetes patients and outpatient or community-dwelling diabetes populations; generalizability toward broader target groups requires further verification.

Second, the finalized models incorporate 12 routine clinical and laboratory

variables, which cannot capture all risk factors correlated with DR status. Input features are predominantly static measurements collected during single hospitalization events, without integration of longitudinal dynamic clinical information. In addition, subgroup-stratified model performance has not been assessed.

Conclusion

Single-center temporal external validation was performed for existing RF and XGBoost prediction models, with model parameters, input features, preprocessing workflows and classification thresholds kept fixed. Comparable discriminative performance was achieved for the two algorithms within the internal test set. Higher ROC-AUC was documented for RF relative to XGBoost within the temporal external validation cohort enrolled from June to October 2025. Systematic underestimation of predicted DR probability and divergent miscalibration patterns were demonstrated for both models. As validation data were sourced from one single-center setting with limited time span, further validation across diverse healthcare institutions and longer observation periods is required to confirm model real-world performance for target populations.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Committee of Nanjing Drum Tower Hospital, the Affiliated Hospital of Nanjing University Medical School. All analyses were performed on retrospective anonymised routine clinical data. The ethics committee waived the requirement for individual patient informed consent. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Consent for publication

Not applicable. No identifiable individual patient data, clinical images or videos are presented within this manuscript.

Availability of data and materials

The raw patient dataset cannot be publicly released due to institutional data-management regulations and constraints defined within the ethical approval. De-identified analytical datasets may be obtained from the corresponding author upon reasonable formal request and compliance with local institutional governance requirements.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Clinical Research Special Fund of Nanjing Drum Tower Hospital, the Affiliated Hospital of Nanjing University Medical School (Grant No. 2021-LCYJ-PY-33). The funding body had no role in study design, data collection, statistical analysis, manuscript writing or the decision to submit for publication.

Authors' contributions

Zhou Yiting conceived the study, collected clinical data, conducted statistical analyses, generated figures and tables, and drafted the original manuscript. Gao Nan conceived the study, developed machine-learning models, performed data quality control and statistical analysis, and revised the manuscript. Luo Yafei organised study datasets, performed literature retrieval, optimised figures and tables, and revised the manuscript. Li Lintong screened patient cases and organised references, and reviewed the manuscript. Sun Lingjun supervised the project, critically revised the manuscript and acquired research resources. Ji Cheng supervised the overall research project, designed the study framework and critically revised the manuscript. All authors read and approved the final submitted manuscript.

Acknowledgements

We thank the clinical staff of Department of Endocrinology, Nanjing Drum Tower Hospital for routine clinical data collection. During the preparation of this work, a large-language model was used only for English-language polishing; no large-language model was used to generate study hypotheses, raw data or analytical results.

References

[1] Teo ZL, Tham YC, Yu M, et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. Ophthalmology. 2021;128(11):1580–1591. doi:10.1016/j.ophtha.2021.04.027.

[2] American Diabetes Association Professional Practice Committee for Diabetes. 12. Retinopathy, neuropathy, and foot care: Standards of Care in Diabetes—2026. Diabetes Care. 2026;49(Suppl 1):S261–S276. doi:10.2337/dc26-S012.

[3] Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. doi:10.1023/A:1010933404324.

[4] 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. New York: ACM; 2016. p. 785–794. doi:10.1145/2939672.2939785.

[5] Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. 2024;384:e074820. doi:10.1136/bmj-2023-074820.

[6] Huang H, Wu Y, Ye H, et al. Risk prediction models for diabetic retinopathy: a systematic review. Front Endocrinol (Lausanne). 2025;16:1556049. doi:10.3389/fendo.2025.1556049.

[7] Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med. 2021;385(3):283–286. doi:10.1056/NEJMc2104626.

[8] World Health Organization. Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation. Part 1: Diagnosis and classification of diabetes mellitus. Geneva: World Health Organization; 1999.

[9] Wilkinson CP, Ferris FL 3rd, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology. 2003;110(9):1677–1682. doi:10.1016/S0161-6420(03)00475-5.

[10] van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw. 2011;45(3):1–67. doi:10.18637/jss.v045.i03.

[11] Van Calster B, Collins GS, Vickers AJ, et al. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. Lancet Digit Health. 2025;7(12):100916. doi:10.1016/j.landig.2025.100916.

[12] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–845.

[13] Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med. 1998;17(8):857–872. doi:10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E.

[14] Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. doi:10.1186/s12916-019-1466-7.

[15] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–574. doi:10.1177/0272989X06295361.

[16] Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28(25):3083–3107. doi:10.1002/sim.3697.

[17] Usman TM, Saheed YK, Nsang A, et al. A systematic literature review of machine learning based risk prediction models for diabetic retinopathy progression. Artif Intell Med. 2023;143:102617. doi:10.1016/j.artmed.2023.102617.

[18] Wan X, Zhang R, Wang Y, et al. Predicting diabetic retinopathy based on routine laboratory tests by machine learning algorithms. Eur J Med Res. 2025;30:183. doi:10.1186/s40001-025-02442-5.

[19] He L, Zhang M, Wang X, et al. Machine learning-based prediction of diabetic retinopathy using clinlabomics: a multi-center study. BMC Med Inform Decis Mak. 2026;26:236. doi:10.1186/s12911-026-03524-y.

Table 1. Comparison of the 12 locked model input variables and DR prevalence between the model development cohort and temporal external validation cohort

Variable	Model development cohort (n = 3,874)	Temporal external validation cohort (n = 670)	|SMD|
UmAlb-24h（mg/24h）	17.95 [8.50, 58.98]	18.10 [8.70, 59.65]	0.004
apoB（g/L）	0.84 [0.69, 1.01]	0.84 [0.69, 1.01]	0.002
apoA-I（g/L）	1.04 [0.92, 1.20]	1.06 [0.89, 1.24]	0.018
DBIL（μmol/L）	2.60 [1.90, 3.40]	2.50 [1.90, 3.40]	0.019
BMI（kg/m^2）	24.80 [22.73, 27.06]	24.89 [22.80, 27.15]	0.009
Urea（mmol/L）	5.60 [4.60, 6.70]	5.60 [4.60, 6.70]	0.001
GA（%）	9.09 [6.39, 12.74]	9.05 [6.37, 12.48]	0.049
TP（g/L）	68.20 [64.45, 72.00]	68.25 [64.43, 71.88]	0.009
TyG-BMI	234.45 [207.18, 274.12]	222.49 [197.98, 250.83]	0.419
Hb（g/L）	143.00 [131.00, 154.00]	143.00 [131.00, 154.00]	0.002
GGT（U/L）	25.16 [18.10, 39.30]	25.10 [18.20, 39.55]	0.004
C-peptide（pmol/L）	585.53 [402.11, 807.59]	585.00 [403.00, 805.25]	0.000
DR，n（%）	1499（38.69）	265（39.55）	0.018

Continuous variables are presented as median (P25, P75), and DR is presented as n (%). SMD, standardized mean difference; absolute SMD values are reported. An $|SMD| \geq 0.10$ was considered indicative of a potentially relevant distributional difference. For continuous variables, SMDs were calculated using the means and pooled standard deviations of the two cohorts; for the categorical variable, the SMD was calculated using the difference in proportions and the pooled standard deviation.

Table 2. Classification performance of RF and XGBoost at the locked classification thresholds in the internal test set

Metric	RF (threshold = 0.380)	XGBoost (threshold = 0.396)
Confusion matrix (TP/FN/FP/TN)	346 / 104 / 116 / 597	336 / 114 / 102 / 611
Accuracy (95% CI)	0.811（0.787–0.832）	0.814（0.791–0.836）
Sensitivity (95% CI)	0.769（0.728–0.805）	0.747（0.705–0.785）
Specificity (95% CI)	0.837（0.808–0.863）	0.857（0.829–0.881）
PPV（95%CI）	0.749（0.707–0.786）	0.767（0.725–0.804）
NPV（95%CI）	0.852（0.823–0.876）	0.843（0.814–0.867）
F1 score	0.759	0.757
False-negative rate (95% CI)	0.231（0.195–0.272）	0.253（0.215–0.295）

TP, true positive; FN, false negative; FP, false positive; TN, true negative; PPV, positive predictive value; NPV, negative predictive value. The classification thresholds for RF (0.380) and XGBoost (0.396) were determined and locked during training. The 95% CIs for accuracy, sensitivity, specificity, PPV, NPV, and false-negative rate were calculated using the Wilson score method. The F1 score is reported as a point estimate.

Table 3. Classification performance of RF and XGBoost at the locked classification thresholds in the temporal external validation cohort

Metric	RF (threshold = 0.380)	XGBoost (threshold = 0.396)
Confusion matrix (TP/FN/FP/TN)	155 / 110 / 57 / 348	137 / 128 / 51 / 354
Accuracy (95% CI)	0.751（0.717–0.782）	0.733（0.698–0.765）
Sensitivity (95% CI)	0.585（0.525–0.643）	0.517（0.457–0.576）
Specificity (95% CI)	0.859（0.822–0.890）	0.874（0.838–0.903）
PPV（95%CI）	0.731（0.668–0.786）	0.729（0.661–0.787）
NPV（95%CI）	0.760（0.719–0.797）	0.734（0.693–0.772）
F1 score	0.650	0.605
False-negative rate (95% CI)	0.415（0.357–0.475）	0.483（0.424–0.543）

TP, true positive; FN, false negative; FP, false positive; TN, true negative; PPV, positive predictive value; NPV, negative predictive value. The classification thresholds for RF and XGBoost were 0.380 and 0.396, respectively. Both thresholds were determined from cross-validated out-of-fold predictions during training and were locked before validation; no threshold re-optimization was performed in the temporal external validation cohort. Except for the F1 score, values in parentheses are Wilson 95% CIs.

Table 4. Changes in probability-based model performance from internal testing to temporal external validation (descriptive analysis)

Metric	RF, internal test	RF, temporal external validation	XGBoost, internal test	XGBoost, temporal external validation	Δ (external − internal; RF/XGBoost)
ROC-AUC（95%CI）	0.880（0.859-0.901）	0.807（0.773-0.840）	0.877（0.856-0.899）	0.783（0.746-0.819）	-0.073 / -0.095
PR-AUC	0.859	0.781	0.865	0.766	-0.078 / -0.099
Brier score	0.1378	0.1757	0.1276	0.1840	+0.0379 / +0.0564

Δ was defined as the value in temporal external validation minus that in internal testing. The 95% CIs for ROC-AUC were estimated using the DeLong method. PR-AUC and Brier score are reported as point estimates. Between-cohort changes were compared descriptively only, without statistical inference.

552 Table 5. Calibration measures for RF and XGBoost in the temporal external validation cohort

Measure	RF	XGBoost
Observed DR prevalence	0.396（265/670）	0.396（265/670）
Mean predicted probability	0.333	0.298
Calibration intercept	0.733（0.468-0.998）	0.338（0.106-0.570）
Calibration slope	1.544（1.295-1.794）	0.668（0.551-0.785）
Calibration-in-the-large (CITL)	0.328（0.158-0.498）	0.756（0.554-0.957）

553 Observed DR prevalence represents the observed proportion of patients with DR, and mean predicted probability
554 represents the mean model-predicted probability. The calibration intercept and calibration slope were estimated
555 using a logistic recalibration model. CITL represents overall calibration with the calibration slope fixed at 1. Under
556 ideal calibration, the calibration intercept and CITL are 0 and the calibration slope is 1. Brier scores in the temporal
557 external validation cohort were 0.1757 for RF and 0.1840 for XGBoost; the corresponding values in the internal
558 test set were 0.1378 and 0.1276.

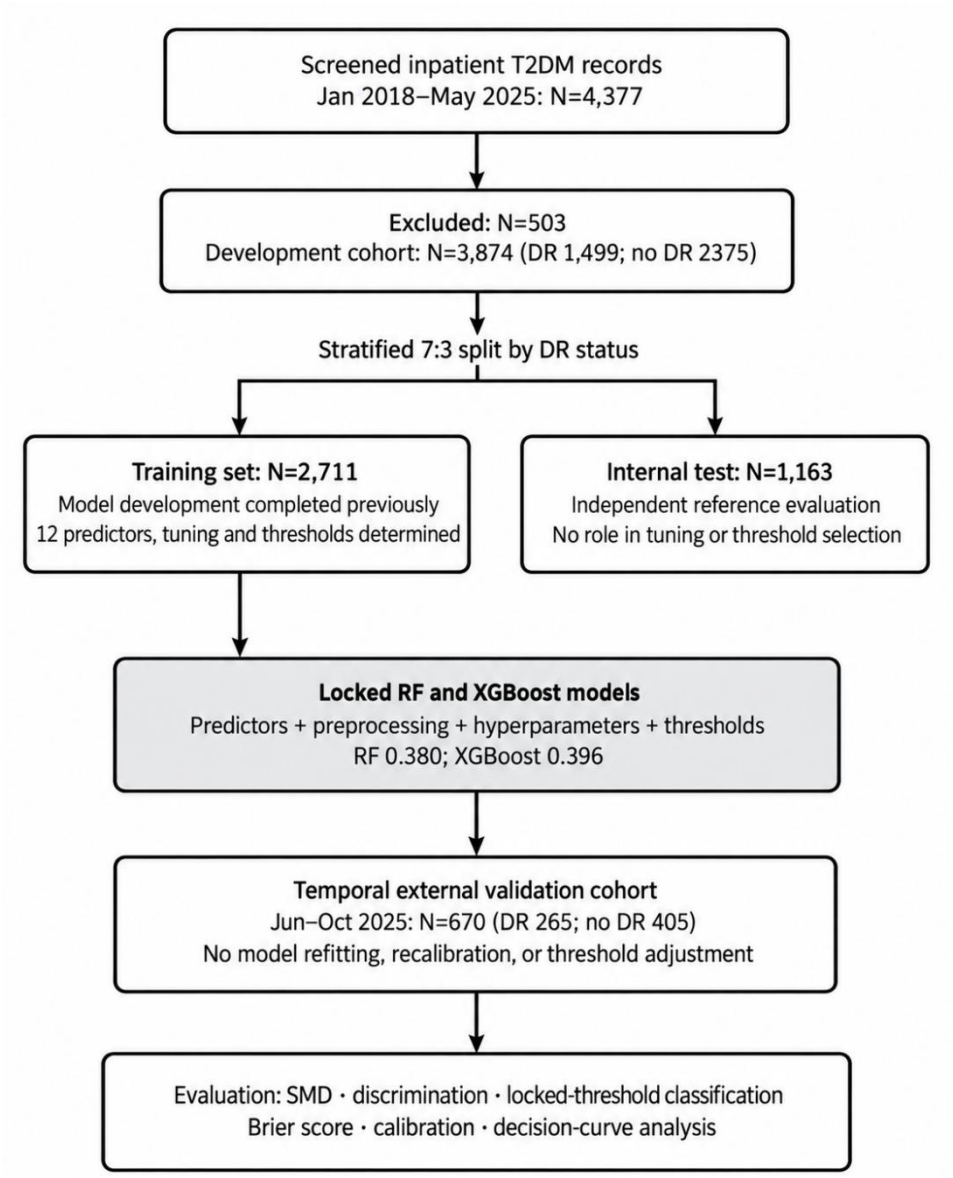


Figure 1. Participant selection, data splitting, and temporal external validation workflow.

The model development cohort comprised 3,874 patients, including 2,711 in the training set and 1,163 in the independent internal test set. The temporal external validation cohort included 670 patients, of whom 265 had DR and 405 did not.

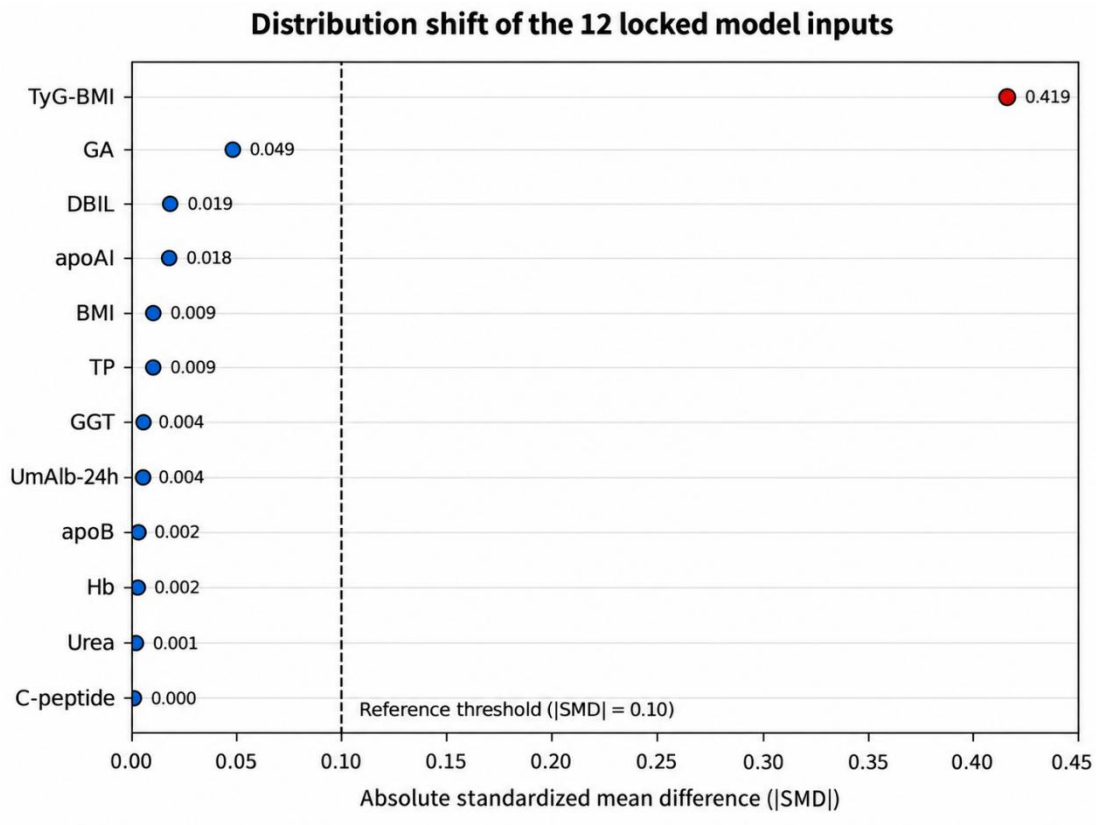


Figure 2. Absolute standardized mean differences for the 12 locked model input variables between the development cohort (n = 3,874) and temporal external validation cohort (n = 670).

The dashed vertical line indicates the reference threshold of |SMD| = 0.10. Only TyG-BMI exceeded this threshold.

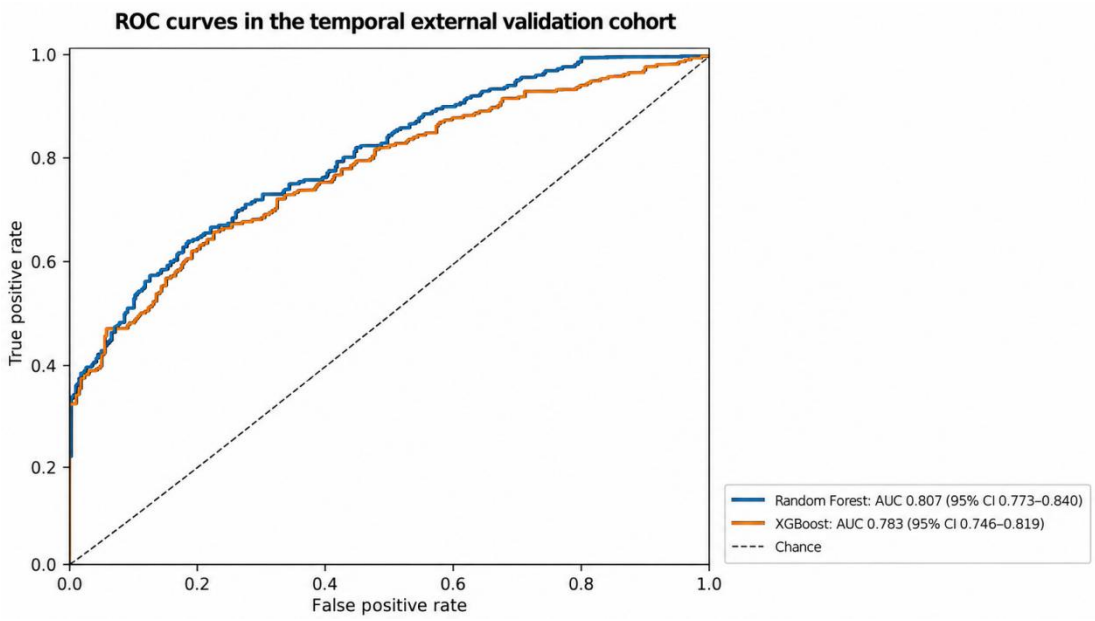
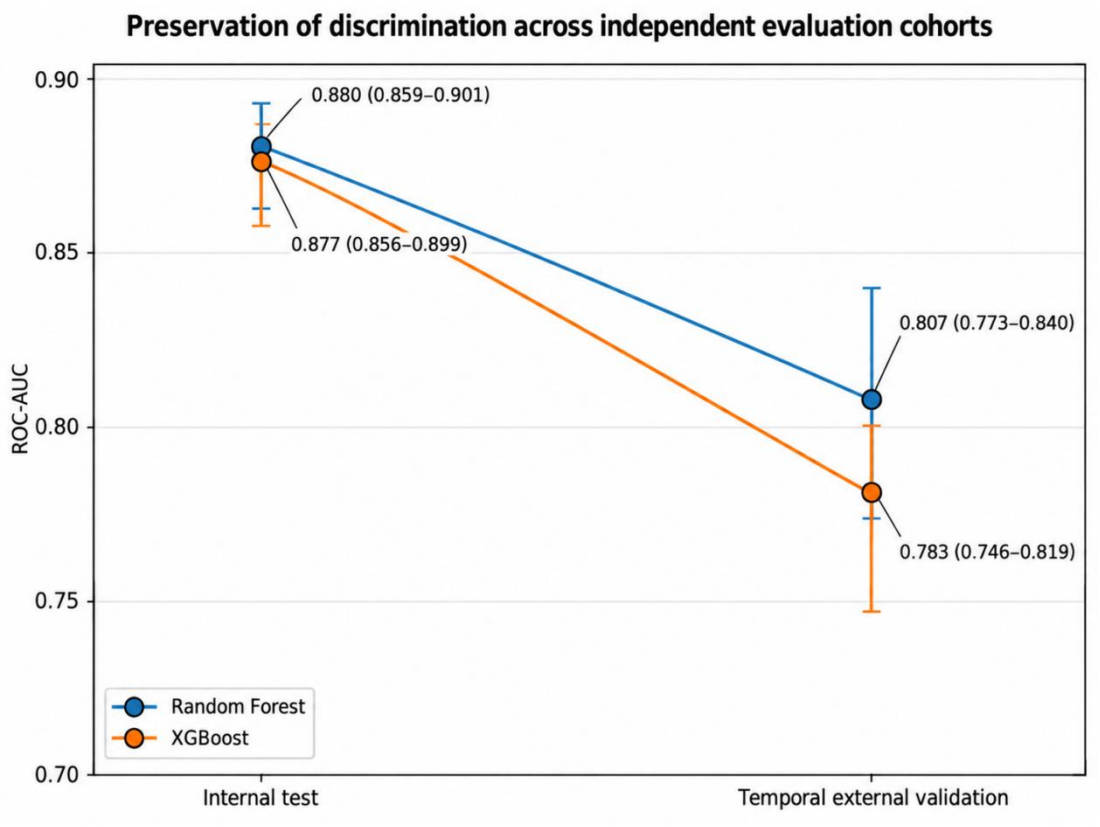


Figure 3. Receiver operating characteristic curves for RF and XGBoost in the temporal external validation cohort (n = 670; 265 patients with DR).

The ROC-AUC was 0.807 (95% CI, 0.773–0.840) for RF and 0.783 (95% CI, 0.746–0.819) for XGBoost. The dashed diagonal line represents chance-level discrimination. The paired DeLong method yielded an AUC difference of 0.024 (95% CI, 0.005–0.043; P = 0.015).



574

575 **Figure 4. Descriptive changes in ROC-AUC from internal testing (n = 1,163; 450 patients with DR) to**
576 **temporal external validation (n = 670; 265 patients with DR) for RF and XGBoost.**
577 Error bars represent DeLong 95% CIs. Between-cohort differences are shown descriptively only; no statistical
578 inference was performed.

Supplementary Material

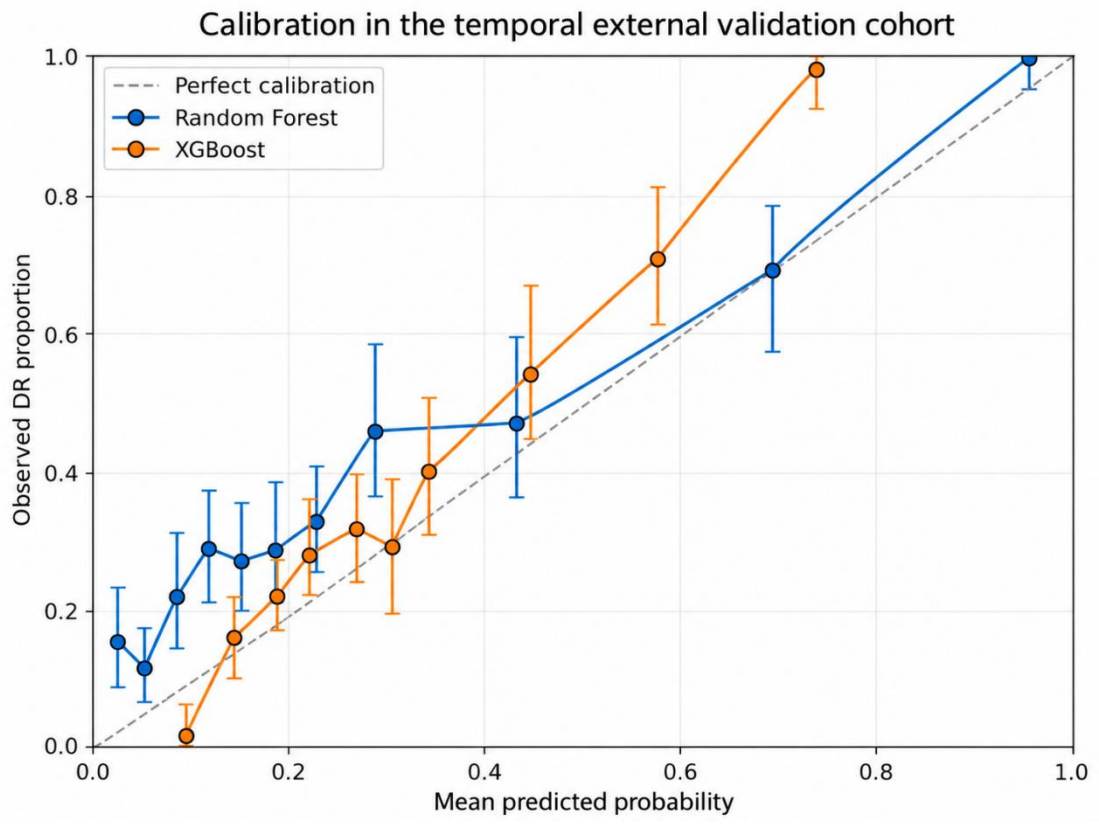
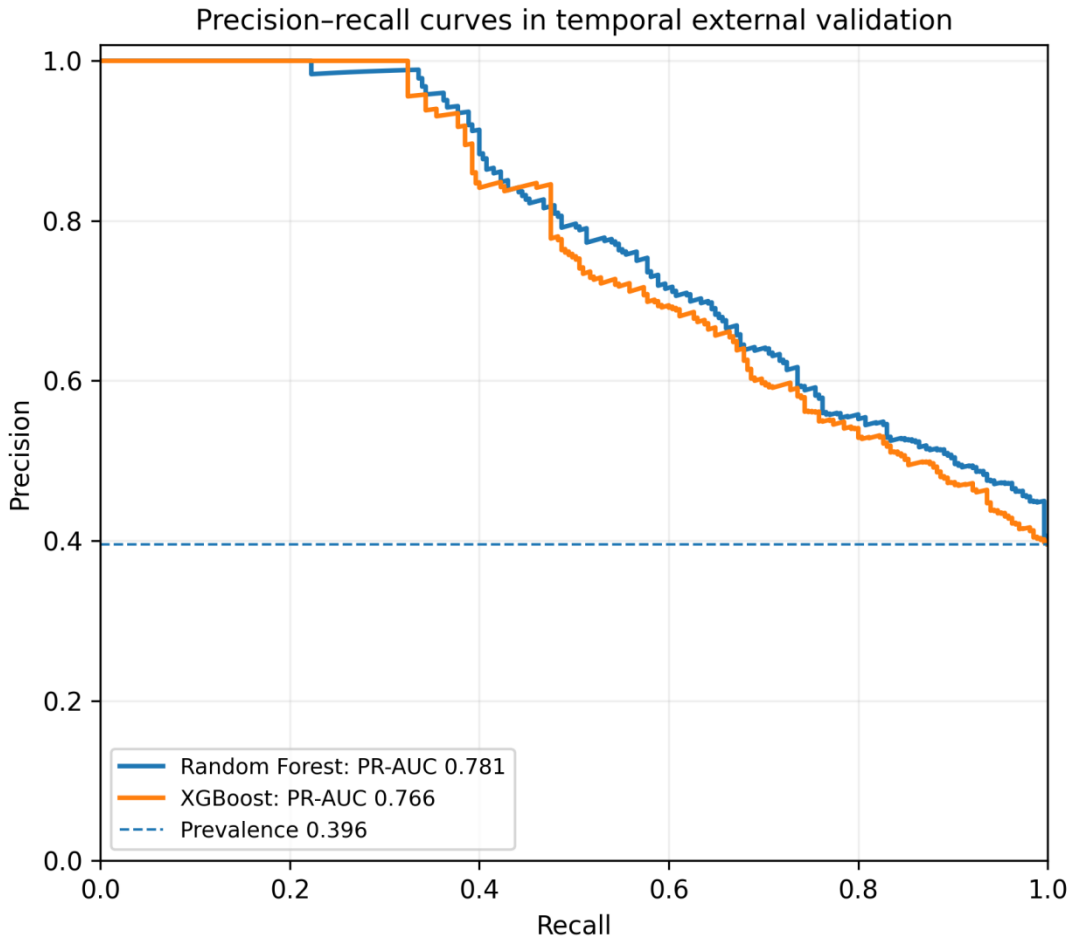


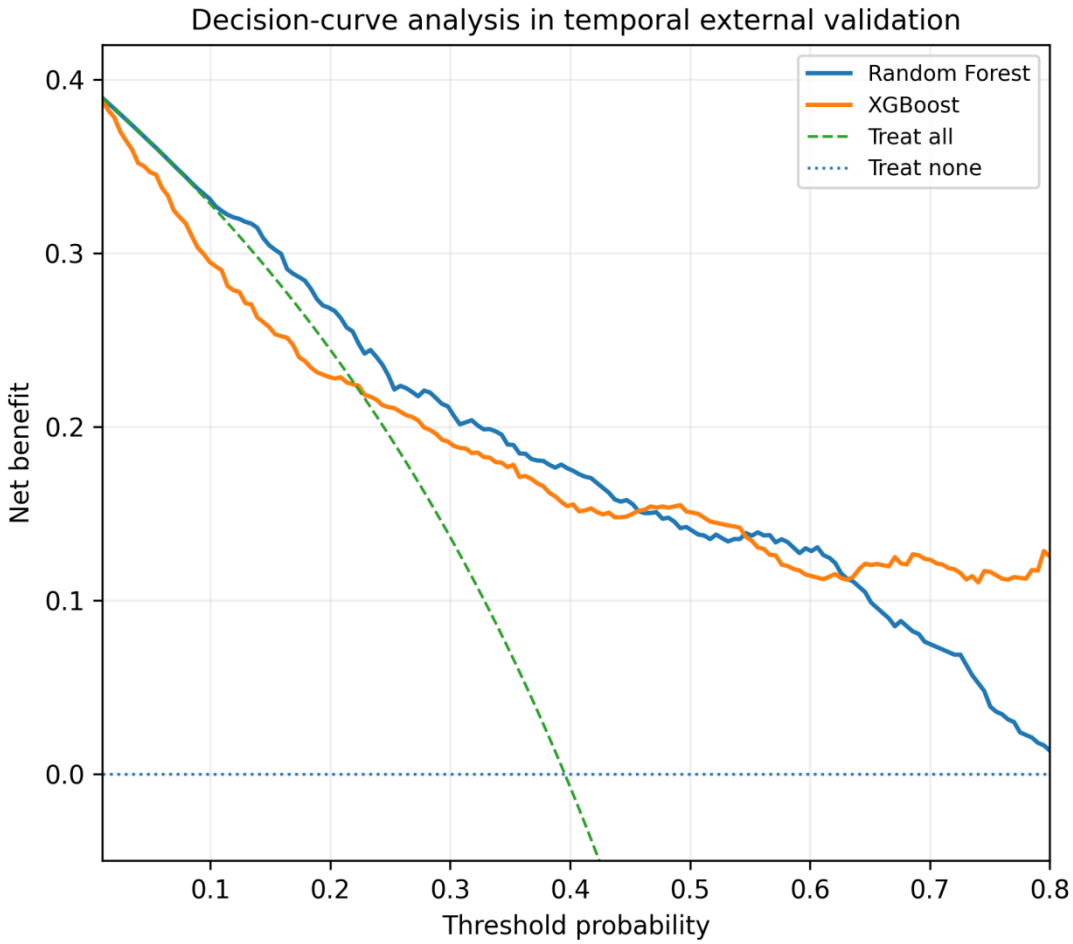
Figure S1. Calibration plots for RF and XGBoost in the temporal external validation cohort (n = 670; 265 patients with DR).

Patients were grouped into deciles according to predicted probability. Error bars represent Wilson 95% CIs for the observed proportions, and the diagonal line represents perfect calibration.



585

586 **Figure S2. Precision-recall curves for RF and XGBoost in the temporal external validation cohort (n = 670;**
587 **265 patients with DR). PR-AUC was 0.781 for RF and 0.766 for XGBoost.**
588 The horizontal reference line represents the DR prevalence in the cohort (39.55%).



589

590 **Figure S3. Decision curve analysis of RF and XGBoost in the temporal external validation cohort (n = 670;**
591 **265 patients with DR).**

592 The x-axis represents threshold probability and the y-axis represents net benefit. The treat-all and treat-none
593 strategies are shown as reference strategies.