

Explainable machine learning estimates clinically significant prostate cancer risk in men with PSA 4–20 ng/mL

Sheng Xie

Affiliated Hospital of North Sichuan Medical College

Qiang Wang

Affiliated Hospital of North Sichuan Medical College

Xiao Wang

Affiliated Hospital of North Sichuan Medical College

Xiaolu Jiang

Affiliated Hospital of North Sichuan Medical College

Huiyu Gao

Affiliated Hospital of North Sichuan Medical College

Yaxin Luo

Affiliated Hospital of North Sichuan Medical College

Le Fan

Affiliated Hospital of North Sichuan Medical College

Qingzhao Li

Affiliated Hospital of North Sichuan Medical College

Haozheng Wang

Affiliated Hospital of North Sichuan Medical College

Xiang Cai

Affiliated Hospital of North Sichuan Medical College

Wenhao Zhou

Affiliated Hospital of North Sichuan Medical College

Tao Wu

alhawking@163.com

Affiliated Hospital of North Sichuan Medical College

Article

Keywords: Clinically significant prostate cancer, Prostate-specific antigen, Machine learning, Risk prediction, Hematological markers, Bootstrap internal validation

Posted Date: October 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11018779/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Men with prostate-specific antigen (PSA) values of 4–20 ng/mL often have overlapping risks of benign disease, clinically insignificant prostate cancer, and clinically significant prostate cancer (csPCa). We retrospectively studied 859 men who underwent pathological evaluation; 76 had csPCa. Twelve prediction algorithms were assessed with 500 bootstrap resamples. In every resample, LASSO feature selection, hyperparameter tuning, and model fitting used only the bootstrap training sample, and performance was estimated from patient-level out-of-bag predictions. C5.0 had the highest AUROC point estimate (0.843, 95% CI 0.801–0.886), an AUPRC of 0.421, and a Brier score of 0.066. Calibration was broadly concordant at low-to-moderate predicted risks but uncertain in the sparse high-risk tail. At a 0.10 risk threshold, the model would avoid 544.8 biopsies per 1,000 men while missing 10.5 csPCa cases per 1,000. PSA-derived measures supplied most of the linear predictive signal, whereas the clearest additional gain arose with nonlinear integration. C5.0 may support pre-biopsy triage, but multicenter external validation, recalibration assessment, and prospective impact studies are required before clinical use.

Introduction

Prostate cancer (PCa) is one of the most common malignancies in men worldwide. As screening, imaging, and pathological assessment have changed, early detection has moved away from simply finding more PCa. The more useful goal is to identify clinically significant prostate cancer (csPCa), which is more likely to alter treatment or prognosis, while reducing unnecessary tests and biopsies for benign disease or low-risk tumors [1–7]. A pre-biopsy risk tool therefore needs more than good discrimination. Its probability estimates should be credible, and its clinical benefit should be clear in the population in which the tool will be used.

Serum PSA remains central to PCa risk assessment, but it is more organ-specific than cancer-specific. This limitation is especially evident between 4 and 20 ng/mL, where benign prostatic hyperplasia, inflammation, clinically insignificant PCa, and csPCa can have similar PSA values. Free PSA (fPSA), the free-to-total PSA ratio (f/t PSA), prostate volume, and PSA density (PSAD) add useful information, but a small set of linear markers may still miss important differences between patients. Multiparametric magnetic resonance imaging (mpMRI) and targeted biopsy have improved the diagnostic pathway, and major trials have shown that MRI-guided strategies can increase csPCa detection while sparing some biopsies [8–11]. Blood- and urine-based biomarkers may help as well, although cost, assay platforms, and local access vary [6, 12]. A low-cost model based on routine clinical data may therefore have a role before mpMRI, after MRI when further risk stratification is needed, or in settings where advanced testing is less accessible.

Routine blood tests also contain information on inflammation, immune-cell composition, and nutritional status. Measures such as the neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), monocyte-to-lymphocyte ratio (MLR), systemic immune-inflammation index (SII), systemic inflammation response index (SIRI), pan-immune-inflammation value (PIV), prognostic nutritional index (PNI), hemoglobin-albumin-lymphocyte-platelet (HALP) score, and advanced lung cancer inflammation index (ALI) can all be calculated from standard blood counts or biochemical tests. Some have been associated with PCa risk [13–16], yet their added value in the PSA gray zone has been inconsistent. They may also carry much the same information as PSA-derived measures, individual blood-cell counts, or nutritional markers. The practical question is not whether one composite index is significant on its own. It is whether these readily available features improve individual risk estimates when they are tested together without information leakage.

Machine-learning methods can capture nonlinear patterns and interactions, and they have been used with routine laboratory data, radiomics, and multimodal information for PCa prediction [16–22]. However, model performance can look too optimistic when feature selection, tuning, and evaluation are not separated within resampling. Current guidance for prediction models, including TRIPOD + AI, stresses a clear boundary between model development, internal validation, calibration, clinical utility, and interpretation [23–30]. We therefore compared 12 algorithms in men with PSA 4–20 ng/mL. In each bootstrap resample, LASSO selection, tuning, and fitting were repeated within the training sample, and only out-of-bag

(OOB) predictions contributed to internal-validation performance. We examined discrimination and calibration, compared successive levels of clinical and PSA-related information, quantified potential biopsy trade-offs with decision curve analysis (DCA), and used SHAP to explain one representative final model.

Results

Study population and baseline characteristics

The final cohort contained 859 men with PSA 4–20 ng/mL. Of these, 765 (89.1%) had no PCa, 76 (8.8%) had csPCa, and 18 (2.1%) had clinically insignificant PCa. Among the 94 men with any PCa, 80.9% had csPCa (Fig. 1b).

Total PSA, f/t PSA, and PSAD differed across the three pathological groups (all $P < 0.001$). Mean tPSA was 10.67 ± 3.30 ng/mL in men with csPCa and 8.64 ± 3.83 ng/mL in men without PCa. Mean f/t PSA was lower in csPCa (0.12 ± 0.03 vs 0.20 ± 0.12), whereas mean PSAD was higher (0.24 ± 0.16 vs 0.16 ± 0.08). The between-group difference in fPSA was close to, but did not reach, statistical significance ($P = 0.055$). Age, body mass index (BMI), prostate volume, blood-cell counts, biochemical measures, composite inflammatory-nutritional indices, and the recorded clinical histories did not show significant overall differences (Table 2).

Feature-selection stability and the representative final model

Because LASSO selection was repeated inside each bootstrap sample, the feature set used for internal validation was not fixed. The selection-frequency analysis showed that f/t PSA and PSAD were chosen in almost every resample. HALP, tPSA, and hemoglobin were also selected often. Creatinine, blood urea nitrogen, and monocyte count had intermediate selection frequencies, while PNI, MLR, SIRI, NLR, and PIV were less stable (Fig. 2a). Thus, PSA-related variables supplied the most consistent signal, although routine hematological or inflammatory-nutritional measures added information in some resamples.

For the separate full-data interpretation branch, LASSO retained five variables: f/t PSA, tPSA, PSAD, monocyte count, and hemoglobin. The full-data cross-validation curve and coefficient paths are shown in Fig. 2b,c. These analyses describe the representative final model only; they did not contribute to the internal-validation performance estimates.

Internal-validation performance of the 12 models

Several tree-based and ensemble models showed relatively strong discrimination on patient-level OOB predictions (Fig. 3a). C5.0 had the highest AUROC point estimate at 0.843, followed by AdaBoost (0.820), gradient boosting machine (GBM; 0.818), and XGBoost (0.815). AUROCs were 0.807 for both naive Bayes and classification and regression tree (CART), 0.802 for random forest, 0.796 for the neural network, 0.782 for the support vector machine, 0.767 for k-nearest neighbors, 0.765 for the multilayer perceptron, and 0.733 for logistic regression. The 95% confidence intervals (CIs) for C5.0, AdaBoost, GBM, and XGBoost overlapped appreciably (Fig. 3b). These data therefore support C5.0 as the model with the highest point estimate and a favorable balance across metrics, but they do not establish statistical superiority over every competing algorithm.

For C5.0, the AUROC was 0.843 (95% CI 0.801–0.886), the area under the precision-recall curve (AUPRC) was 0.421 (0.325–0.536), and the Brier score was 0.066 (0.061–0.070). At the Youden cutoff, sensitivity was 0.750 (0.658–0.842), specificity 0.765 (0.736–0.794), positive predictive value (PPV) 0.237 (0.205–0.271), negative predictive value (NPV) 0.969 (0.958–0.981), Matthews correlation coefficient (MCC) 0.326 (0.261–0.392), and F1 score 0.360 (0.314–0.408) (Table 3). Because csPCa prevalence was only 8.8%, an AUPRC of 0.421 was about 4.8 times the event rate and indicates substantial enrichment of csPCa among higher-ranked predictions.

Threshold-dependent behavior differed between models (Table 3). CART had the highest sensitivity, 0.908 (0.829–0.961), the highest NPV, 0.985 (0.973–0.994), and the lowest Brier score, 0.065 (0.059–0.071), but its specificity was only 0.591 and its PPV was 0.177. Random forest showed a similar sensitivity-oriented pattern. The support vector machine had the highest specificity, 0.801 (0.771–0.828), and PPV, 0.239 (0.201–0.281), but its sensitivity was 0.645. C5.0 combined the highest AUROC, AUPRC, MCC, and F1 point estimates with a low Brier score and a more even balance of sensitivity and specificity. We therefore carried C5.0 forward for calibration, clinical-utility, progressive-model, and interpretation analyses.

Calibration and clinical decision performance

The smoothed OOB calibration curve for C5.0 stayed close to the ideal line through the densely populated low-to-moderate risk range. At higher predicted probabilities, the smooth rose above the ideal line and the confidence band widened (Fig. 3c). Few patients were present in this high-risk tail, so the divergence is better read as greater calibration uncertainty than as evidence of uniform good calibration or systematic miscalibration across the whole range. The corresponding OOB calibration curves for the three leading models are shown in Fig. 3d.

DCA showed positive net benefit for C5.0 and GBM through most threshold probabilities from 0.01 to 0.30. C5.0 was among the better-performing models at most thresholds, generally exceeded the treat-none strategy, and was above treat-all across much of the clinically relevant range. AdaBoost approached or slightly crossed below zero net benefit around some thresholds between approximately 0.22 and 0.27 (Fig. 4a).

The threshold-consequence analysis translated these curves into clinical counts (Fig. 4b). At a risk threshold of 0.05, 406.3 biopsies per 1,000 men could be avoided (349/859 in this cohort), while 3.5 csPCa cases per 1,000 would be missed (3/859). At 0.10, 544.8 biopsies per 1,000 could be avoided (468/859), with 10.5 missed csPCa cases per 1,000 (9/859). Raising the threshold to 0.20 increased avoided biopsies to 814.9 per 1,000 (700/859), but missed csPCa also rose to 34.9 per 1,000 (30/859). At 0.30, 938.3 biopsies per 1,000 could be avoided (806/859), while 58.2 csPCa cases per 1,000 would be missed (50/859). There was therefore no single threshold that could be chosen merely by maximizing the number of avoided biopsies.

Progressive models and SHAP interpretation

In the progressive-model comparison, AUROCs were 0.350 for basic clinical logistic regression, 0.742 after PSA-derived variables and PSAD were added, 0.733 for nested-LASSO logistic regression, and 0.843 for integrated C5.0 (Fig. 5a). The first large gain came from PSA/PSAD information. Data-driven LASSO selection within a linear logistic model did not improve AUROC beyond the prespecified PSA/PSAD model. The next clear gain appeared with nonlinear C5.0. DCA showed the same broad pattern: the basic clinical model quickly lost net benefit, the two linear models retained some benefit at low-to-moderate thresholds, and C5.0 generally remained higher and more stable from 0.01 to 0.30 (Fig. 5b).

Continuous net reclassification improvement (NRI) and integrated discrimination improvement (IDI) supported this interpretation (Fig. 5c,d). Adding PSA/PSAD to the basic clinical model produced positive NRI and IDI with 95% CIs above zero. Nested-LASSO logistic regression had a positive NRI relative to the PSA/PSAD model, but the IDI CI crossed zero. By contrast, C5.0 showed positive improvements in both NRI and IDI relative to the PSA/PSAD and nested-LASSO logistic models, with CIs generally not crossing zero. The data therefore suggest that PSA/PSAD supplied the main linear predictive information, while much of the later gain came from nonlinear integration of multiple features.

SHAP analysis of the representative full-data C5.0 model included only its five retained variables (Fig. 5e,f). By mean absolute SHAP value, f/t PSA contributed most, followed by tPSA, PSAD, monocyte count, and hemoglobin. Lower f/t PSA generally shifted predictions toward higher risk, and higher tPSA and PSAD also tended to increase predicted risk. Monocyte count and hemoglobin had smaller effects and more between-patient variation. HALP, PNI, NLR, PLR, SII, SIRI, and PIV did not

enter the representative final C5.0 model and should not be described as its main explanatory features. HALP was nevertheless selected in many bootstrap resamples, which suggests partial overlap with PSA-related or individual hematological variables rather than a complete absence of predictive information.

Discussion

This study developed and internally validated a pre-biopsy csPCa risk model for men with PSA 4–20 ng/mL. The model comparison revealed clear trade-offs. CART favored sensitivity, whereas the support vector machine favored specificity and positive confirmation. C5.0 had the highest AUROC, AUPRC, MCC, and F1 point estimates and kept a low Brier score with a more balanced sensitivity-specificity profile. PSA/PSAD variables were the dominant and most stable source of linear predictive information. Nested-LASSO logistic regression did not improve AUROC beyond the prespecified PSA/PSAD model, while a clearer second gain appeared after nonlinear C5.0 integration. The clinical-consequence analysis also showed why a single performance number is not enough: at a 0.10 threshold, about 544.8 biopsies per 1,000 men could be avoided with 10.5 csPCa cases missed, whereas at 0.20 the number of avoided biopsies rose to 814.9 but missed csPCa increased to 34.9 per 1,000.

The PSA gray zone has long complicated biopsy decisions. Current guidelines and major diagnostic studies place greater emphasis on finding csPCa than on maximizing overall PCa detection [3–11]. In our cohort, men with csPCa had higher tPSA and PSAD and lower f/t PSA. These same variables dominated bootstrap selection and the final SHAP analysis. The pattern is consistent with the established clinical value of PSA-derived measures in this setting and helps explain why AUROC increased from 0.350 to 0.742 when PSA/PSAD information was added to the basic clinical model.

The inflammatory-nutritional findings need a more cautious reading. HALP was selected fairly often across bootstrap samples, so it may contribute supplementary information in some resamples. Yet HALP, PNI, and the other composite indices were absent from the final five-variable C5.0 model. Routine hematological variables and composite scores probably carry partly redundant information, and different resamples may represent that information with different features. Some of it may also overlap with PSA-derived measures, monocyte count, or hemoglobin. SHAP shows how the representative final model used its inputs; it does not identify independent risk factors or causal mechanisms [31].

The progressive comparison also argues against a simple 'more variables are better' explanation. Nested-LASSO logistic regression had a slightly lower AUROC than the prespecified PSA/PSAD model (0.733 vs 0.742), and its IDI relative to the PSA/PSAD model did not show stable positive evidence. C5.0, in contrast, improved more consistently across AUROC, NRI, IDI, and DCA. Nonlinear relations, threshold effects, or interactions may therefore matter more than the linear addition of extra variables in this dataset. Other machine-learning approaches have shown promise with routine laboratory data, radiomics, and multimodal PCa information [16–22]. Still, the AUROC CIs for C5.0, AdaBoost, GBM, and XGBoost overlapped substantially. Our findings support nonlinear integration as a useful strategy in this cohort; they do not show that C5.0 will be superior in other populations.

Class imbalance is another reason to look beyond AUROC. Only 8.8% of the cohort had csPCa, whereas C5.0 reached an AUPRC of 0.421, about 4.8 times the event rate. NPV was 0.969, but PPV remained only 0.237. In a low-prevalence setting, this profile is more consistent with a triage or rule-out-oriented adjunct than with a stand-alone confirmatory test. The differences in sensitivity, specificity, and PPV across algorithms reinforce the same point: a clinically useful model should be judged across discrimination, precision-recall performance, probabilistic error, calibration, and net benefit rather than AUROC alone.

Calibration deserves particular caution. Predicted and observed risks were broadly similar where most patients lay, but the data became sparse at the highest predicted risks and the confidence band widened. The current cohort therefore does not support a claim of good calibration across the full risk range [29]. External assessment, and possibly recalibration, will be especially important at higher predicted risks. DCA was generally favorable for C5.0 between thresholds of 0.01 and 0.30 [30], but the threshold-consequence results show the clinical price of moving that threshold. Lower thresholds better suit a

sensitivity-oriented pathway that prioritizes csPCa detection. Higher thresholds spare more biopsies, but they require acceptance of more missed disease. A clinical pathway should therefore specify what happens below the threshold, such as repeat assessment or mpMRI triage, rather than treating a low predicted risk as a reason to stop evaluation.

A strength of this study is that feature selection remained inside every bootstrap iteration. LASSO selection, tuning, and model training used only the corresponding training sample, while performance came only from OOB patients and was then aggregated at the patient level. This design reduces the optimism that can arise when the full dataset influences feature selection before validation. We also kept the full-data LASSO, final C5.0 model, and SHAP analysis in a separate interpretation branch. That separation follows the reporting principles emphasized by TRIPOD + AI and related prediction-model guidance [23–28].

Several limitations remain. This was a single-center retrospective study with only 76 csPCa events, so transportability across regions, pathological spectra, and care pathways is uncertain. Repeated bootstrap validation improves the rigor of internal assessment, but it cannot replace external validation. We used complete-case analysis and excluded 27 men with missing candidate predictors or outcome data, which may have introduced selection bias. Outcome verification could come from biopsy or postoperative pathology, so the verification pathway was not identical for every patient. The available predictors did not include digital rectal examination, family history, previous negative biopsy, PI-RADS, quantitative MRI features, or some contemporary molecular biomarkers. Calibration was uncertain in the sparsely populated high-risk tail, and NRI and IDI estimates also warrant caution because the number of events was limited. Finally, we did not prespecify paired superiority tests between C5.0 and the other leading algorithms, so point estimates and CI overlap should not be used as evidence of statistically significant pairwise differences. SHAP, likewise, describes one representative full-data model rather than a common mechanism shared by all 500 bootstrap-specific models.

These findings should be viewed as internal evidence from the model-development stage. The next step is temporal and multicenter geographic validation, with explicit assessment of calibration-in-the-large, calibration slope, and any need for recalibration. Clinically relevant thresholds should then be reassessed in the target population, alongside the associated numbers of avoided biopsies and missed csPCa cases. Comparisons with mpMRI, PI-RADS, and contemporary biomarkers will also be important. Until such work is available, C5.0 is best positioned as an adjunct for pre-biopsy risk stratification rather than an independent diagnostic test.

Methods

Study design and participants

This single-center retrospective study developed and internally validated a diagnostic prediction model. Data came from the electronic medical record, laboratory information system, imaging records, and pathology reports of the Affiliated Hospital of North Sichuan Medical College. The index time was the most recent complete clinical assessment before pathological diagnosis, and all candidate predictors were available before the outcome was ascertained. Reporting followed the TRIPOD+AI framework [23].

We consecutively screened men who underwent pathological evaluation from January 2018 through December 2025, had a total PSA (tPSA) concentration of 4–20 ng/mL before pathological assessment, and had a definitive pathological result. We excluded men with previous PCa or related treatment, a recent urological procedure likely to materially alter PSA, acute infection or urinary retention, use of a 5 α -reductase inhibitor that could not be adequately accounted for, duplicate records, or missing data for the outcome or any candidate predictor. Only the first eligible record for each patient was retained. Of 886 screened men, 27 had incomplete data and were excluded, which left 859 complete cases (Fig. 1a).

Outcome and candidate predictors

The primary outcome was pathologically confirmed csPCa, defined as International Society of Urological Pathology (ISUP) grade group ≥ 2 or Gleason score $\geq 3+4$. The non-csPCa category included benign prostate pathology and ISUP grade group 1 PCa. If both biopsy and postoperative pathology were available, the higher-grade result defined the outcome.

Candidate predictors covered demographics and clinical history, tPSA, fPSA, f/t PSA, prostate volume, PSAD, complete blood count variables, biochemical measures, and the composite inflammatory-nutritional indices NLR, PLR, MLR, SII, SIRI, PIV, PNI, HALP, and ALI. The f/t PSA ratio was fPSA divided by tPSA. Prostate volume was calculated as length $\times$ width $\times$ height $\times \pi/6$, and PSAD as tPSA divided by prostate volume. Table 1 gives the formulas for the composite indices.

Model development and nested-bootstrap internal validation

All models used the complete-case cohort; we did not perform multiple imputation. Internal validation used 500 bootstrap resamples. For each resample, 859 observations were drawn with replacement from the original cohort to form the training sample, and patients who were not sampled formed the corresponding OOB validation set. LASSO feature selection was repeated from scratch within every bootstrap training sample. Model-specific preprocessing, hyperparameter tuning, and model fitting then used that training sample alone, and predicted probabilities were generated only for the corresponding OOB patients.

We compared 12 candidate algorithms: logistic regression, k-nearest neighbors, support vector machine, neural network, multilayer perceptron, random forest, naive Bayes, CART, GBM, AdaBoost, XGBoost, and C5.0. A patient could appear in the OOB set for many bootstrap iterations. Internal-validation performance was therefore estimated from the mean of all OOB probabilities available for each patient, not from apparent predictions from a model refitted to the full dataset.

After internal validation was complete, LASSO and the model with the highest internally validated AUROC point estimate were fitted once to the full dataset. This separate branch supplied a reproducible representative model for interpretation and SHAP analysis (Fig. 1a). The full-data LASSO, final C5.0 fit, and SHAP values did not contribute to the internal-validation performance estimates.

Performance, calibration, and clinical utility

Discrimination was assessed with the area under the receiver operating characteristic curve (AUROC) and AUPRC, while overall probabilistic error was assessed with the Brier score. For patient-level OOB predictions, each model received its own cutoff from the Youden index. We then calculated sensitivity, specificity, PPV, NPV, MCC, F1 score, accuracy, and the confusion matrix. Ninety-five percent CIs for AUROC, AUPRC, Brier score, and threshold-dependent measures came from 2,000 patient-level bootstrap resamples stratified by outcome.

Calibration of C5.0 was described from patient-level OOB predictions. A binomial generalized additive model provided a smoothed calibration curve and a 95% confidence band. We also displayed mean predicted probability, observed event rate, and Wilson 95% CI for five equally sized risk groups. The smooth was restricted to the observed range of predicted probabilities and was not extrapolated into unsupported high-risk regions.

Clinical net benefit was assessed with DCA. The primary curve covered threshold probabilities from 0.01 to 0.30 and included treat-all and treat-none reference strategies. For the leading model, we also calculated biopsies avoided and csPCa cases missed per 1,000 men across thresholds from 0.01 to 0.30 in 0.005 increments. Thresholds of 0.05, 0.10, 0.20, and 0.30 were highlighted so that the trade-off between fewer biopsies and missed csPCa could be read directly.

Progressive models and incremental predictive value

Four models represented successive levels of information. M1 was a basic clinical logistic regression model with age, BMI, hypertension, diabetes, smoking history, and alcohol-use history. M2 added tPSA, fPSA, f/t PSA, prostate volume, and PSAD. M3 was the nested-LASSO logistic regression model from the 12-model comparison: within every bootstrap training sample, LASSO selected features from the full candidate set before logistic regression was fitted. M4 was C5.0, which had the highest internally validated AUROC point estimate. M1 and M2 generated OOB predictions under the same outer bootstrap sampling procedure as the main analysis; M3 and M4 reused their corresponding patient-level OOB predictions.

M3 is not a strict blockwise model in which inflammatory-nutritional variables were simply added to M2. Because each bootstrap-specific LASSO fit could add or remove different predictors, M3 represents data-driven multivariable logistic regression after LASSO selection. It cannot estimate a fixed-block independent effect for inflammatory-nutritional indices or support causal claims. The four models were compared with AUROC, DCA, continuous NRI, and IDI. Ninety-five percent CIs for NRI and IDI were obtained from 2,000 patient-level bootstrap resamples stratified by outcome.

Model interpretation and statistical analysis

The representative full-data C5.0 model was interpreted with SHAP. Mean absolute SHAP values and a beeswarm plot described the relative contribution and direction of each variable in the model output. Because this analysis refers to one representative final model, whereas internal validation combined different feature sets and fitted models across 500 bootstrap iterations, SHAP values should not be read as a single mechanism shared by all internally validated models or as evidence of causality [31].

Continuous variables were summarized as mean $\pm$ standard deviation or median (interquartile range), as appropriate, and categorical variables as counts and percentages. Between-group comparisons used one-way analysis of variance, the Kruskal–Wallis test, Pearson's χ^2 test, or Fisher's exact test according to variable type and distribution. All tests were two-sided, and $P < 0.05$ was considered statistically significant. Analyses were performed in R version 4.4.2 (R Foundation for Statistical Computing, Vienna, Austria).

Declarations

Ethics approval and informed consent

The Ethics Committee of the Affiliated Hospital of North Sichuan Medical College approved the study (approval no. 2026EF285-1). The study was conducted in accordance with the Declaration of Helsinki and applicable institutional and national ethical requirements. Because this retrospective study used data that were anonymized before analysis, the Ethics Committee waived the requirement for informed consent.

Generative AI-assisted language editing

During manuscript preparation, the authors used OpenAI ChatGPT (GPT-5.6 Sol) for Chinese-to-English translation, language editing, and adaptation to journal format. The tool was not used to generate or alter study data, perform statistical analyses, train or evaluate prediction models, select reported results, or make clinical decisions. The authors reviewed and revised the generated text and take responsibility for the final manuscript.

Data availability

The datasets analyzed during the current study contain patient-level clinical information and are not publicly available because of institutional ethics and privacy restrictions. Data access, if permitted, is subject to the applicable institutional and ethics requirements and should be discussed with the corresponding author.

Funding

This work was supported by the Health Commission of Sichuan Province Medical Science and Technology Program (24LCYJPT15), Medical Research Project of Sichuan Medical Association (S2024066), Sichuan Medical Science and Technology Innovation Research Association Project (2025YCZD007), and Sichuan Provincial Science and Technology-Education Joint Fund of the Sichuan Science and Technology Department (2026NSFSC2081).

Author contributions

SX conceived and designed the study. SX and QW analyzed and interpreted the data. XW, HW, XJ, HG, YL, LF, and QL curated the data. SX and QW drafted the manuscript. TW and SX critically revised the manuscript for important academic content. TW, XC, and WZ supervised the study. TW acquired funding. All authors reviewed and approved the final manuscript and agree to be accountable for the integrity of the work.

Additional information

Competing interests

The authors declare no competing interests.

Correspondence and requests for materials

Correspondence and requests for materials should be addressed to Tao Wu (alhawking@163.com).

References

1. Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263 (2024). <https://doi.org/10.3322/caac.21834>
2. Sung, H. et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021). <https://doi.org/10.3322/caac.21660>
3. Grossman, D. C. et al. Screening for prostate cancer: US Preventive Services Task Force recommendation statement. *JAMA* **319**, 1901–1913 (2018). <https://doi.org/10.1001/jama.2018.3710>
4. Wei, J. T. et al. Early detection of prostate cancer: AUA/SUO guideline part I: Prostate cancer screening. *J. Urol.* **210**, 46–53 (2023). <https://doi.org/10.1097/JU.0000000000003491>
5. Wei, J. T. et al. Early detection of prostate cancer: AUA/SUO guideline part II: Considerations for a prostate biopsy. *J. Urol.* **210**, 54–63 (2023). <https://doi.org/10.1097/JU.0000000000003492>
6. Moses, K. A. et al. Prostate cancer early detection, version 1.2023, NCCN clinical practice guidelines in oncology. *J. Natl. Compr. Canc. Netw.* **21**, 236–246 (2023). <https://doi.org/10.6004/jnccn.2023.0014>
7. Cornford, P. et al. EAU-EANM-ESTRO-ESUR-ISUP-SIOG guidelines on prostate cancer—2024 update. Part I: Screening, diagnosis, and local treatment with curative intent. *Eur. Urol.* **86**, 148–163 (2024). <https://doi.org/10.1016/j.eururo.2024.03.027>
8. Kasivisvanathan, V. et al. MRI-targeted or standard biopsy for prostate-cancer diagnosis. *N. Engl. J. Med.* **378**, 1767–1777 (2018). <https://doi.org/10.1056/NEJMoa1801993>
9. Rouvière, O. et al. Use of prostate systematic and targeted biopsy on the basis of multiparametric MRI in biopsy-naïve patients (MRI-FIRST): A prospective, multicentre, paired diagnostic study. *Lancet Oncol.* **20**, 100–109 (2019). [https://doi.org/10.1016/S1470-2045\(18\)30569-2](https://doi.org/10.1016/S1470-2045(18)30569-2)
10. van der Leest, M. et al. Head-to-head comparison of transrectal ultrasound-guided prostate biopsy versus multiparametric prostate resonance imaging with subsequent magnetic resonance-guided biopsy in biopsy-naïve men with elevated prostate-specific antigen: A large prospective multicenter clinical study. *Eur. Urol.* **75**, 570–578 (2019). <https://doi.org/10.1016/j.eururo.2018.11.023>

11. Ahdoot, M. et al. MRI-targeted, systematic, and combined biopsy for prostate cancer diagnosis. *N. Engl. J. Med.* **382**, 917–928 (2020). <https://doi.org/10.1056/NEJMoa1910038>
12. Robinson, H. S., Lee, S. S., Barocas, D. A. & Tosoian, J. J. Evaluation of blood and urine based biomarkers for detection of clinically-significant prostate cancer. *Prostate Cancer Prostatic Dis.* **28**, 45–55 (2025). <https://doi.org/10.1038/s41391-024-00840-0>
13. Zhu, M. et al. Diagnostic efficiency of pan-immune-inflammation value to predict prostate cancer in patients with prostate-specific antigen between 4 and 20 ng/mL. *J. Clin. Med.* **12**, 820 (2023). <https://doi.org/10.3390/jcm12030820>
14. Kaya, C. et al. HALP score and albumin levels in men with prostate cancer and benign prostate hyperplasia. *Int. J. Clin. Pract.* **75**, e13766 (2021). <https://doi.org/10.1111/ijcp.13766>
15. Sun, J., Song, Y., Tan, Y., Ma, L. & Xie, H. Predictive value of inflammatory marker PLR in prostate cancer. *BMC Urol.* **25**, 188 (2025). <https://doi.org/10.1186/s12894-025-01882-9>
16. Satır, A., Üstündağ, Y., Yeşil, M. R. & Huysal, K. Prediction of prostate cancer from routine laboratory markers with automated machine learning. *J. Clin. Lab. Anal.* **39**, e25143 (2025). <https://doi.org/10.1002/jcla.25143>
17. Deng, X. et al. A machine learning model for the prediction of prostate cancer in patients with low prostate-specific antigen levels: A multicenter retrospective study. *Front. Oncol.* **12**, 985940 (2022). <https://doi.org/10.3389/fonc.2022.985940>
18. Zhong, J. G. et al. Predicting prostate cancer in men with PSA levels of 4-10 ng/mL: MRI-based radiomics can help junior radiologists improve the diagnostic performance. *Sci. Rep.* **13**, 4846 (2023). <https://doi.org/10.1038/s41598-023-31869-1>
19. Hyndman, M. E. et al. Development of an effective predictive screening tool for prostate cancer using the ClarityDX machine learning platform. *npj Digit. Med.* **7**, 163 (2024). <https://doi.org/10.1038/s41746-024-01167-9>
20. Cai, J. C. et al. Fully automated deep learning model to detect clinically significant prostate cancer at MRI. *Radiology* **312**, e232635 (2024). <https://doi.org/10.1148/radiol.232635>
21. Saha, A. et al. Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): An international, paired, non-inferiority, confirmatory study. *Lancet Oncol.* **25**, 879–887 (2024). [https://doi.org/10.1016/S1470-2045\(24\)00220-1](https://doi.org/10.1016/S1470-2045(24)00220-1)
22. Zhao, W. et al. Interpretable machine learning model for predicting clinically significant prostate cancer: Integrating intratumoral and peritumoral radiomics with clinical and metabolic features. *BMC Med. Imaging* **24**, 353 (2024). <https://doi.org/10.1186/s12880-024-01548-2>
23. Collins, G. S. et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024). <https://doi.org/10.1136/bmj-2023-078378>
24. Moons, K. G. M. et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **388**, e082505 (2025). <https://doi.org/10.1136/bmj-2024-082505>
25. Efthimiou, O. et al. Developing clinical prediction models: A step-by-step guide. *BMJ* **386**, e078276 (2024). <https://doi.org/10.1136/bmj-2023-078276>
26. Collins, G. S. et al. Evaluation of clinical prediction models (part 1): From development to external validation. *BMJ* **384**, e074819 (2024). <https://doi.org/10.1136/bmj-2023-074819>
27. Riley, R. D. et al. Evaluation of clinical prediction models (part 2): How to undertake an external validation study. *BMJ* **384**, e074820 (2024). <https://doi.org/10.1136/bmj-2023-074820>
28. Riley, R. D. et al. Evaluation of clinical prediction models (part 3): Calculating the sample size required for an external validation study. *BMJ* **384**, e074821 (2024). <https://doi.org/10.1136/bmj-2023-074821>
29. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L. & Steyerberg, E. W. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **17**, 230 (2019). <https://doi.org/10.1186/s12916-019-1466-7>

30. Vickers, A. J., van Calster, B. & Steyerberg, E. W. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn. Progn. Res.* **3**, 18 (2019). <https://doi.org/10.1186/s41512-019-0064-7>
31. Stiglic, G. et al. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **10**, e1379 (2020). <https://doi.org/10.1002/widm.1379>

Tables

Table 1. Definitions and formulas for composite inflammatory-nutritional indices

Index	Formula	Interpretation
NLR	Neutrophil count / lymphocyte count	Relative neutrophil and lymphocyte levels
PLR	Platelet count / lymphocyte count	Relative platelet and lymphocyte levels
MLR	Monocyte count / lymphocyte count	Relative monocyte and lymphocyte levels
SII	Platelet count × neutrophil count / lymphocyte count	Systemic immune-inflammation index
SIRI	Neutrophil count × monocyte count / lymphocyte count	Systemic inflammation response index
PIV	Neutrophil count × platelet count × monocyte count / lymphocyte count	Pan-immune-inflammation value
PNI	Albumin (g/L) + 5 × lymphocyte count (10 ⁹ /L)	Nutritional and immune-status index
HALP	Hemoglobin × albumin × lymphocyte count / platelet count	Composite hematological, nutritional, and immune index
ALI	BMI × albumin / NLR	Composite body-mass, nutritional, and inflammatory index

Table 2. Baseline characteristics of men with PSA 4–20 ng/mL

Variable	Non-PCa (n=765)	csPCa (n=76)	Clinically insignificant PCa (n=18)	P value
Age, years	71.30 ± 7.73	70.36 ± 7.34	68.83 ± 8.53	0.259
BMI, kg/m ²	23.44 ± 3.38	23.65 ± 3.35	23.45 ± 2.82	0.877
tPSA, ng/mL	8.64 ± 3.83	10.67 ± 3.30	9.23 ± 3.99	<0.001
fPSA, ng/mL	1.56 ± 0.91	1.30 ± 0.51	1.56 ± 0.57	0.055
f/t PSA	0.20 ± 0.12	0.12 ± 0.03	0.18 ± 0.07	<0.001
Prostate volume, mL	59.80 ± 23.25	54.51 ± 20.78	59.85 ± 24.99	0.162
PSAD	0.16 ± 0.08	0.24 ± 0.16	0.17 ± 0.08	<0.001
Neutrophil count, 10 ⁹ /L	4.33 ± 2.02	4.19 ± 2.19	3.90 ± 1.27	0.583
Lymphocyte count, 10 ⁹ /L	1.53 ± 0.61	1.51 ± 0.48	1.59 ± 0.70	0.873
Monocyte count, 10 ⁹ /L	0.45 ± 0.19	0.43 ± 0.13	0.44 ± 0.15	0.497
Platelet count, 10 ⁹ /L	182.86 ± 65.57	184.63 ± 53.35	183.78 ± 72.12	0.973
Hemoglobin, g/L	131.68 ± 17.01	134.89 ± 16.64	131.06 ± 23.94	0.291
Albumin, g/L	40.29 ± 3.80	40.89 ± 3.77	40.34 ± 3.16	0.425
NLR	3.35 ± 3.07	3.13 ± 2.33	2.83 ± 1.50	0.652
PLR	134.76 ± 73.57	133.70 ± 56.32	129.94 ± 79.37	0.956
MLR	0.33 ± 0.19	0.32 ± 0.16	0.32 ± 0.16	0.723
SII	620.43 ± 616.99	578.52 ± 438.72	566.96 ± 575.28	0.797
SIRI	1.62 ± 1.68	1.51 ± 1.82	1.31 ± 1.05	0.653
PIV	309.17 ± 371.01	282.77 ± 331.47	287.19 ± 394.56	0.817
PNI	47.94 ± 5.19	48.45 ± 5.14	48.31 ± 5.03	0.690
HALP	49.58 ± 27.08	51.02 ± 31.18	52.66 ± 31.78	0.823
ALI	396.13 ± 227.37	414.64 ± 223.14	418.00 ± 203.82	0.740
AST, U/L	23.84 ± 17.01	22.80 ± 7.42	23.62 ± 7.94	0.868
ALT, U/L	21.00 ± 14.80	21.35 ± 10.27	23.22 ± 12.61	0.801
AST/ALT	1.57 ± 6.53	1.18 ± 0.37	1.15 ± 0.33	0.842
Total bilirubin, µmol/L	14.18 ± 6.15	13.74 ± 6.07	12.46 ± 5.19	0.431
Blood urea nitrogen, mmol/L	6.03 ± 2.68	6.05 ± 2.34	5.83 ± 1.54	0.948
Creatinine, µmol/L	84.03 ± 56.56	76.08 ± 27.14	85.26 ± 47.59	0.473
Hypertension				0.938
No	579 (75.7%)	57 (75.0%)	13 (72.2%)	
Yes	186 (24.3%)	19 (25.0%)	5 (27.8%)	
Diabetes				0.841
No	671 (87.7%)	66 (86.8%)	15 (83.3%)	
Yes	94 (12.3%)	10 (13.2%)	3 (16.7%)	

Variable	Non-PCa (n=765)	csPCa (n=76)	Clinically insignificant PCa (n=18)	P value
Smoking history				0.919
No	448 (58.6%)	43 (56.6%)	10 (55.6%)	
Yes	317 (41.4%)	33 (43.4%)	8 (44.4%)	
Alcohol-use history				0.143
No	392 (51.2%)	39 (51.3%)	5 (27.8%)	
Yes	373 (48.8%)	37 (48.7%)	13 (72.2%)	

Note: Continuous variables are presented as mean ± standard deviation and categorical variables as n (%). Continuous variables were compared with one-way analysis of variance or the Kruskal–Wallis test according to distribution; categorical variables were compared with Pearson's χ^2 test or Fisher's exact test. BMI, body mass index; PCa, prostate cancer; csPCa, clinically significant prostate cancer; tPSA, total prostate-specific antigen; fPSA, free prostate-specific antigen; f/t PSA, free-to-total PSA ratio; PSAD, PSA density. P values are for overall three-group comparisons.

Table 3. Patient-level OOB predictive performance of the 12 machine-learning models

Model	AUROC (95% CI)	AUPRC (95% CI)	Brier score (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	MCC (95% CI)	F1 score (95% CI)
C5.0	0.843 (0.801– 0.886)	0.421 (0.325– 0.536)	0.066 (0.061– 0.070)	0.750 (0.658– 0.842)	0.765 (0.736– 0.794)	0.237 (0.205– 0.271)	0.969 (0.958– 0.981)	0.326 (0.261– 0.392)	0.360 (0.314– 0.408)
AdaBoost	0.820 (0.776– 0.862)	0.415 (0.319– 0.517)	0.073 (0.068– 0.077)	0.829 (0.750– 0.908)	0.695 (0.663– 0.728)	0.209 (0.186– 0.236)	0.977 (0.966– 0.988)	0.311 (0.255– 0.370)	0.333 (0.299– 0.373)
GBM	0.818 (0.770– 0.862)	0.402 (0.303– 0.504)	0.067 (0.060– 0.073)	0.842 (0.763– 0.921)	0.668 (0.635– 0.701)	0.198 (0.176– 0.220)	0.978 (0.966– 0.989)	0.299 (0.245– 0.350)	0.320 (0.287– 0.354)
XGBoost	0.815 (0.769– 0.860)	0.352 (0.262– 0.464)	0.069 (0.065– 0.073)	0.776 (0.684– 0.855)	0.731 (0.699– 0.762)	0.219 (0.191– 0.247)	0.971 (0.959– 0.981)	0.310 (0.249– 0.367)	0.341 (0.301– 0.382)
NB	0.807 (0.763– 0.848)	0.301 (0.217– 0.394)	0.073 (0.067– 0.078)	0.763 (0.671– 0.855)	0.719 (0.687– 0.750)	0.209 (0.182– 0.238)	0.969 (0.957– 0.981)	0.293 (0.231– 0.353)	0.328 (0.287– 0.370)
CART	0.807 (0.755– 0.852)	0.407 (0.307– 0.505)	0.065 (0.059– 0.071)	0.908 (0.829– 0.961)	0.591 (0.558– 0.623)	0.177 (0.162– 0.194)	0.985 (0.973– 0.994)	0.285 (0.238– 0.324)	0.297 (0.272– 0.321)
RF	0.802 (0.758– 0.844)	0.393 (0.300– 0.499)	0.068 (0.064– 0.072)	0.895 (0.816– 0.961)	0.581 (0.544– 0.616)	0.172 (0.157– 0.188)	0.983 (0.971– 0.993)	0.271 (0.227– 0.311)	0.288 (0.264– 0.313)
NN	0.796 (0.750– 0.843)	0.304 (0.223– 0.405)	0.071 (0.067– 0.076)	0.789 (0.697– 0.882)	0.655 (0.623– 0.688)	0.182 (0.160– 0.205)	0.970 (0.957– 0.983)	0.260 (0.204– 0.317)	0.296 (0.262– 0.331)
SVM	0.782 (0.732– 0.831)	0.236 (0.187– 0.313)	0.074 (0.071– 0.077)	0.645 (0.539– 0.750)	0.801 (0.771– 0.828)	0.239 (0.201– 0.281)	0.959 (0.947– 0.970)	0.297 (0.223– 0.371)	0.349 (0.293– 0.406)
KNN	0.767 (0.715– 0.818)	0.264 (0.188– 0.358)	0.073 (0.069– 0.077)	0.789 (0.684– 0.882)	0.628 (0.596– 0.660)	0.171 (0.150– 0.193)	0.969 (0.954– 0.982)	0.241 (0.181– 0.298)	0.281 (0.247– 0.314)
MLP	0.765 (0.714– 0.812)	0.254 (0.184– 0.351)	0.074 (0.069– 0.079)	0.671 (0.566– 0.776)	0.711 (0.679– 0.741)	0.184 (0.156– 0.213)	0.957 (0.943– 0.970)	0.232 (0.166– 0.298)	0.289 (0.246– 0.333)
LR	0.733 (0.683– 0.780)	0.216 (0.156– 0.301)	0.078 (0.073– 0.083)	0.803 (0.711– 0.895)	0.587 (0.553– 0.622)	0.159 (0.141– 0.177)	0.968 (0.954– 0.983)	0.223 (0.167– 0.275)	0.265 (0.235– 0.294)

Note: Values are point estimates (95% CIs). AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve; PPV, positive predictive value; NPV, negative predictive value; MCC, Matthews correlation coefficient. Threshold-dependent measures were calculated from patient-level aggregated OOB predictions using each model's Youden cutoff. The 95% CIs were estimated with 2,000 patient-level bootstrap resamples stratified by outcome. Brier scores were calculated from the original OOB probabilities without recalibration on the same data.

Figures

Figure 1

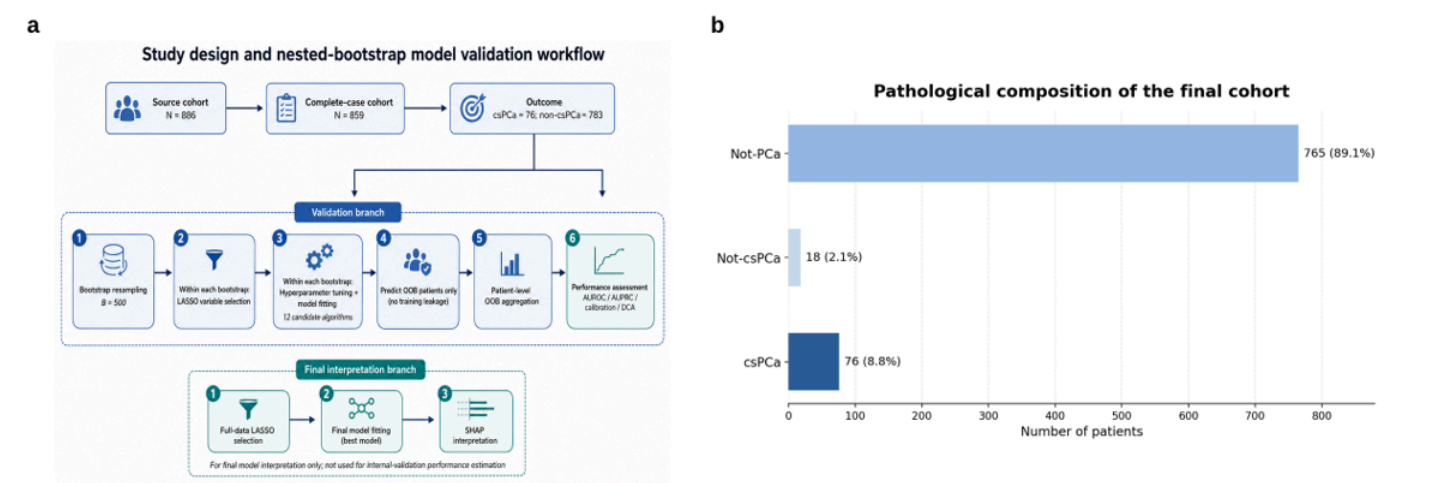
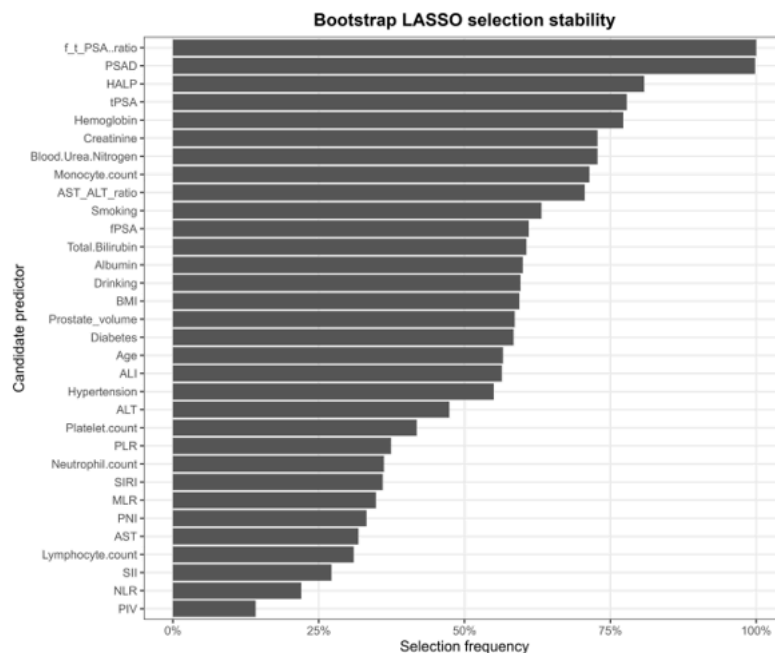


Figure 1

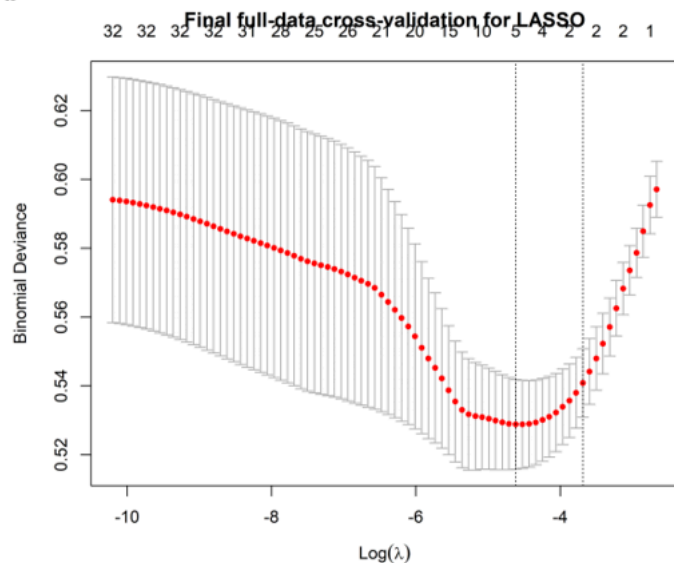
Study design and cohort composition. (a) Study design and nested-bootstrap model-validation workflow. LASSO selection, hyperparameter tuning, and training were repeated within each bootstrap training sample, and only OOB predictions were used to estimate internal-validation performance. The separate final-interpretation branch was used only for the representative full-data model and SHAP analysis. (b) Pathological composition of the final cohort.

Figure 2

a



b



c

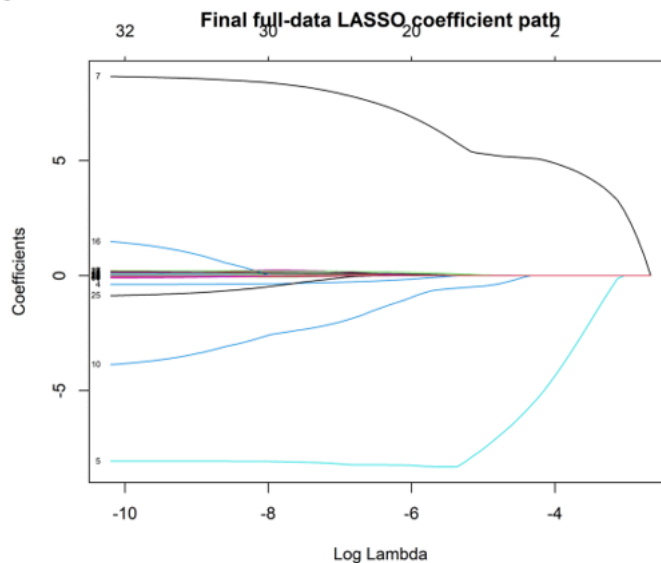


Figure 2

Feature-selection stability and full-data LASSO diagnostics. (a) Frequency with which each candidate predictor was selected by LASSO across 500 bootstrap resamples. (b) Cross-validation curve for the final full-data LASSO model. (c) Coefficient paths for the final full-data LASSO model. Panels b and c describe the representative final model only and were not used to estimate internal-validation performance.

Figure 3

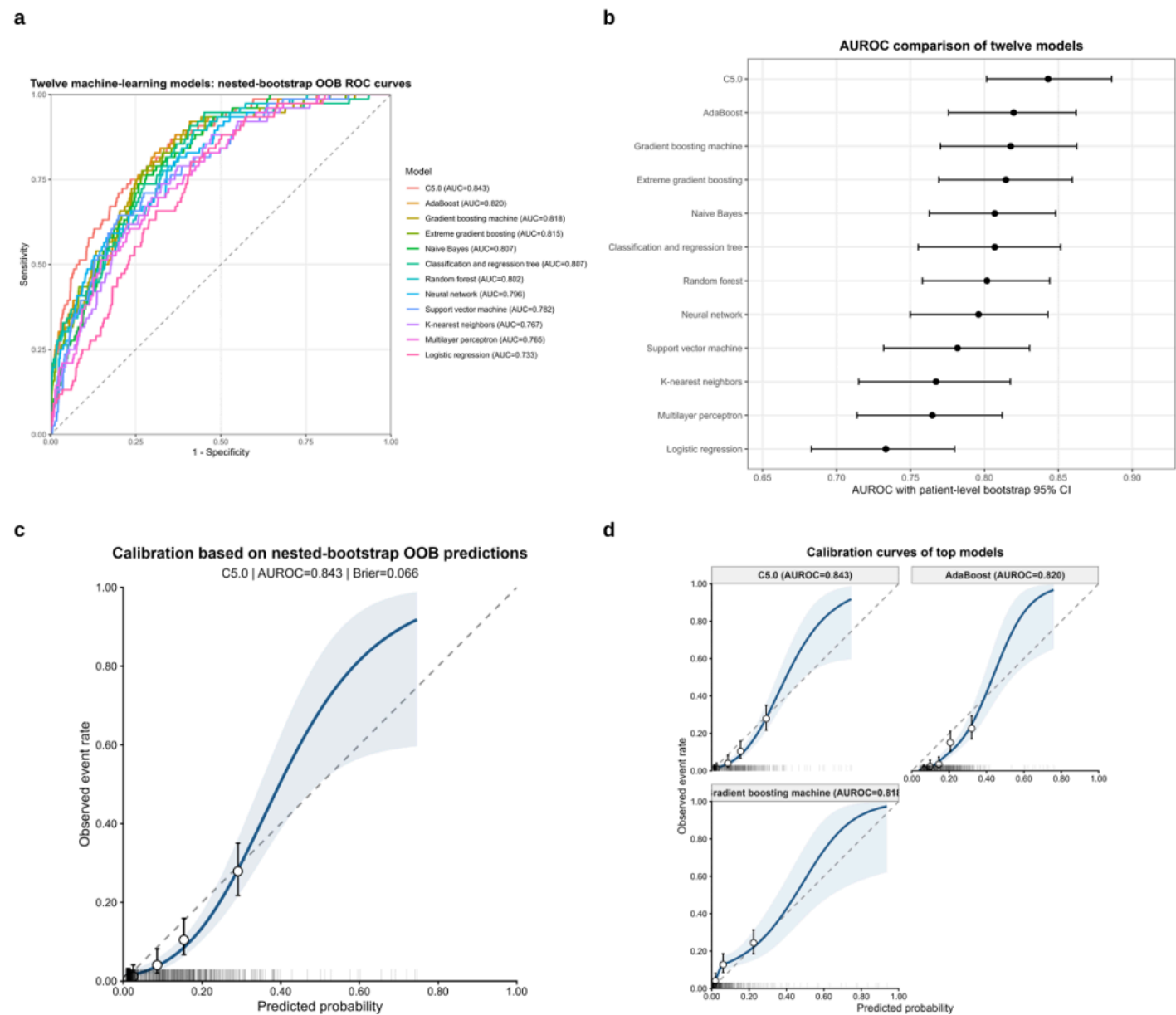


Figure 3

Discrimination and calibration of the candidate models. (a) Patient-level nested-bootstrap OOB receiver operating characteristic curves for the 12 machine-learning models. (b) AUROC point estimates and patient-level bootstrap 95% CIs. (c) Smoothed calibration curve for C5.0 based on patient-level OOB predictions; the solid line is the binomial GAM smooth, the shaded region is the 95% confidence band, points and error bars show observed event rates and Wilson 95% CIs in five equally sized risk groups, and the dashed line indicates ideal calibration. (d) Patient-level OOB smoothed calibration curves for the three leading models.

Figure 4

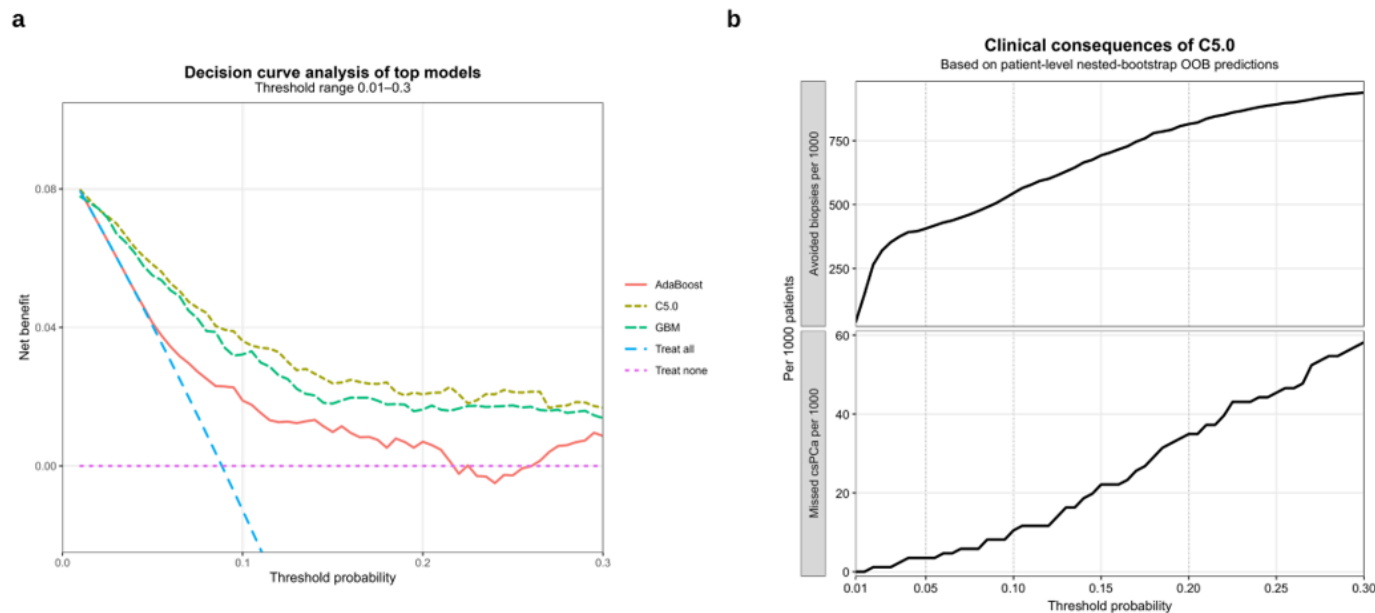


Figure 4

Clinical utility and threshold consequences. (a) Decision curve analysis for the better-performing models across threshold probabilities of 0.01–0.30. (b) Biopsies avoided and csPCa cases missed per 1,000 men across C5.0 risk thresholds from 0.01 to 0.30.

Figure 5

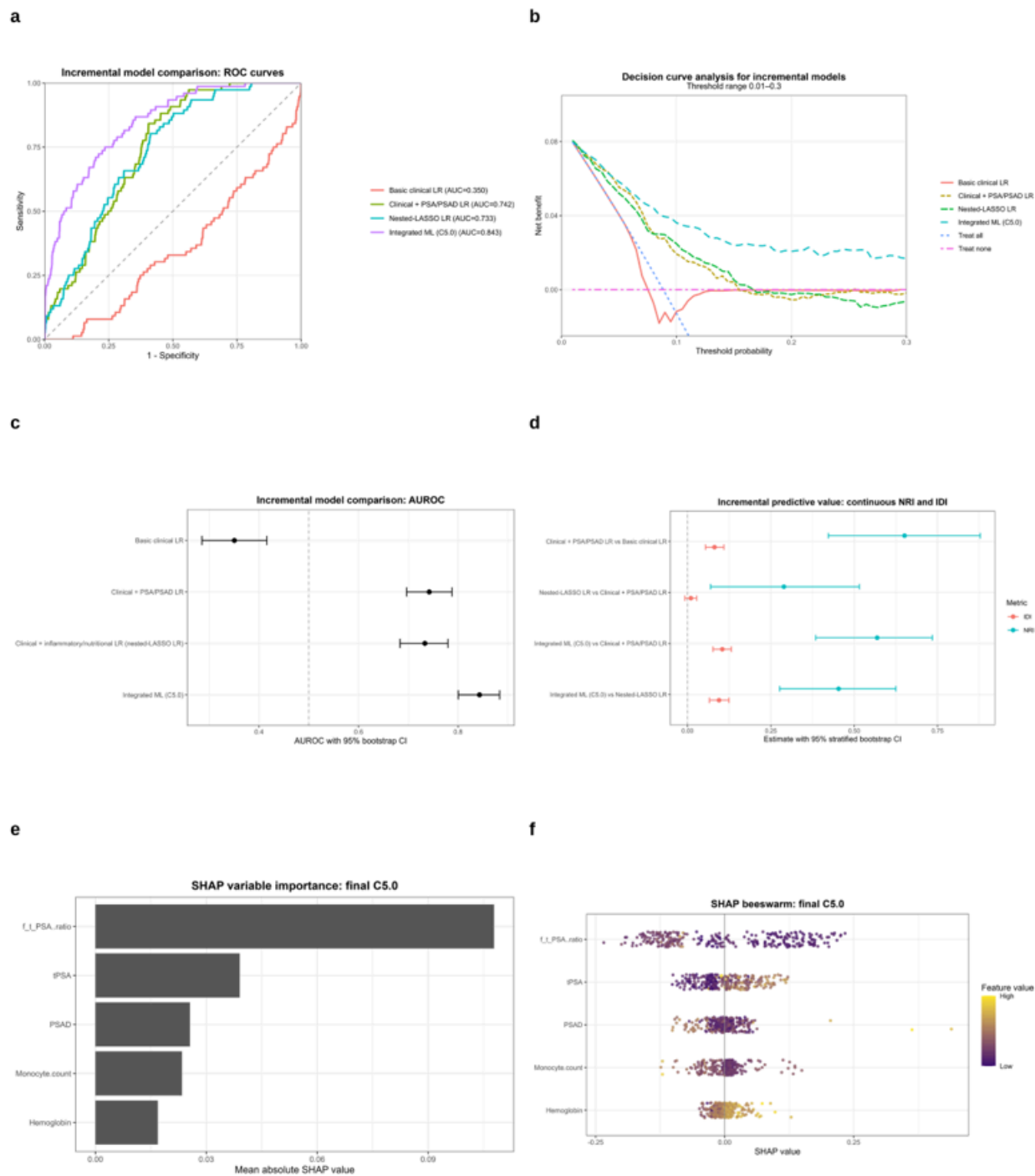


Figure 5

Progressive-model performance, incremental value, and interpretation of the representative C5.0 model. (a) Receiver operating characteristic curves for the four progressive models. (b) Decision curve analysis for the same models; M3 denotes nested-LASSO logistic regression and is not a strict blockwise clinical-plus-inflammatory/nutritional model. (c) AUROC point estimates and 95% CIs. (d) Continuous NRI and IDI with patient-level stratified bootstrap 95% CIs. (e) Mean absolute SHAP values for the representative full-data C5.0 model. (f) SHAP beeswarm plot. Panels e and f describe one representative final model and do not imply a common variable mechanism across all 500 bootstrap-specific models.