HHS Public Access

Author manuscript

Eur Urol Oncol. Author manuscript; available in PMC 2026 February 08.

Published in final edited form as:

Eur Urol Oncol. 2026 February ; 9(1): 142–149. doi:10.1016/j.euo.2025.01.005.

Comparative Performance of Machine Learning Models in Reducing Unnecessary Targeted Prostate Biopsies

Fuyao Chen[a,b,†], **Roxana Esmaili**[c,d,†], **Ghazal Khajir**[e], **Tal Zeevi**[a], **Moritz Gross**[c,d], **Michael Leapman**[e], **Preston Sprenkle**[e], **Amy C. Justice**[f,g], **Sandeep Arora**[c], **Jeffrey C. Weinreb**[c], **Michael Spektor**[c], **Steffan Huber**[c], **Peter A. Humphrey**[h], **Angelique Levi**[h], **Lawrence H. Staib**[a,c,i], **Rajesh Venkataraman**[j], **Darryl T. Martin**[e], **John A. Onofrey**[a,c,e,*]

[a]Department of Biomedical Engineering, Yale University, New Haven, CT, USA

[b]Medical Scientist Training Program, Yale School of Medicine, New Haven, CT, USA

[c]Department of Radiology and Biomedical Imaging, Yale University, New Haven, CT, USA

[d]Charité Center for Diagnostic and Interventional Radiology, Charité-Universitätsmedizin Berlin, Berlin, Germany

[e]Department of Urology, Yale University, New Haven, CT, USA

[f]VA Connecticut Healthcare, West Haven, CT, USA

[g]Department of Internal Medicine, Yale University, New Haven, CT, USA

[h]Department of Pathology, Yale University, New Haven, CT, USA

[i]Department of Electrical Engineering, Yale University, New Haven, CT, USA

[j]Eigen Health, Grass Valley, CA, USA

Abstract

[*]Corresponding author. Department of Radiology and Biomedical Imaging, Department of Urology, Department of Biomedical Engineering, PO Box 208042, New Haven, CT 06520-8042, USA. john.onofrey@yale.edu (J.A. Onofrey).

[†]These authors are joint first authors.

Author contributions: John A. Onofrey had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Chen, Esmaili, Onofrey.

Acquisition of data: Chen, Esmaili, Khajir, Sprenkle, Venkataraman, Martin, Onofrey.

Analysis and interpretation of data: Chen, Esmaili, Khajir, Zeevi, Gross, Leapman, Sprenkle, Justice, Arora, Weinreb, Spektor, Huber, Humphrey, Levi, Staib, Venkataraman, Martin, Onofrey.

Drafting of the manuscript: Chen, Esmaili, Khajir, Zeevi, Gross, Leapman, Sprenkle, Justice, Arora, Weinreb, Spektor, Huber, Humphrey, Levi, Staib, Venkataraman, Martin, Onofrey.

Critical revision of the manuscript for important intellectual content: Chen, Esmaili, Khajir, Zeevi, Gross, Leapman, Sprenkle, Justice, Arora, Weinreb, Spektor, Huber, Humphrey, Levi, Staib, Venkataraman, Martin, Onofrey.

Statistical analysis: Chen, Esmaili, Khajir, Zeevi, Gross, Onofrey.

Obtaining funding: Venkataraman, Onofrey.

Administrative, technical, or material support: Chen, Esmaili, Khajir, Zeevi, Gross, Onofrey.

Supervision: Onofrey.

Other: None.

Background and objective: Conventional core needle biopsy for prostate cancer diagnosis can lead to diagnostic uncertainty and complications, prompting exploration of alternative risk assessment approaches that use clinical and imaging features. Our aim was to evaluate the effectiveness of machine learning (ML) models in reducing unnecessary biopsies.

Methods: We conducted a retrospective analysis of data for 1884 patients across two academic centers who underwent prostate magnetic resonance imaging and biopsy between 2016 and 2020 or 2004 and 2011. Twelve ML models were assessed for prediction of clinically significant prostate cancer (csPCa; Gleason grade group ≥2) using combinations of clinical features, including patient age, prostate-specific antigen level and density, Prostate Imaging-Reporting and Data System/Likert score, lesion volume, and gland volume. The models were trained and validated using a tenfold split for intrasite, intersite, and combined-site data sets. Model effectiveness was evaluated using the area under the receiver operating characteristic curve and decision curve analysis.

Key findings and limitations: The best-performing ML model would reduce the number of biopsies by 13.07% at a false-negative rate of 1.91%. Performance was consistent across sites, although the study is limited by the small number of centers and the absence of specific clinical data.

Conclusions and clinical implications: ML-enhanced clinical models provide an effective and generalizable approach for prediction of csPCa using standard clinical data. These models allow personalized risk assessment and follow-up, support clinical decision-making, and improve workflow efficiency.

Patient summary: Models that are enhanced by machine learning can predict the severity of prostate cancer and help doctors in tailoring treatments for individual patients. This approach can simplify health care decisions and improve clinical efficiency.

Abstract

Our results demonstrate that a machine learning model comprising standard clinical data such as patient age, prostate-specific antigen, and imaging results could reduce unnecessary prostate biopsies by 13.07%, with a false-negative rate of 1.91% for clinically significant prostate cancer. Model performance was consistent across centers. This approach could personalize risk assessment and enhance the efficiency of clinical workflows.

Keywords

Prostate cancer; Multiparametric magnetic resonance imaging; Prostate Imaging-Reporting and Data System; Prostate-specific antigen density; Unnecessary biopsies; Machine learning

1. Introduction

Prostate cancer (PCa) is one of the most common cancers among men worldwide [1,2], prompting early detection efforts via prostate-specific antigen (PSA) screening [3]. While multiparametric magnetic resonance imaging (mpMRI) augments and significantly enhances diagnosis in comparison to screening via PSA alone [3,4], needle biopsy remains the gold standard [3]. However, biopsies are prone to errors: only 0.45% of prostatic tissue is sampled

[5] and 50% of the tumors detected are indolent, so there are risks of both overtreatment and missing cancer heterogeneity, with discrepancies from prostatectomy pathology in 40% of cases [6,7]. The invasive nature of biopsy is associated with procedural complications such as hematuria (up to 11.6%), hematochezia (up to 1.6%), hemospermia (up to 12%), and potential for infection (up to 2%) [8], besides discomfort and anxiety.

Risk stratification studies have proposed combination of mpMRI with clinical factors to predict PCa significance and reduce unnecessary biopsies [9–18]. Threshold strategies that combine current diagnostic tools have been based on PSA, PSA density (PSAd), and MRI markers [9]. However, in comparison to clinician-defined diagnostic strategies, machine learning (ML) techniques may identify disease patterns not immediately apparent to clinicians. While numerous ML approaches have been explored [10–18], studies directly comparing the clinical effectiveness, generalizability, and interpretability of ML models in the PCa diagnostic workflow remain scarce, hindering their clinical implementation in the contemporary era for patients undergoing PCa screening and mpMRI. The aim of our study was to evaluate the performance of diverse ML models previously applied for the diagnosis of clinically significant PCa (csPCa) [9–18] and assess the predictive power of different combinations of clinical features in two independent cohorts from two different clinical centers.

2. Patients and methods

2.1. Study population

This retrospective multicenter study included two cohorts of patients: one from Yale New Haven Hospital (YNHH) between 2016 and 2020, and one from University of California-Los Angeles Health (UCLA) between 2004 and 2011, who underwent mpMRI before MRI-targeted transrectal ultrasound (TRUS) fusion biopsy (Fig. 1). For the YNHH cohort, the study was approved by the institutional review board of Yale University School of Medicine with waiver of informed consent (reference 2000024079) and was compliant with the Health Insurance Portability and Accountability Act. For the UCLA cohort, the public data set is available under approval from the UCLA institutional review board with informed consent obtained from patients [19–21]. For both cohorts, pathology findings on biopsy are reported as a single Gleason grade group [22] assigned per patient. For each biopsy sample, both data sets included the Prostate Imaging-Reporting and Data System (PI-RADS) score [23] (or a similar Likert scale score for the UCLA cohort [19–21]), gland volume, lesion volume, PSA value, PSAd, and patient age.

2.2. Imaging with mpMRI

The mpMRI protocol at both institutions included T2-weighted (T2W) imaging, diffusion-weighted imaging with apparent diffusion coefficient maps, and dynamic contrast-enhanced imaging, collected via a 3.0-T scanner (Siemens, Erlangen, Germany). In both cohorts the prostate and suspected lesion regions were contoured on T2W images via semi-automated segmentation using ProFuseCAD software (Eigen Health, Grass Valley, CA, USA) for prostate biopsy planning. Visible MRI lesions were scored according to PI-RADS v2 [23] on a scale from 1 to 5 points and the lesions were annotated by radiologists at YNHH.

A 5-point Likert scale with close correspondence to PI-RADS v2 was used for the UCLA cohort [19–21], with a score of ≥3 considered to correlate with the risk of csPCa.

2.3. Histopathology

For both cohorts, patients underwent MRI-TRUS fusion targeted biopsy using an Artemis biopsy device (Eigen Health). YNHH and UCLA patients underwent systematic or targeted biopsy or both. For the present study we did not consider systematic biopsies. For suspicious lesions on MRI (PI-RADS/Likert score ≥3), between one and five cores were taken from each lesion. Histopathology evaluation according to the Gleason grading system [22] was performed by pathologists for all needle biopsy cores, with each Gleason grade score per core subsequently converted to a Gleason grade group (GG) value. If more than one biopsy was taken from a lesion, the highest GG was used.

2.4. Evaluation and statistical analysis

Descriptive statistics were calculated for the overall study cohort and for the UCLA and YNHH cohorts individually. Continuous variables are reported as the median with interquartile range. Comparisons between the cohorts were conducted using the Mann-Whitney U test for continuous variables, and Pearson's χ^2 test for categorical variables. The predictive power of the models was assessed using the area under the receiver operating characteristic curve (AUC) and the clinical utility in terms of the net benefit ratio was evaluated via decision curve analysis. AUC values were compared using the DeLong test [24]. All tests were two-sided, with significance set at $p < 0.05$. For multiple comparisons during tenfold cross-validation, Bonferroni correction was applied to the significance level ($p < 0.005$). Statistical analyses were performed with Python v3.8.8 and R v4.2.1.

2.5. Data preparation

For each cohort, data were randomly divided into independent training and testing subsets using ten-fold cross validation across subjects (Scikit-learn Library 0.24.2, Python v3.8.8). For the YNHH cohort, the training set for each iteration contained 989 patients (1372–1390 lesions), while the testing set contained 110 patients (139–165 lesions). For the UCLA cohort, the training set comprised 709 patients (1091–1116 lesions) and the testing set comprised 79 patients (108–133 lesions) per iteration. Lesions from the same patients were exclusive to each iteration. Eight feature groups (FGs) were evaluated, comprising combinations of patient age, PSA value, PSAd, PI-RADS/Likert score, lesion volume, and gland volume (Fig. 2A). FGs were constructed to reflect the temporal availability of data before prostate biopsy, as a patient proceeded from baseline (FG1) to PSA screening (FG2) to mpMRI (FG3, FG4) and image review by radiologists (FG5, FG6) and eventually repeat PSA screening for those on active surveillance (FG7, FG8).

2.6. Machine learning

The primary outcome was the detection rate for csPCa, defined as GG ≥2. Twelve ML algorithms were applied to both cohorts under tenfold cross-validation using eight defined feature groups (Fig. 2A): logistic regression (LR), random forest (RF), k-nearest neighbors (kNN), naïve Bayes (NB), extreme gradient boosting (XGBoost), multilayer perceptron

(MLP), Gaussian process classifier (GPC), linear support vector machine (LSVM), radial basis function kernel support vector machine (RBF-SVM), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and decision tree (DT). These models were selected to ensure diverse representation of different ML approaches and their important variations. Further model details and implementation are described in the Supplementary material.

To assess model generalizability, intrasite, intersite, and combined-site analyses were performed. For intrasite analysis, the ML models were trained and tested on data from a single center (either YNHH or UCLA). For intersite analysis, the models were tested on the center that was not involved in model training. The best-performing models on intrasite analysis were used for intersite analysis. For the combined-site analysis, the training data from both YNHH and UCLA were combined to train the selected ML models, which were then tested on YNHH and UCLA combined and separate testing sets.

3. Results

3.1. Patient characteristics

A total of 3201 lesion biopsy results from a single visit were collected for 1518 patients from YNHH. The UCLA cohort involved 1233 lesion biopsies from 791 patients. Patients with missing data for PSAd, PI-RADS/Likert score, age, or lesion/gland volume were excluded, leaving 1099 patients (1537 lesion biopsies) from YNHH and 785 patients (1224 lesion biopsies) from UCLA for analysis (Fig. 1). The prevalence of csPCa (Table 1) was 40.4% (444 of 1099 patients) in the YNHH cohort and 38.59% (303 of 785 patients) in the UCLA cohort (Pearson's χ^2 test, $p = 0.46$).

3.2. Model evaluation

Feature coefficients and intercepts were obtained for selected models with linear decision boundaries (LR and LSVM), while feature importance results were obtained for the RF model (Supplementary Tables 1–9). For both FG5 and FG6, PSAd and PI-RADS scores had the highest importance scores for the RF model, and the relative importance of features was stable across sites. While PSAd and PI-RADS/Likert scores had the highest coefficients for YNHH-trained LR and LSVM models, PSA and PI-RADS/Likert scores had the highest coefficients for the UCLA-trained LR model. For the UCLA-trained LR and LSVM models, the weights for PSA varied significantly between validation iterations.

3.3. Intrasite analysis

For ML models trained and tested on the YNHH cohort, the highest csPCa discriminatory ability was observed for the RF model (AUC 0.82, 95% confidence interval [CI] 0.79–0.85) and the LSVM model (AUC 0.82, 95% CI 0.80–0.84) with FG5. The GPC and MLP models had the highest AUC (0.82, 95% CI 0.80–0.84) with FG6 (Fig. 2B and Supplementary Fig. 1B), significantly outperforming other FGs ($p < 0.005$). These models consistently outperformed the other ML methods, except the LR model, with both FG5 and FG6 ($p < 0.005$). Model performance improved with addition of clinical features (sequentially PSA, gland volume, PSAd, PI-RADS/Likert score, and lesion volume), as reflected by

increases in AUC from FG1 to FG5. For the UCLA cohort, the RF (AUC 0.79, 95% CI 0.73–0.85) and GPC (AUC 0.79, 95% CI 0.74–0.84) models had the highest AUC values for FG6, outperforming the NB, XGBoost, LSVM, QDA, and RBF-SVM models ($p < 0.005$; Fig. 2B). However, their AUC values did not significantly differ from those with FG5. Decision curve analysis (Fig. 2C and Supplementary Fig. 1C) showed that the LR, RF, GPC, MLP, and LSVM models offer a net clinical benefit in comparison to a biopsy-all strategy when the risk threshold exceeds 10% for YNHH and 15% for UCLA. The clinical benefit improves with the addition of more clinical features, particularly in FG5 and FG6, for which the LR, RF, GPC, MLP, and LSVM models showed the most consistent performance across the two cohorts.

3.4. Intersite and multisite analyses

When trained with combined data, the RF model achieved an AUC of 0.79 (95% CI 0.72–0.83) for the UCLA testing set, 0.81 (95% CI 0.78–0.82) for the YNHH testing set, and 0.80 (95% CI 0.78–0.82) for the combined set when applied to FG6 (Fig. 3). Similarly, the GPC and MLP models showed consistent results (AUC 0.80, 95% CI 0.78–0.81) for the combined testing set. The LSVM model yielded an overall AUC of 0.80 (95% CI 0.78–0.81) for the combined testing set, but the performance was greater for the YNHH subset (AUC 0.82, 95% CI 0.80–0.84) than for the UCLA subset (AUC 0.77, 95% CI 0.73–0.82). The LR model yielded an overall AUC of 0.79 (95% CI 0.77–0.81) for the combined testing set.

3.5. Proportion of biopsies avoided

Figure 4 shows the proportion of biopsies that could be avoided according to selected models for FG6 at different missed case rates. Results derived using clinician-defined strategies based on PSAd and PI-RADS scores [9] are plotted for comparison. The RF, GPC, and MLP models outperformed the clinician-defined strategies, especially at a missing case threshold of <2%. For instance, the MLP model would avoid 13.07% of biopsies while missing 1.91% of csPCa cases. The performance of the RF, GPC, and MLP models was stable across sites (Supplementary Figs. 3 and 4). Detailed results for the csPCa probability at different missing case thresholds are provided in Supplementary Table 10.

4. Discussion

In this multicenter study, we assessed the utility of 12 ML models across eight combinations of clinical features using intrasite, intersite, and multisite analyses for two independent cohorts from two clinical centers. We comprehensively evaluated the performance of the models to provide insight into the utility of different combinations of clinical features according to the prebiopsy diagnostic workflow.

The RF, GPC, and MLP models applied to FG5 and FG6 (including patient age, PSA, gland volume, PSAd, PI-RADS/Likert score, and lesion volume) had the highest discriminatory ability for csPCa (Supplementary Fig. 1B). The performance increased with addition of more clinical variables (FG1–FG5, sequential addition of PSA, gland volume, PSAd). However, addition of lesion volume in FG6, did not further improve the discriminatory performance for many models (Supplementary Fig. 1B). This might be because of high

inter-rater variability for lesion segmentation given variable lesion sizes [25]. Among all the features, the PI-RADS/Likert score led to the most significant improvement in discriminatory performance, despite temporal shifts and differences in imaging assessment (PI-RADS v2 for YNHH and Likert scale for UCLA) between the centers. While the LR and LSVM models performed well for the YNHH data, their performance was lower for the UCLA cohort (Supplementary Figs. 1, 3, and 4). Feature coefficients for both models trained with UCLA data demonstrated variation between validation iterations (Supplementary Tables 1–9), suggesting potential model vulnerability to data-set shifts.

In terms of clinical utility, the RF, GPC, and MLP models outperformed clinician-defined threshold strategies with clinically reasonable tolerance for missing cases (<2%). Clinician-defined threshold strategies can only yield a specific missing case rate and corresponding avoidable biopsy rate for a given data cohort, whereas ML models have the flexibility to estimate avoidable biopsy rates for any missing case rate. Application of the best-performing models to independent cohort data demonstrated their generalizability (Supplementary Fig. 2), with consistent discriminatory ability and net clinical benefit observed. Combining the same number of YNHH and UCLA patients for training did not fully eliminate the higher performance for the YNHH cohort (Supplementary Fig. 3), possibly because of intrinsic data variation in the UCLA cohort. Among the best-performing models, RF showed consistent feature importance and thus provides intrinsic model interpretability. For the GPC and MLP models, other inspection techniques such as premutation importance can be used for model understanding.

Our study, which used fully data-driven approaches and data from standard-of-care practice, demonstrated robust performance of ML models for independent cohorts from two clinical centers. Minimal impact of interinstitutional heterogeneity on the applicability of selected models indicates their potential for use in multicenter clinical trials.

Several prediction models for PCa have been proposed. The readily available Prostate Cancer Prevention Trial Risk Calculator 2.0 (PCPTRC2) [26], although externally validated using ten international cohorts, does not include imaging biomarkers. While Falagario et al [9] combined PSAd and PI-RADS scores, their risk stratification strategies were clinician-defined instead of data-driven. A few studies applied an LR model to clinical features that included age, family history, race and ethnicity, PSAd, and PI-RADS scores [10–12]. The performance results for our models are in line with these studies (AUC 0.80–0.88), but direct comparison is challenging given differences in the clinical features used. Furthermore, these models were developed using data from small single-center cohorts (393–1159 patients) [11,12] and were independently validated in even smaller cohorts (198–251 patients) [11,12] or lacked external validation [10]. Limited cohort sizes may result in overfitting to the development cohort, especially with the assumption of linearity between dependent and independent variables [27,28].

Beyond LR models, Rustam et al [13] compared SVM and NB classifiers for patient age, PSA, and blood cell counts, achieving PCa detection accuracy of 83.3% and 78.3%, respectively. Xiao et al [14] developed an RF algorithm using transrectal ultrasound data, achieving 83.1% accuracy. Chang et al [15] assessed a kNN model for PCa detection in a

cohort with benign prostate hyperplasia, with 81.7% accuracy. Perera et al [16] used a GPC model for PSA kinetics and obtaining AUC values ranging from 0.71 to 0.89 for csPCa prediction. Citak-Er et al [17] achieved sensitivity of 86.51% and specificity of 63.99% for an LDA model with mpMRI data. Finne et al [18] observed a false-positive rate of 19% for PSA results at 95% sensitivity using an MLP model.

All of these models were developed with either all of the clinical parameters available or with specific clinical data, whereas our study provide insights into different combinations of features according to their temporal availability and demonstrates the utility of ML models at different stages before prostate biopsy (Supplementary Fig. 1). Furthermore, previous studies did not investigate challenges common to ML, such as data-set shifts, whereby the distribution of data changes significantly between training and testing phases [27,28]. Our study provides a comprehensive evaluation and direct comparison of diverse ML models, with assessment of their clinical effectiveness, generalizability, and interpretability, and optimization of their use in modern workflows for PCa diagnosis.

For clinical implementation, all the data used for the models are readily available in electronic health records and could potentially be extracted automatically. Our best-performing ML model would avoid 13.07% of biopsies at a clinically reasonable (<2%) missing case rate (Fig. 4). Should the clinician and patient wish to jointly decide whether to proceed with biopsy, the models can be used in the shared decision-making process: an individual patient's data can be entered into the model, which will provide an estimated csPCa probability and the associated missing csPCa case rate (Fig. 4). The physician and patient can then discuss whether the risk level is acceptable should they choose not to proceed with biopsy.

Our analysis included only two data cohorts, so further assessment in additional cohorts with stratification for differences in patient populations, imaging, and biopsy protocols is needed to further determine the generalizability of the model. Given the limited number of cohorts, simple model hyperparameter selection was conducted for parameters with good performance in both cohorts (Supplementary Figs. 5–11). For potential multicenter trials, a validation set from each participating centers could be used for further hyperparameter optimization to potentially improve performance. We did not include race and ethnicity data [29], family history, and prior biopsy status [30], which were used in previous studies [10–12], because of unavailability for the UCLA cohort. Exploration of additional clinical information, such as urological follow-up, and imaging biomarkers, such as radiomics, is warranted to enhance model performance. Comparison to other clinical tools such as the PCPTRC2 risk calculator [26] could also offer insights into the utility of ML approaches. Our study sets a foundation for future studies on correlation of biopsy and radical prostatectomy pathology findings using ML models. We hope to eventually incorporate the validated model into a clinical decision support system for which the standard-of-care data used for its development are readily available.

5. Conclusions

Our retrospective multicenter study demonstrated that ML models comprising combinations of PSAd, PI-RADS/Likert score, patient age, lesion volume, and gland volume can aid in personalized decision-making regarding the need for prostate biopsy. The best-performing model would reduce unnecessary biopsies by 13.07% at the risk of missing 1.91% of csPCa cases. Our results show consistent performance across data sets, which confirms the generalizability of the models and supports their potential utility in multicenter trials in which interinstitutional and temporal variations may exist.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding/Support and role of the sponsor:

Research reported in this publication was supported by the National Cancer Institute of the US National Institutes of Health (NIH) under award number R42 CA224888. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work also was supported by the Office of the Assistant Secretary of Defense for Health Affairs via Prostate Cancer Research Program award W81XWH-20–1-0297. The opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense. Moritz Gross was supported by a travel stipend from the Rolf W. Günther Foundation for Radiological Sciences and Roxana Esmaili by a travel stipend from Ingeborg Rapoport Fellowship for their travel to Yale University. Fuyao Chen was supported by NIH Medical Scientist Training Program training grant T32GM007205. The sponsors played no role in the design and conduct of the study, and collection, management, analysis, and interpretation of the data.

Financial disclosures:

John A. Onofrey certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (eg, employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: None.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209–49. [PubMed: 33538338]
2. Bergengren O, Pekala KR, Matsoukas K, et al. 2022 update on prostate cancer epidemiology and risk factors—a systematic review. Eur Urol 2023;84:191–206. [PubMed: 37202314]
3. Mottet N, van den Bergh RCN, Briers E, et al. EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer—2020 update. Part 1: screening, diagnosis, and local treatment with curative intent. Eur Urol 2021;79:243–62. 10.1016/j.eururo.2020.09.042 [PubMed: 33172724]
4. Fazekas T, Shim SR, Basile G, et al. Magnetic resonance imaging in prostate cancer screening: a systematic review and meta-analysis. JAMA Oncol 2024;10:745–54. [PubMed: 38576242]
5. Ukimura O, Faber K, Gill IS. Intraprostatic targeting. Curr Opin Urol 2012;22:97. [PubMed: 22249373]
6. Bostwick DG. Evaluating prostate needle biopsy: therapeutic and prognostic importance. CA Cancer J Clin 1997;47:297–319. [PubMed: 9314824]
7. Isariyawongse BK, Sun L, Bañez LL, et al. Significant discrepancies between diagnostic and pathologic Gleason sums in prostate cancer: the predictive role of age and prostate-specific antigen. Urology 2008;72:882–6. [PubMed: 18384857]

8. Zattoni F, Rajwa P, Miszczyk M, et al. Transperineal versus transrectal magnetic resonance imaging–targeted prostate biopsy: a systematic review and meta-analysis of prospective studies. Eur Urol Oncol 2024;7:1303–12. [PubMed: 39095298]
9. Falagario UG, Jambor I, Lantz A, et al. Combined use of prostate-specific antigen density and magnetic resonance imaging for prostate biopsy decision planning: a retrospective multi-institutional study using the Prostate Magnetic Resonance Imaging Outcome Database (PROMOD). Eur Urol Oncol 2021;4:971–9. [PubMed: 32972896]
10. Radtke JP, Wiesenfarth M, Kesch C, et al. Combined clinical parameters and multiparametric magnetic resonance imaging for advanced risk modeling of prostate cancer—patient-tailored risk stratification can reduce unnecessary biopsies. Eur Urol 2017;72:888–96. [PubMed: 28400169]
11. Mehralivand S, Shih JH, Rais-Bahrami S, et al. A magnetic resonance imaging-based prediction model for prostate biopsy risk stratification. JAMA Oncol 2018;4:678–85. [PubMed: 29470570]
12. van Leeuwen PJ, Hayen A, Thompson JE, et al. A multiparametric magnetic resonance imaging-based risk model to determine the risk of significant prostate cancer prior to biopsy. BJU Int 2017;120:774–81. [PubMed: 28207981]
13. Rustam Z, Darmawan NA, Hartini S, Aurelia JE. Support vector machines and naïve Bayes classifier for classifying a prostate cancer. In: Kacprzyk J, Balas VE, Ezziyyani M, editors. Proceedings of Advanced Intelligent Systems for Sustainable Development (AI2SD 2020). Cham, Switzerland: Springer International Publishing; 2022. p. 854–60.
14. Xiao LH, Chen PR, Gou ZP, et al. Prostate cancer prediction using the random forest algorithm that takes into account transrectal ultrasound findings, age, and serum levels of prostate-specific antigen. Asian J Androl 2017;19:586–90. [PubMed: 27586028]
15. Chang CC, Chiou JK, Lin CJ, et al. Machine-learning algorithm-based risk prediction and screening-detected prostate cancer in a benign prostate hyperplasia cohort. Anticancer Res 2024;44:1683–93. [PubMed: 38537959]
16. Perera M, Smith L, Thompson I, et al. Advancing traditional prostate-specific antigen kinetics in the detection of prostate cancer: a machine learning model. Eur Urol Focus 2022;8:1204–10. [PubMed: 34920976]
17. Citak-Er F, Vural M, Acar O, Esen T, Onay A, Ozturk-Isik E. Final Gleason score prediction using discriminant analysis and support vector machine based on preoperative multiparametric MR imaging of prostate cancer at 3T. Biomed Res Int 2014;2014:690787. [PubMed: 25544944]
18. Finne P, Finne R, Auvinen A, et al. Predicting the outcome of prostate biopsy in screen-positive men by a multilayer perceptron network. Urology 2000;56:418–22. [PubMed: 10962306]
19. Natarajan S, Priester A, Margolis D, Huang J, Marks L. Prostate MRI and ultrasound with pathology and coordinates of tracked biopsy (Prostate-MRI-US-Biopsy) (version 2). Data set. Cancer Imaging Archive 2020. 10.7937/TCIA.2020.A61IOC1A
20. Sonn GA, Natarajan S, Margolis DJ, et al. Targeted biopsy in the detection of prostate cancer using an office based magnetic resonance ultrasound fusion device. J Urol 2013;189:86–92. [PubMed: 23158413]
21. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Dig Imaging 2013;26:1045–57.
22. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system. Am J Surg Pathol 2016;40:244–52. [PubMed: 26492179]
23. Barentsz JO, Weinreb JC, Verma S, et al. Synopsis of the PI-RADSs v2 guidelines for multiparametric prostate magnetic resonance imaging and recommendations for use. Eur Urol 2016;69:41–9. [PubMed: 26361169]
24. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45. [PubMed: 3203132]
25. Jeganathan T, Salgues E, Schick U, et al. Inter-rater variability of prostate lesion segmentation on multiparametric prostate MRI. Biomedicines. 2023;11:3309. [PubMed: 38137530]

26. Ankerst DP, Hoefler J, Bock S, et al. Prostate Cancer Prevention Trial risk calculator 2.0 for the prediction of low- versus high-grade prostate cancer. Urology 2014;83:1362–8. [PubMed: 24862395]
27. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: logistic regression. Perspect Clin Res 2017;8:148–51. [PubMed: 28828311]
28. Stoltzfus JC. Logistic regression: a brief primer. Acad Emerg Med 2011;18:1099–104. [PubMed: 21996075]
29. Klebaner D, Courtney PT, Garraway IP, et al. Association of health-care system with prostate cancer-specific mortality in African American and non-Hispanic white men. J Natl Cancer Inst 2021;113:1343–51. [PubMed: 33892497]
30. Presti JC. Prostate biopsy: current status and limitations. Rev Urol 2007;9:93–8. [PubMed: 17934565]

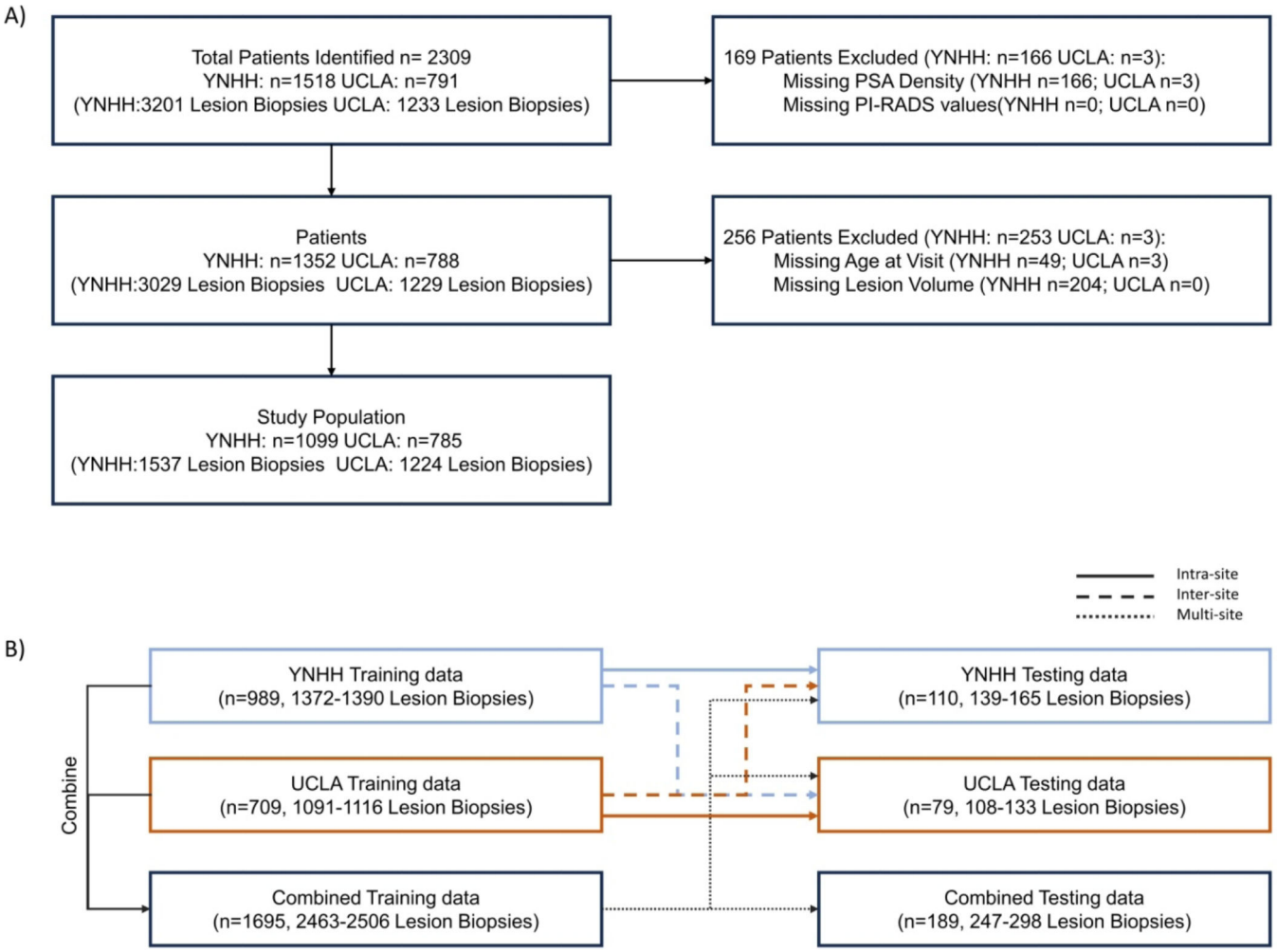
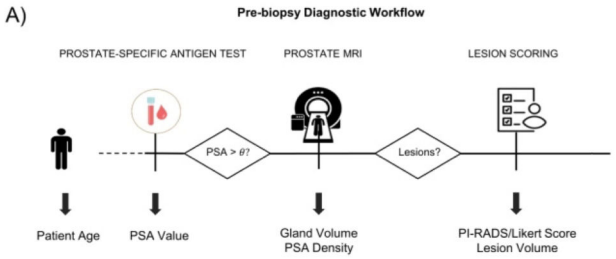


Fig. 1 –.

(A) Patient inclusion and exclusion criteria. (B) Outline of tenfold cross-validation for intrasite, intersite, and multisite training and testing. UCLA = University of California-Los Angeles; YNHH = Yale New Haven Hospital; PSA = prostate-specific antigen; PI-RADS = Prostate Imaging-Reporting and Data System.



Feature Group (FG)	Clinical Features
1	Patient Age
2	Patient Age, PSA Value
3	Patient Age, PSA Value, Gland Volume
4	Patient Age, PSA Value, Gland Volume, PSA Density
5	Patient Age, PSA Value, Gland Volume, PSA Density, PI-RADS/Likert Score
6	Patient Age, PSA Value, Gland Volume, PSA Density, PI-RADS/Likert Score, Lesion Volume
7	PSA Density
8	PSA Value, PSA Density

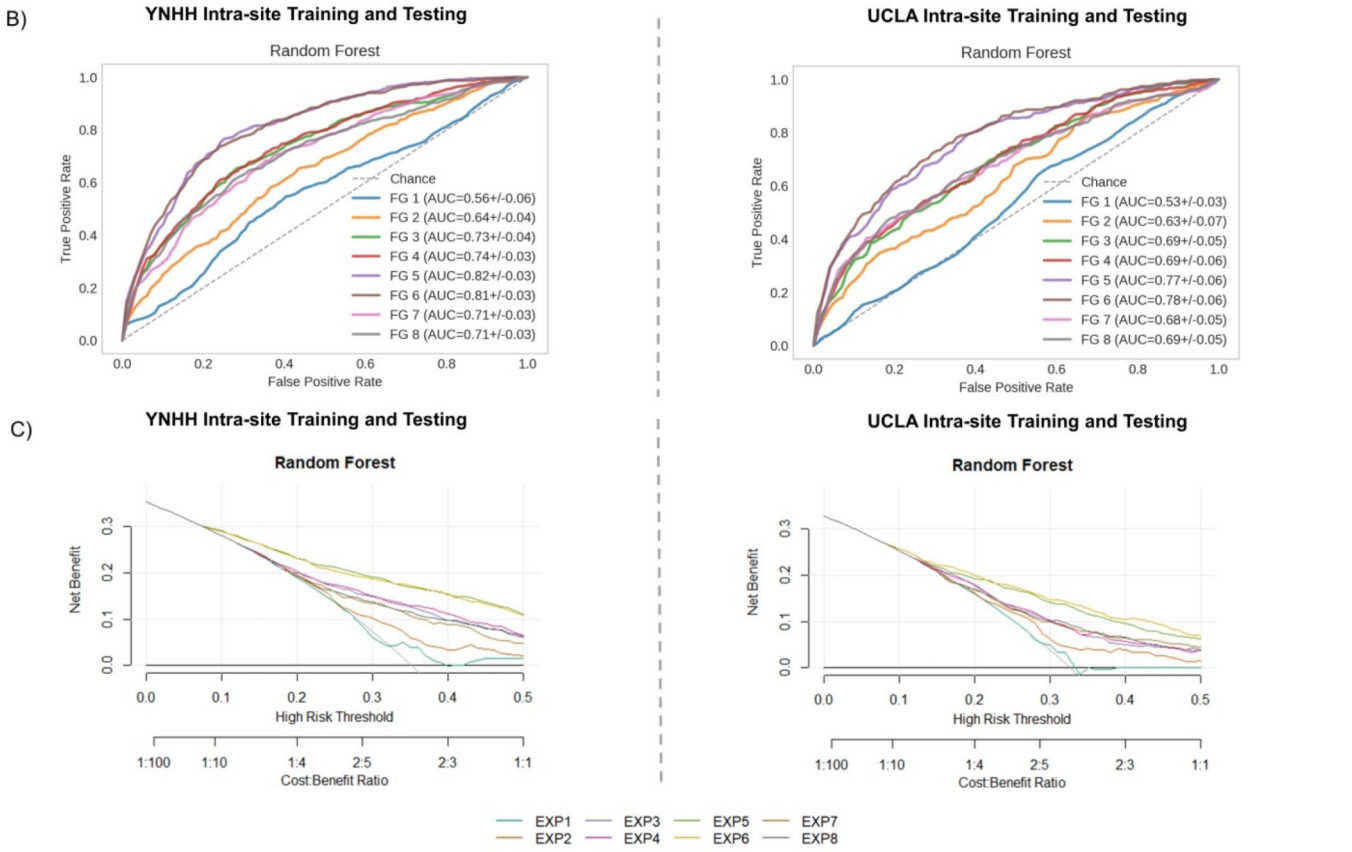


Fig. 2 –.

(A) Eight combinations of clinical features, organized by temporal availability along the prebiopsy diagnostic workflow, used to train the machine learning models. (B) Receiver operating characteristic (ROC) curves and (C) decision curve analysis for random forest models comprising eight clinical features that were trained and tested in the YNHH and UCLA cohorts via tenfold cross-validation. Supplementary Figure 1 shows results for all 12 models. AUC = area under the ROC curve; MRI = magnetic resonance imaging; UCLA = University of California-Los Angeles; YNHH = Yale New Haven Hospital.

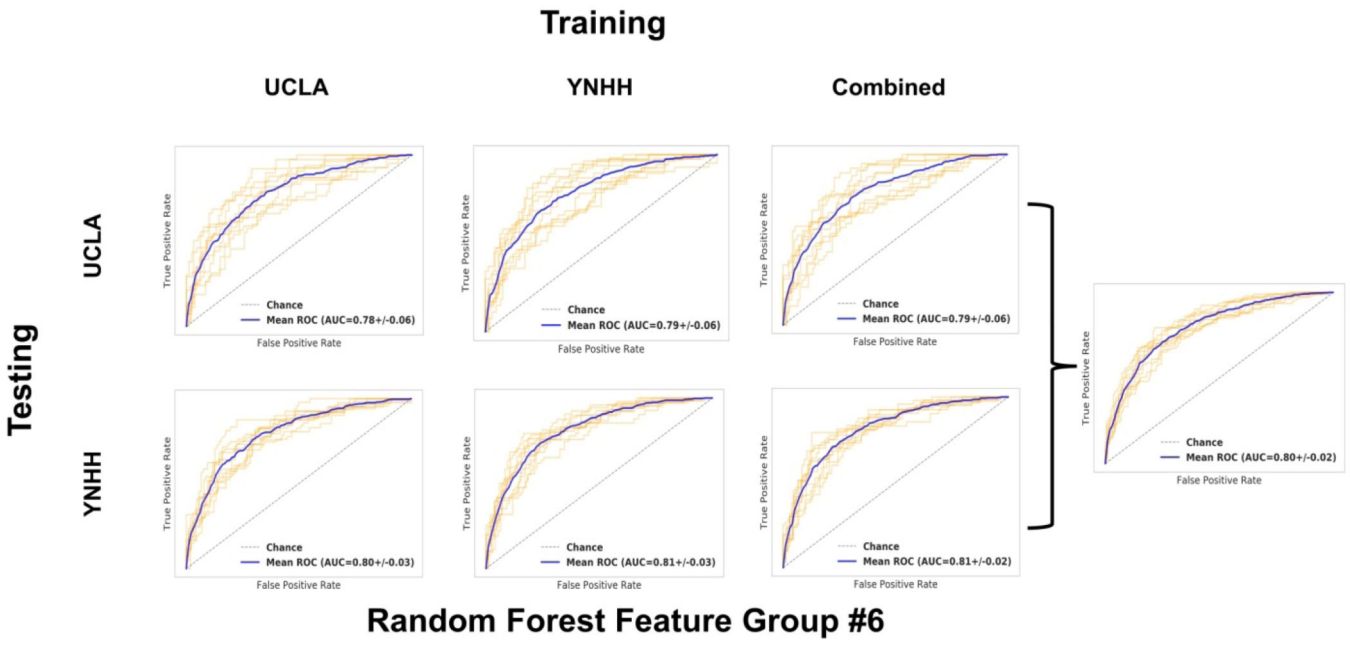


Fig. 3 –.

Comparison of receiver operating characteristic (ROC) curves for intrasite, intersite, and combined-site analyses for the random forest model. The 95% confidence interval (mean ± standard deviation) was calculated via tenfold cross-validation. Supplementary Figure 2 shows results for five selected models. AUC = area under the ROC curve; UCLA = University of California-Los Angeles; YNHH = Yale New Haven Hospital.

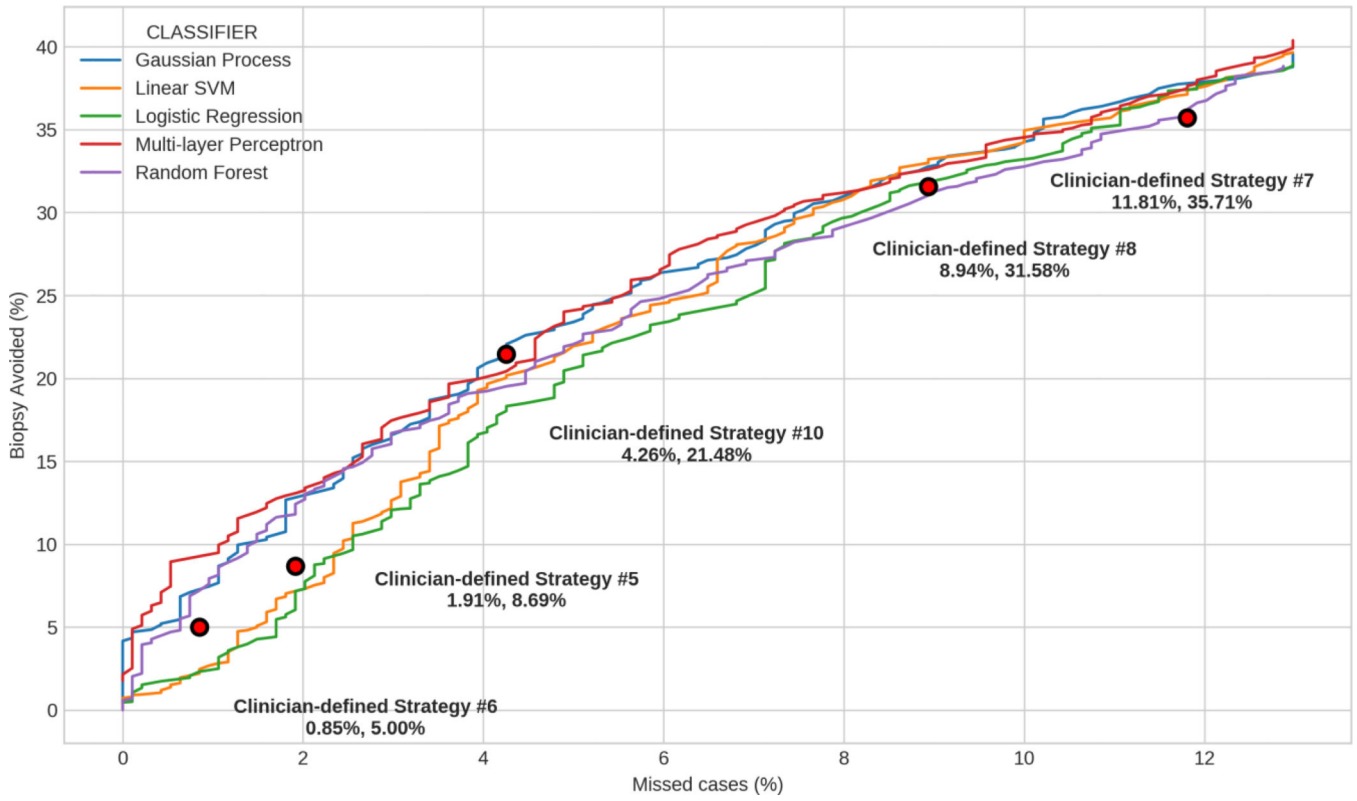


Fig. 4 –.

Proportion of case missed and biopsies avoided predicted by the machine learning models using the combined-site training and testing analysis with feature group 6. The results are compared to results derived using the clinician-defined strategies from Falagario et al. [9] (red circles), with biopsy indication as follows: strategy 5, PI-RADS 3–5 and/or PSA density >0.15 ng/ml/cm^3; strategy 6, PI-RADS 3–5 and/or PSA density >0.10 ng/ml/cm^3; strategy 7: PI-RADS 4–5 and/or PI-RADS 3 if PSA density >0.2 ng/ml/cm^3; strategy 8, PI-RADS 4–5 and/or PI-RADS 3 if PSA density >0.15 ng/ml/cm^3; strategy 10, PI-RADS 4–5 and/or PI-RADS 3 if PSA density >0.10 ng/ml/cm^3 and/or PSA density >0.2 ng/ml/cm^3. PI-RADS = Prostate Imaging-Reporting and Data System; PSA = prostate-specific antigen; SVM = support vector machine.

Table 1 –

Patient characteristics

Parameter	Overall cohort	UCLA cohort	YNHH cohort	*p* value
Patients (*n*)	1884	785	1099	
Median age, yr (IQR)	66.0 (60.4–71.0)	65.0 (61.0–71.0)	66.5 (60.4–71.7)	0.052 [a]
Median PSA, ng/ml (IQR)	6.9 (4.9–10.4)	6.9 (4.8–10.4)	6.9 (4.9–10.3)	0.996 [a]
Median PSA density, ng/ml^2 (IQR)	0.1 (0.1–0.2)	0.1 (0.1–0.2)	0.1 (0.1–0.2)	<0.001 [a]
Lesions (*n*)	2761	1224	1537	
PI-RADS/Likert score, *n* (%)				<0.001 [b]
<3	385 (13.9)	138 (11.3)	247 (16.1)	
≥3	2376 (86.1)	1086 (88.7)	1290 (83.9)	
Gleason grade group, *n* (%)				<0.001 [b]
0	1217 (44.1)	594 (48.5)	623 (40.5)	
1	604 (21.9)	230 (18.8)	374 (24.3)	
≥2	940 (34.0)	400 (32.7)	540 (35.1)	
Median lesion volume, ml (IQR)	0.4 (0.2–0.9)	0.3 (0.1–0.8)	0.4 (0.2–1.0)	<0.001 [a]
Median gland volume, ml (IQR)	46.8 (33.8–65.2)	42.0 (31.8–60.3)	50.0 (37.0–69.0)	<0.001 [a]

IQR = interquartile range; PSA = prostate-specific antigen; PI-RADS = Prostate Imaging-Reporting and Data System; UCLA = University of California-Los Angeles; YNHH = Yale New Haven Hospital.

[a]To assess data consistency, the data sets were compared using a Mann-Whitney U test for continuous variables and

[b]Pearson χ^2 test for categorical variables.