Development, multi-institutional external validation, and algorithmic audit of an artificial intelligence-based Side-specific Extra-Prostatic Extension Risk Assessment tool (SEPERA) for patients undergoing radical prostatectomy: a retrospective cohort study

Jethro C C Kwong, Adree Khondker, Eric Meng, Nicholas Taylor, Cynthia Kuk, Nathan Perlis, Girish S Kulkarni, Robert J Hamilton, Neil E Fleshner, Antonio Finelli, Theodorus H van der Kwast, Amna Ali, Munir Jamal, Frank Papanikolaou, Thomas Short, John R Srigley, Valentin Colinet, Alexandre Peltier, Romain Diamand, Yolene Lefebvre, Qusay Mandoorah, Rafael Sanchez-Salas, Petr Macek, Xavier Cathelineau, Martin Eklund, Alistair E W Johnson, Andrew Feifer, Alexandre R Zlotta**

Summary

Background Accurate prediction of side-specific extraprostatic extension (ssEPE) is essential for performing nerve-sparing surgery to mitigate treatment-related side-effects such as impotence and incontinence in patients with localised prostate cancer. Artificial intelligence (AI) might provide robust and personalised ssEPE predictions to better inform nerve-sparing strategy during radical prostatectomy. We aimed to develop, externally validate, and perform an algorithmic audit of an AI-based Side-specific Extra-Prostatic Extension Risk Assessment tool (SEPERA).

Methods Each prostatic lobe was treated as an individual case such that each patient contributed two cases to the overall cohort. SEPERA was trained on 1022 cases from a community hospital network (Trillium Health Partners; Mississauga, ON, Canada) between 2010 and 2020. Subsequently, SEPERA was externally validated on 3914 cases across three academic centres: Princess Margaret Cancer Centre (Toronto, ON, Canada) from 2008 to 2020; L'Institut Mutualiste Montsouris (Paris, France) from 2010 to 2020; and Jules Bordet Institute (Brussels, Belgium) from 2015 to 2020. Model performance was characterised by area under the receiver operating characteristic curve (AUROC), area under the precision recall curve (AUPRC), calibration, and net benefit. SEPERA was compared against contemporary nomograms (ie, Sayyid nomogram, Soeterik nomogram [non-MRI and MRI]), as well as a separate logistic regression model using the same variables included in SEPERA. An algorithmic audit was performed to assess model bias and identify common patient characteristics among predictive errors.

Findings Overall, 2468 patients comprising 4936 cases (ie, prostatic lobes) were included in this study. SEPERA was well calibrated and had the best performance across all validation cohorts (pooled AUROC of 0·77 [95% CI 0·75–0·78] and pooled AUPRC of 0·61 [0·58–0·63]). In patients with pathological ssEPE despite benign ipsilateral biopsies, SEPERA correctly predicted ssEPE in 72 (68%) of 106 cases compared with the other models (47 [44%] in the logistic regression model, none in the Sayyid model, 13 [12%] in the Soeterik non-MRI model, and five [5%] in the Soeterik MRI model). SEPERA had higher net benefit than the other models to predict ssEPE, enabling more patients to safely undergo nerve-sparing. In the algorithmic audit, no evidence of model bias was observed, with no significant difference in AUROC when stratified by race, biopsy year, age, biopsy type (systematic only *vs* systematic and MRI-targeted biopsy), biopsy location (academic *vs* community), and D'Amico risk group. According to the audit, the most common errors were false positives, particularly for older patients with high-risk disease. No aggressive tumours (ie, grade >2 or high-risk disease) were found among false negatives.

Interpretation We demonstrated the accuracy, safety, and generalisability of using SEPERA to personalise nerve-sparing approaches during radical prostatectomy.

Funding None.

Lancet Digit Health 2023; 5: e435–45

Published Online
May 19, 2023
https://doi.org/10.1016/S2589-7500(23)00067-5

*Joint senior authors

Division of Urology, Department of Surgery (J C C Kwong MD, N Perlis MD, Prof G S Kulkarni MD, R J Hamilton MD, Prof N E Fleshner MD, Prof A Finelli MD, M Jamal MD, F Papanikolaou MD, T Short MD, A Feifer MD, Prof A R Zlotta MD) **and Department of Laboratory Medicine and Pathobiology** (Prof T H van der Kwast MD, Prof J R Srigley MD)**, University of Toronto, Toronto, ON, Canada; Division of Urology, Department of Surgery, Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada** (J C C Kwong, N Perlis, Prof G S Kulkarni, R J Hamilton, Prof N E Fleshner, Prof A Finelli, Prof A R Zlotta)**; Temerty Centre for AI Research and Education in Medicine** (J C C Kwong, Prof G S Kulkarni, A E W Johnson DPhil) **and Temerty Faculty of Medicine** (A Khondker BHSc, N Taylor BSc)**, University of Toronto, Toronto, ON, Canada; Faculty of Medicine, Queen's University, Kingston, ON, Canada** (E Meng BSc)**; Division of Urology, Department of Surgery, Mount Sinai Hospital, Sinai Health System, Toronto, ON, Canada** (C Kuk MSc, Prof A R Zlotta)**; Laboratory Medicine Program, Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada**

Introduction

Accurate identification of extraprostatic extension (ie, tumours extending beyond the prostatic capsule) is an essential part of surgical planning and counselling in patients with localised prostate cancer. Preservation of the adjacent neurovascular bundles at the time of radical prostatectomy is associated with a lower risk of post-operative erectile dysfunction and urinary incontinence.[1]

(Prof T H van der Kwast); **Institute for Better Health, Trillium Health Partners, Mississauga, ON, Canada** (A Ali BSc, A Feifer); **Division of Urology, Department of Surgery** (V Colinet MD, Prof A Peltier MD, R Diamand MD) **and Department of Medical Imagery** (Y Lefebvre MD), **Jules Bordet Institute, Brussels, Belgium; Division of Urology, Department of Surgery, L'Institut Mutualiste Montsouris, Paris, France** (Q Mandoorah MD, R Sanchez-Salas MD, P Macek MD, Prof X Cathelineau MD); **Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden** (Prof M Eklund PhD); **Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada** (A E W Johnson); **Vector Institute, Toronto, ON, Canada** (A E W Johnson)

Correspondence to:
Prof Alexandre R Zlotta, Division of Urology, Department of Surgery, Mount Sinai Hospital, Sinai Health System, Toronto, ON M5T 3L9, Canada
alexandre.zlotta@sinaihealth.ca

Research in context

Evidence before this study

We searched PubMed, Embase, MEDLINE, and the Cochrane Library on Apr 24, 2022, for available studies to date. Search terms were "extraprostatic extension", "extracapsular extension", "machine learning", and "artificial intelligence". No language restrictions were applied. A total of 18 studies investigated the application of artificial intelligence (AI) in determining the risk of extraprostatic extension. The most common approach was the use of radiomics on prostate MRIs. Area under the receiver operating characteristic curve of existing AI models ranged from 0·68 to 0·88 for overall risk and 0·72 to 0·81 for side-specific extraprostatic extension (ssEPE). However, 14 (78%) of 18 studies lacked external validation, 16 (89%) lacked an assessment of model bias, and 13 (72%) lacked sufficient description of their model development or hyperparameter tuning. Only ten (56%) of 18 studies included a reference standard for performance comparisons with their AI models. Only one (6%) of 18 studies provided a sample size calculation, which was our previous work describing our initial experience in developing an AI model to predict ssEPE. Collectively, these issues raise concerns about the overall safety and generalisability of these AI models in real-world clinical practice.

Added value of this study

Our study applied current best practices in AI to thoroughly investigate the accuracy, generalisability, and safety of our AI model (ie, Side-specific Extra-Prostatic Extension Risk Assessment tool [SEPERA]) to predict the risk of ssEPE. Using an algorithmic audit, we verified the safety of SEPERA by demonstrating that it was less prone to bias than existing nomograms and that no aggressive tumours were missed by our model. Additionally, we highlighted the clinical utility of SEPERA by presenting challenging clinical scenarios (ie, predicting contralateral ssEPE in unilateral high-risk disease or ssEPE in the context of benign ipsilateral biopsies) in which our model outperformed existing nomograms.

Implications of all the available evidence

SEPERA could be used to help inform surgical planning and patient counselling for patients with localised prostate cancer. Accurate predictions of ssEPE by SEPERA might help personalise surgical approach in nerve-sparing, manage postoperative expectations in keeping with optimal oncological control, and potentially minimise postoperative functional decline.

However, nerve-sparing, especially in patients with a high risk of extraprostatic extension, can increase the likelihood of positive surgical margins, cancer recurrence, and subsequently worse oncological outcomes.[2] To help guide decision making of when to safely perform nerve-sparing surgery, several nomograms have been developed to predict either overall or side-specific extraprostatic extension (ssEPE) risk. More recently, the adoption of prostate MRI in the preoperative workup has led to the inclusion of additional MRI information in several models, although with mixed results.[3]

In recent years, artificial intelligence (AI) has emerged as a promising tool to provide accurate and individualised risk estimates in medicine and urology.[4–6] However, AI models to date have been fraught with methodological issues including an absence of standardised reporting to enhance reproducibility and comparability, limited external validation to evaluate generalisability, poor evaluation of model bias to identify suboptimal performance in clinically relevant subgroups, and an absence of in-depth investigations of predictive errors to understand model behaviour.[7–9] These drawbacks have raised concerns regarding the safety and applicability of AI models in real-world clinical practice.[10]

See **Online** for appendix

To address these limitations in this study, we aimed to comprehensively assemble current AI best practices to develop SEPERA (Side-specific Extra-Prostatic Extension Risk Assessment tool). Firstly, we developed SEPERA using a standardised reporting framework designed for AI studies in urology. Secondly, we externally validated SEPERA on adequately powered, multi-institutional cohorts and compared its performance against contemporary nomograms using clinically relevant metrics. Finally, we performed an algorithmic audit to assess model fairness and to identify common characteristics among predictive errors to recognise their implications if SEPERA is used in clinical practice.[10]

Methods

Study design and problem

This study was done in accordance with the STREAM-URO framework, a standardised reporting framework we have previously developed for AI studies in urology.[7] This project is a supervised binary classification problem to determine the risk of ssEPE using available clinical, pathological, and MRI reported information. Each prostatic lobe was treated as an individual case such that each patient contributed two cases to the overall cohort.

A total of 611 cases (ie, prostatic lobes) with 184 ssEPE cases were required to satisfy the criteria outlined by Riley and colleagues,[11] assuming a 30% incidence of ssEPE and a maximum of 12 features (variables) included in SEPERA (appendix p 2).

This study was approved by the Research Ethics Board at the University Health Network, Canada (research ethics board number 20-6038). The need for consent was waived as only de-identified, retrospective data were used, which would not affect standard of care for these patients.

Data sources and eligibility criteria

The training cohort comprised 1022 cases treated at Trillium Health Partners (Mississauga, Canada) between 2010 and 2020. This data source comes from the largest community-based hospital system in Canada, which includes two community hospitals, and was selected as the training cohort to reflect the average risk profile of patients with prostate cancer to improve generalisability.

The validation cohorts included a total of 3914 cases across three academic centres: Princess Margaret Cancer Centre, University Health Network (Toronto, ON, Canada), from 2008 to 2020 (validation cohort 1; 2300 cases); L'Institut Mutualiste Montsouris (Paris, France), from 2010 to 2020 (validation cohort 2; 1352 cases); and Jules Bordet Institute (Brussels, Belgium), from 2015 to 2020 (validation cohort 3; 262 cases). Since validation cohort 1 is a tertiary referral centre for robotic-assisted radical prostatectomy, this cohort was further subdivided based on where the biopsy was done (at the University Health Network *vs* at a community hospital).

Patients were included regardless of preoperative imaging (MRI *vs* no MRI), biopsy method (systematic only *vs* systematic and MRI-targeted biopsy), or surgical approach (open *vs* robotic-assisted radical prostatectomy). For all cases, systematic and MRI-targeted biopsy cores were assigned to three standardised sites (base, middle, and apex). Since multiple cores were often taken at each site, the core with the highest International Society of Urological Pathology (ISUP) grade—ie, the classification system used to grade the aggressiveness of prostate cancer—at each site was recorded. All MRIs were assessed by dedicated uroradiologists at each institution.

Patients were excluded if they underwent transperineal biopsy or previously received radiotherapy or androgen deprivation therapy. Patients who did not have any site-specific biopsy information or available pathology reports for the prostatectomy specimen (ie, unknown outcome) were excluded (appendix p 14).

Data abstraction, processing, and outcome definition

Data were manually extracted from the electronic medical record using a standardised data form, as previously described.[12] No discrete imputation methods were used for missing data in SEPERA, as it was capable of handling missing data automatically. Multiple imputation was used for missing data for the other nomograms (appendix p 3).

The outcome of interest (label) was the presence of ssEPE, defined as tumour that has extended beyond the ipsilateral prostatic capsule in the radical prostatectomy specimen (pT3a disease). All prostatectomy specimens were reviewed by dedicated uropathologists at each institution. A data dictionary describing all features and labels is included in the appendix (pp 4–5).

SEPERA development and explanations

SEPERA was developed using XGBoost (version 1.5.0). We trained this gradient-boosted ensemble machine learning model that sequentially builds decision trees such that each subsequent tree reduces the misclassification error of previous trees.[13] XGBoost manages missing data by using a sparsity-aware split finding algorithm to automatically determine the best imputation value for missing data. Stratified ten-fold cross-validation (based on presence of ssEPE) of the training cohort was used for hyperparameter tuning and feature selection using mean area under the receiver operating characteristic curve (AUROC) as the optimisation metric. Additional information about the hyperparameter search space is provided in the appendix (p 6). Different combinations of features were selected for model training based on clinical judgement. The model was then retrained on the entire training cohort using the final set of hyperparameters and features that yielded the highest mean AUROC. The so-called black-box nature of SEPERA was interrogated using SHapley Additive exPlanations (SHAP) to help understand which features were most important overall and the individual effect of each feature on the probability of ssEPE.[14]

Reference standards

SEPERA was compared against contemporary clinicopathological-based and MRI-based models that have previously been externally validated. The Sayyid nomogram is a biopsy-derived model developed based on patients treated at the University Health Network (Canada) from 2009 to 2015.[15] This nomogram predicts ssEPE using age, prostate-specific antigen (PSA), prostate volume, palpable nodule on digital rectal examination, hypoechoic nodule on transrectal ultrasound, percentage of positive cores, maximum core involvement, and highest ISUP grade.[15] On external validation, the Sayyid nomogram has demonstrated an AUROC of 0·74–0·75.[12]

The Soeterik nomogram refers to several clinicopathological-based and MRI-based models developed from patient data collected at the Canisius Wilhelmina Hospital (Nijmegen, Netherlands) from 2014 to 2018.[16] Their clinicopathological model (ie, model 1, herein referred to as the Soeterik non-MRI model) predicts ssEPE based on PSA density (PSA concentration divided by prostate volume), digital rectal examination local staging, highest ISUP grade, and percentage of positive cores, whereas their best MRI-based model (ie, model 2, herein referred to as the Soeterik MRI model) uses PSA density, multiparametric MRI-based local staging, and highest ISUP grade. On external validation, the non-MRI model had an AUROC of 0·77–0·80 and the MRI model had an AUROC of 0·77–0·83.[16,17]

Finally, a separate logistic regression model was developed using the same features included in SEPERA to determine the iterative improvements provided by an AI-based approach (appendix p 7).

Model evaluation

Model performance was characterised by discrimination, calibration, and clinical utility across the training and

	Training cohort	Validation 1 cohort	Validation 2 cohort	Validation 3 cohort	p value
Clinical features					
Number of patients	511	1150	676	131	··
Number of prostatic lobes	1022	2300	1352	262	··
Age at biopsy	62 (57–66)	62 (58–67)	71 (66–73)	66 (61–71)	<0·0001
Race					<0·0001
White	686 (67·1%)	1802 (78·3%)	1290 (95·4%)	216 (82·4%)	··
Black	76 (7·4%)	114 (5·0%)	30 (2·2%)	32 (12·2%)	··
Hispanic	96 (9·4%)	190 (8·3%)	14 (1·0%)	14 (5·3%)	··
Asian	164 (16·0%)	194 (8·4%)	18 (1·3%)	0	··
Prostate volume (mL)	34 (26–44)	36 (28–46)	40 (30–50)	35 (28–51)	<0·0001
Serum PSA (ng/mL)	7·1 (5·5–9·5)	6·6 (5·0–9·3)	7·7 (5·8–11·4)	7·8 (6·0–11·3)	<0·0001
PSA density (ng/mL^2)	0·21 (0·15–0·31)	0·18 (0·12–0·27)	0·19 (0·13–0·28)	0·23 (0·13–0·32)	<0·0001
Clinical stage					<0·0001
T1	642 (62·8%)	1536 (66·8%)	1310 (96·9%)	184 (70·2%)	··
T2	368 (36·0%)	673 (29·3%)	20 (1·5%)	70 (26·7%)	··
T3	12 (1·2%)	26 (1·1%)	20 (1·5%)	6 (2·3%)	··
Unknown	0	65 (2·8%)	2 (0·1%)	2 (0·8%)	··
D’Amico risk group					<0·0001
Low	250 (24·5%)	172 (7·5%)	219 (16·2%)	64 (24·4%)	··
Intermediate	600 (58·7%)	1862 (81·0%)	970 (71·7%)	124 (47·3%)	··
High	172 (16·8%)	266 (11·6%)	163 (12·1%)	74 (28·2%)	··
Global biopsy features					
Percentage Gleason pattern 4 or 5	12·5 (5·0–55·0)	10·0 (5·0–40·0)	10·0 (0·0–40·0)	20·0 (0·0–63·8)	<0·0001
Perineural invasion	462 (45·2%)	938 (40·8%)	348 (25·7%)	95 (36·3%)	<0·0001
Periprostatic fat invasion	12 (1·2%)	26 (1·1%)	19 (1·4%)	6 (2·3%)	0·38
Side-specific features (ie, left or right prostatic lobe)					
Abnormal digital rectal examination	219 (21·4%)	382 (16·6%)	12 (0·9%)	35 (13·4%)	<0·0001
Percentage of positive cores	33·3 (16·7–66·7)	33·3 (16·7–66·7)	33·3 (14·3–61·2)	37·5 (0·0–62·5)	0·61
Highest ISUP grade					<0·0001
Benign	221 (21·6%)	490 (21·3%)	315 (23·3%)	62 (23·7%)	··
1	179 (17·5%)	596 (25·9%)	397 (29·4%)	70 (26·7%)	··
2	370 (36·2%)	789 (34·3%)	445 (32·9%)	52 (19·8%)	··
3	145 (14·2%)	256 (11·1%)	145 (10·7%)	22 (8·4%)	··
4	61 (6·0%)	119 (5·2%)	43 (3·2%)	33 (12·6%)	··
5	46 (4·5%)	47 (2·0%)	7 (0·5%)	3 (1·1%)	··
Unknown	0	3 (0·1%)	0	20 (7·6%)	··
Maximum percentage core involvement	20 (5–50)	20 (5–55)	20 (3–44)	37 (0–64)	0·0086
Base findings					<0·0001
Benign	531 (52·0%)	1115 (48·5%)	672 (49·7%)	135 (51·5%)	··
1	143 (14·0%)	465 (20·2%)	253 (18·7%)	49 (18·7%)	··
2	207 (20·3%)	482 (21·0%)	303 (22·4%)	24 (9·2%)	··
3	73 (7·1%)	133 (5·8%)	95 (7·0%)	15 (5·7%)	··
4	34 (3·3%)	66 (2·9%)	23 (1·7%)	19 (7·3%)	··
5	34 (3·3%)	32 (1·4%)	5 (0·4%)	0	··
Unknown	0	7 (0·3%)	1 (0·1%)	20 (7·6%)	··
Base percentage core involvement	14·5 (24·3)	17·3 (26·3)	15·6 (23·2)	11·5 (23·0)	<0·0001

(Table 1 continues on next page)

validation cohorts. Discriminative performance was determined by AUROC and area under the precision recall curve (AUPRC). The latter compares sensitivity (recall) and positive predictive values (precision) across various thresholds. Differences in AUROC and AUPRC between models were tested using 10 000 bootstrap samples with replacement, and results were presented with 95% CIs. Calibration was determined by comparing the predicted and observed risk of ssEPE by deciles. Clinical utility was assessed by decision curve analysis,

(Continued from previous page)

	Training cohort	Validation 1 cohort	Validation 2 cohort	Validation 3 cohort	p value
Middle finding					<0·0001
Benign	450 (44·0%)	977 (42·5%)	610 (45·1%)	120 (45·8%)	..
1	182 (17·8%)	513 (22·3%)	301 (22·3%)	57 (21·8%)	..
2	255 (25·0%)	449 (19·5%)	320 (23·7%)	38 (14·5%)	..
3	87 (8·5%)	144 (6·3%)	87 (6·4%)	11 (4·2%)	..
4	26 (2·5%)	68 (3·0%)	29 (2·1%)	13 (5·0%)	..
5	22 (2·2%)	25 (1·1%)	4 (0·3%)	3 (1·1%)	..
Unknown	0	124 (5·4%)	1 (0·1%)	20 (7·6%)	..
Middle percentage core involvement	16·5 (24·1)	18·0 (25·9)	16·0 (22·8)	12·3 (23·0)	<0·0001
Apex findings					<0·0001
Benign	581 (56·8%)	1078 (46·9%)	791 (58·5%)	152 (58·0%)	..
1	162 (15·9%)	535 (23·3%)	225 (16·6%)	48 (18·3%)	..
2	168 (16·4%)	372 (16·2%)	235 (17·4%)	21 (8·0%)	..
3	70 (6·8%)	118 (5·1%)	71 (5·3%)	12 (4·6%)	..
4	25 (2·4%)	41 (1·8%)	23 (1·7%)	9 (3·4%)	..
5	16 (1·6%)	21 (0·9%)	6 (0·4%)	0	..
Unknown	0	135 (5·9%)	1 (0·1%)	20 (7·6%)	..
Apex percentage core involvement	14·4 (24·3)	15·3 (24·0)	12·6 (21·5)	7·5 (18·6)	<0·0001
MRI findings					<0·0001
Normal	33 (3·2%)	41 (1·8%)	545 (40·3%)	114 (43·5%)	..
Lesion but no ssEPE	15 (1·5%)	69 (3·0%)	724 (53·6%)	122 (46·6%)	..
ssEPE	2 (0·2%)	8 (0·3%)	83 (6·1%)	26 (9·9%)	..
Preoperative MRI not performed	972 (95·1%)	2182 (94·9%)	0	0	..
ssEPE in the final prostatectomy specimen	327 (32·0%)	660 (28·7%)	375 (27·7%)	47 (17·9%)	0·0001
Benign ipsilateral biopsy but ssEPE	34 (3·3%)	50 (2·2%)	50 (3·7%)	6 (2·3%)	0·037

Data are n, median (IQR), n (%), or mean (SD), unless otherwise specified. PSA density is determined by dividing PSA concentration by prostate volume. D'Amico risk group is a classification system to determine overall aggressiveness of prostate cancer based on PSA concentration, clinical stage, and prostate biopsy grade. The ISUP grade is a standardised grading system for prostate cancer based on histopathological features. Statistical significance for numerical variables were determined by Kruskal-Wallis test or ANOVA, while categorical variables were determined by χ^2 test. All p values were based on whether the features for any cohort differed significantly from the others (ie, training *vs* validation 1 *vs* validation 2 *vs* validation 3). ISUP=International Society of Urological Pathology. PSA=prostate-specific antigen. ssEPE=side-specific extraprostatic extension.

Table 1: **Baseline characteristics at the prostatic lobe level for each cohort**

which measures the net benefit of each model. Net benefit is a weighted combination of true and false positives, where the weight is derived from the threshold probability at which a clinical decision is made (in this case, performing nerve-sparing).[18]

Alternative model development strategies

Additional models were developed to assess the effect of different training cohorts and feature sets on model performance. Firstly, a separate XGBoost model (named Academic SEPERA) was trained on an academic cohort (University Health Network) with the same variables as the final SEPERA model to examine the effect of model training using this dataset. This model was externally validated on the remaining cohorts. Secondly, another XGBoost model (named MRI SEPERA) was trained on the Paris cohort (L'Institut Mutualiste Montsouris). This model also included worst Prostate Imaging Reporting and Data System (PI-RADS) score and MRI findings (ie, normal, lesion but no ssEPE, or ssEPE) to investigate the effect of incorporating MRI-specific features on model performance. MRI SEPERA was externally validated on patients from the remaining cohorts who underwent preoperative MRI. The Trillium Health Partners cohort was excluded in the MRI SEPERA analysis since PI-RADS score was not reported, although the presence of lesions and ssEPE on MRI was mentioned.

Algorithmic audit

An algorithmic audit is a systematic approach previously described by Liu and colleagues to recognise and understand algorithmic errors or inaccurate predictions by AI models.[10] The purpose of this audit is to identify deviations from expected performance to determine the overall safety of SEPERA. We focused on the testing stage of the audit, which includes patient-specific and task-specific subgroup analysis, as well as exploratory error analysis. Subgroup analysis was done by comparing AUROCs across clinically relevant subgroups to ensure that SEPERA was not biased against specific patient populations. The minimum p value for all pairwise comparisons in each subgroup was reported. Patient-specific subgroups included age group, race, and biopsy year. Task-specific subgroups included location of biopsy (academic *vs* community, for validation cohort 1), biopsy method (systematic only *vs* systematic and MRI-targeted, for patients with pre-biopsy MRI), and D'Amico risk group.[19] Exploratory error analysis of SEPERA was done by comparing baseline characteristics of correct predictions, false negatives, and false positives on the validation cohorts to determine specific clinicopathological features that might be more prone to inaccurate predictions. For this analysis, SEPERA was set to a threshold probability to target a sensitivity of 95% in the training cohort.

Role of the funding source

There was no funding source for this study.

Results

Overall, 2468 patients comprising 4936 cases (ie, prostatic lobes) were included in this study, with the prevalence of ssEPE ranging from about 18% to 32% (depending on cohort). The baseline characteristics, according to cohort of origin, are summarised in table 1. These cohorts were diverse and significant differences were observed for almost all features. In total, 140 (2·8%) of 4936 cases had pathological ssEPE despite benign ipsilateral biopsies.

Use of preoperative MRI varied across institutions, ranging from 5% to 100%. Of patients who received preoperative MRI (1782 [36·1%] of 4936 cases), only 64 (13·7%) of 468 cases with pathological ssEPE were suspected to have ssEPE on MRI (sensitivity of 14%). By contrast, 1259 (95·8%) of 1314 cases with organ-confined disease did not demonstrate any evidence of ssEPE on imaging (specificity of 96%; appendix p 8). Overall, ssEPE seen on preoperative MRI corresponded to a positive predictive value of 54% for pathological ssEPE.

SEPERA was trained on 11 (55%) of 20 candidate features (age, PSA density, highest ISUP grade, perineural invasion, percentage of positive cores, percentage of Gleason pattern 4 or 5, maximum percentage core involvement, base finding, base percentage core involvement, middle percentage core involvement, and apex percentage core involvement). Model explanations using SHAP are provided to illustrate which features are most important overall (appendix p 15) and the individual effect of each feature on the probability of ssEPE (appendix p 16). The five most important features are base percentage core involvement, maximum percentage core involvement, perineural invasion, base finding, and percentage of Gleason pattern 4 or 5.

SEPERA performed favourably compared with existing nomograms across all cohorts (table 2). Similarly, SEPERA outperformed the logistic regression model trained on the same features included in SEPERA (appendix p 17). On the ten-fold stratified cross-validation of the training cohort, SEPERA achieved an AUROC of 0·80 (95% CI 0·77–0·82) and AUPRC of 0·69 (0·63–0·72). For the validation cohorts, SEPERA achieved a pooled AUROC of 0·77 (0·75–0·78) and AUPRC of 0·61 (0·58–0·63). SEPERA was also better calibrated on the validation cohorts than the other nomograms (figure 1). Individual calibration curves for each validation cohort are provided in the appendix (p 18).

Regarding alternative model development strategies, Academic SEPERA achieved a pooled AUROC of 0·75 (95% CI 0·73–0·77) and AUPRC of 0·59 (0·55–0·62), with individual validation cohort AUROCs ranging from 0·73 to 0·79 (appendix p 10). MRI SEPERA achieved a pooled AUROC of 0·72 (0·65–0·79) and

	SEPERA	Logistic regression	Sayyid	Soeterik non-MRI	Soeterik MRI
AUROC					
Training cohort	0·80 (0·77–0·82)*	0·80 (0·77–0·82)*	0·77† (0·74–0·79)	0·74‡ (0·71–0·77)	0·70‡ (0·66–0·73)
Validation 1 cohort	0·78 (0·76–0·79)*	0·76‡ (0·74–0·79)	0·77‡ (0·75–0·79)	0·74‡ (0·72–0·76)	0·70‡ (0·67–0·72)
Validation 2 cohort	0·75 (0·73–0·78)*	0·75 (0·72–0·78)*	0·71‡ (0·67–0·74)	0·69‡ (0·66–0·73)	0·69‡ (0·66–0·72)
Validation 3 cohort	0·77 (0·71–0·82)*	0·76 (0·68–0·83)	0·76 (0·67–0·84)	0·72† (0·63–0·80)	0·71† (0·62–0·79)
Combined validation cohort	0·77 (0·75–0·78)*	0·75‡ (0·74–0·77)	0·75‡ (0·73–0·76)	0·72‡ (0·70–0·74)	0·69‡ (0·68–0·71)
AUPRC					
Training cohort	0·69 (0·63–0·72)*	0·68 (0·63–0·72)	0·65† (0·61–0·70)	0·63‡ (0·58–0·67)	0·53‡ (0·47–0·59)
Validation 1 cohort	0·64 (0·61–0·67)*	0·62‡ (0·59–0·65)	0·63‡ (0·59–0·66)	0·57‡ (0·54–0·61)	0·48‡ (0·44–0·51)
Validation 2 cohort	0·57 (0·52–0·62)*	0·55† (0·50–0·59)	0·51‡ (0·47–0·56)	0·50‡ (0·45–0·54)	0·48‡ (0·44–0·52)
Validation 3 cohort	0·47 (0·34–0·59)*	0·46 (0·32–0·61)	0·42 (0·30–0·55)	0·36 (0·26–0·48)	0·35 (0·24–0·47)
Combined validation cohort	0·61 (0·58–0·63)*	0·58‡ (0·55–0·61)	0·57‡ (0·54–0·60)	0·53‡ (0·50–0·55)	0·46‡ (0·44–0·49)

Data are AUROC (95% CI) or AUPRC (95% CI). Performance metrics for the training cohort were determined based on stratified ten-fold cross validation. Statistically significant differences between SEPERA and existing nomograms are shown. All 95% CIs and p values were determined using 10 000 bootstrap samples with replacement. AUROC=area under the receiver operating-characteristic curve. AUPRC=area under the precision-recall curve. SEPERA=Side-specific Extra-Prostatic Extension Risk Assessment tool. *Best performing models for each cohort. †p<0·05. ‡p<0·01.

Table 2: **Discriminative performance of all models based on AUROC and AUPRC**

AUPRC of 0·41 (0·31–0·54), with individual validation cohort AUROCs ranging from 0·70 to 0·74 (appendix p 11). Given these results, the original SEPERA model was selected as the final model for further analysis.

Using the prespecified threshold probability to target a sensitivity of 95% in the training cohort, SEPERA achieved a sensitivity of 93% in the combined validation cohort. Of 106 cases in the validation cohorts with pathological ssEPE despite benign ipsilateral biopsies, ssEPE was correctly predicted in 72 (68%; SEPERA), 47 (44%; logistic regression), 0 (0%; Sayyid), 13 (12%; Soeterik non-MRI), and five (5%; Soeterik MRI) cases.

Decision curve analysis was done to help contextualise the benefits of implementing SEPERA in clinical practice. Overall, SEPERA had a higher net benefit for clinically relevant thresholds between 15% and 30%, which means more patients would safely undergo a nerve-sparing approach if SEPERA was used instead of other nomograms (figure 2). Individual net benefit curves for each validation cohort are provided in the appendix (p 19). Using a threshold probability of 20%, seven (logistic regression), 23 (Sayyid model), 52 (Soeterik non-MRI model), and 75 (Soeterik MRI model) more patients would correctly receive a nerve-sparing prostatectomy per 1000 cases if SEPERA was used instead (appendix p 9).

Patient-specific and task-specific subgroup analysis was performed to evaluate model fairness across clinically relevant subgroups. SEPERA achieved comparable AUROCs across all age groups (0·77 for <55 years, 0·79 for 56–65 years, 0·76 for 66–75 years, and 0·73 for >75 years; p>0·086) and races (0·77 for White, 0·74 for Black, 0·80 for Hispanic, and 0·78 for Asian; p>0·19; figure 3). Furthermore, its performance remained stable across biopsy periods (0·78 for 2010 and before, 0·74 for 2011–12, 0·80 for 2013–14, 0·77 for 2015–16, 0·77 for 2017–18, and 0·76 for 2019–20; p>0·13). In subgroup analyses of validation cohort 1, patients who underwent prostate biopsy at a community site (989 [43%] of 2300 cases) had equivalent AUROCs to those biopsied at the University Health Network (0·77 *vs* 0·78; p=0·76). For patients who underwent preoperative MRI, no difference in AUROC was observed between patients who underwent both systematic and MRI-targeted biopsy (607 [34%] of 1782 cases) or systematic biopsy alone (0·74 *vs* 0·74; p=0·91). Similarly, no difference in AUROC was observed when stratified by D'Amico risk group (0·74 for low risk, 0·75 for intermediate risk, and 0·79 for high risk; p>0·051).

On exploratory error analysis of the combined validation cohort, we found that 96% of errors were false positives. These cases tended to be older patients with higher PSA concentrations, PSA density, percentage of Gleason pattern 4 or 5 disease, and either intermediate-risk or high-risk disease (appendix p 12). Among the false negatives, no cases with an ISUP grade of more than 2 were included.

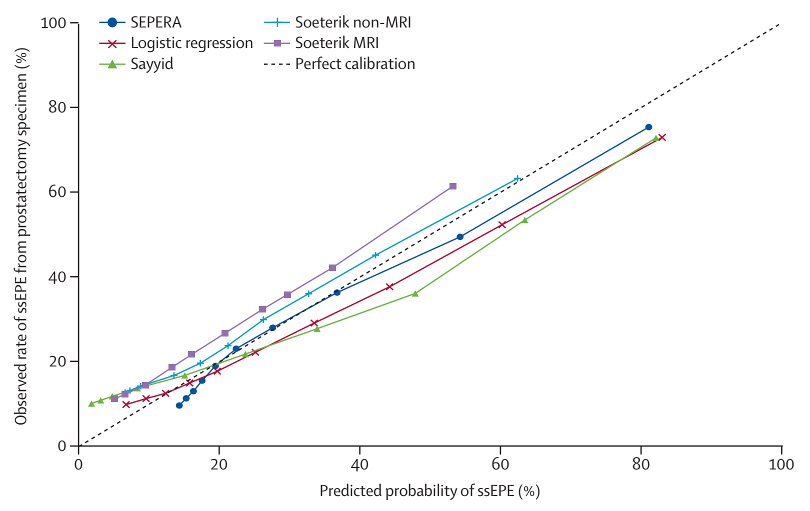

Figure 1: **Calibration of all models on the combined validation cohorts by measuring the degree of agreement between the predicted and observed risk of ssEPE by deciles**

A perfectly calibrated model corresponds to a 45-degree line. ssEPE=side-specific extraprostatic extension. SEPERA=Side-specific Extra-Prostatic Extension Risk Assessment tool.

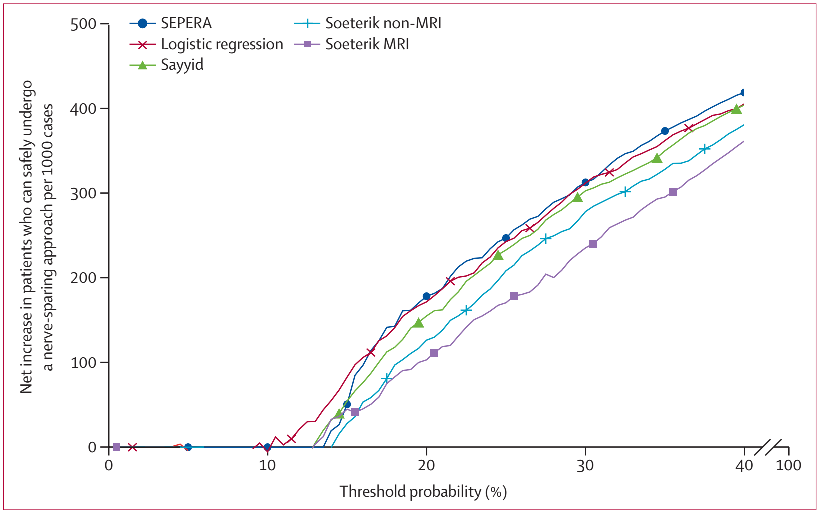

Figure 2: **Net benefit (ie, potential clinical impact) for the combined validation cohort**

Examination was done by comparing the number of patients who can safely undergo nerve-sparing per 1000 cases compared with a so-called non-nerve sparing approach for all strategy. The threshold probability refers to the probability at which nerve-sparing would not be recommended if the model predicted a probability of ssEPE greater than or equal to the threshold (ie, if the threshold probability was set at 20%, then a nerve-sparing approach would not be recommended if the model predicted a ≥20% probability of ssEPE). ssEPE=side-specific extraprostatic extension. SEPERA=Side-specific Extra-Prostatic Extension Risk Assessment tool.

Discussion

Accurate prediction of ssEPE preoperatively in patients undergoing radical prostatectomy is essential to tailor nerve-sparing strategy at the time of surgery. New AI tools should be clinician-friendly and evaluated against existing clinical and logistic regression models.[7] In our study, we have developed and externally validated SEPERA—an explainable AI model to predict ssEPE—on large, multi-institutional cohorts. SEPERA performs favourably, demonstrates better calibration, and provides

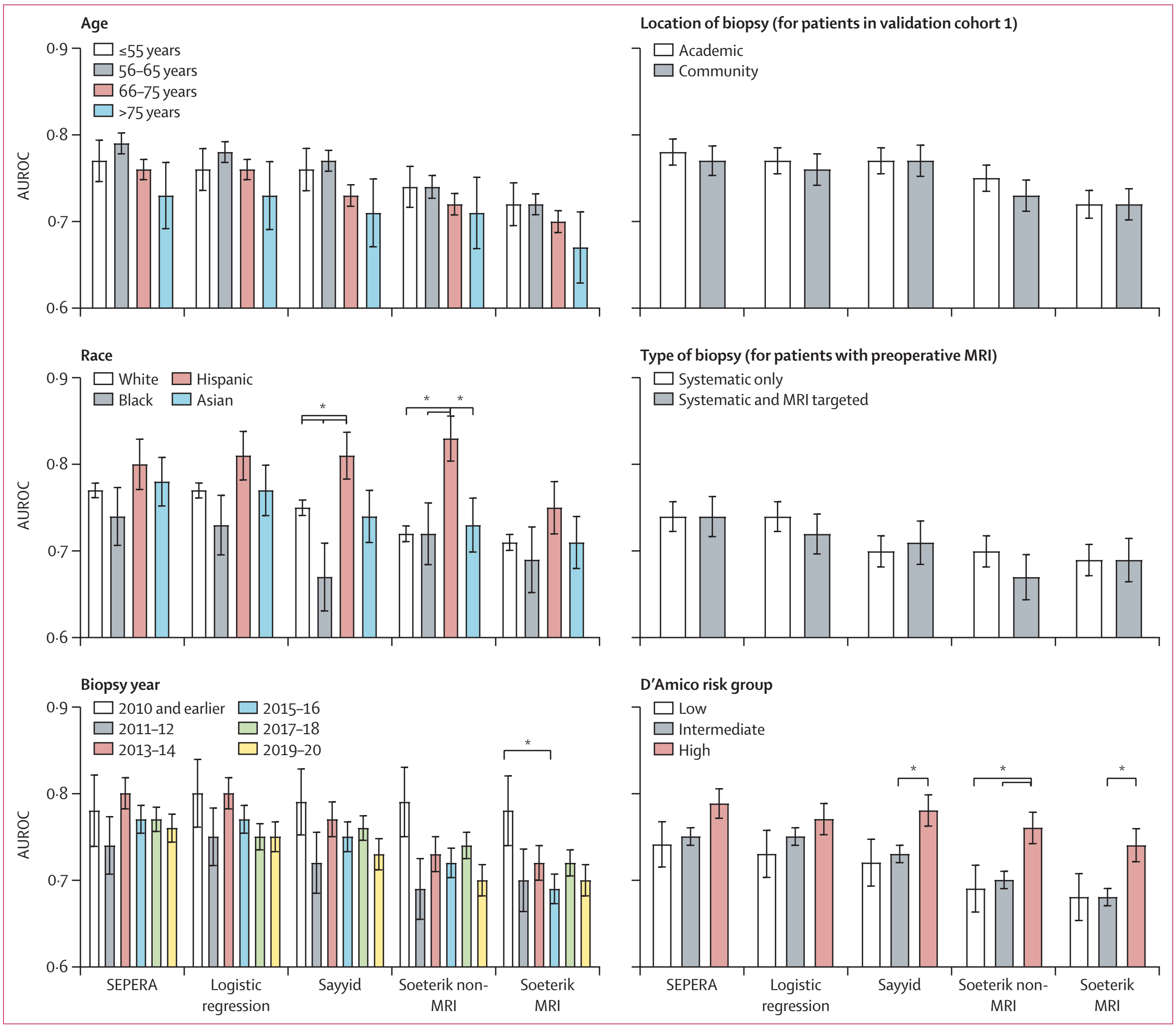

Figure 3: **Bias assessment of all models by comparing AUROC across clinically relevant subgroups**

Error bars are 95% CIs. AUROC=area under the receiver operating characteristic curve. SEPERA=Side-specific Extra-Prostatic Extension Risk Assessment tool. *Significant differences (p<0·05).

greater net benefit than contemporary nomograms and the logistic regression model, which suggests that AI approaches might provide some additional performance benefit when applied to the appropriate clinical context. SEPERA is an important step towards integrating AI into clinical practice at the time of patient counselling to better inform clinical decision making and operative planning.

SEPERA might also be useful in selecting appropriate candidates for contralateral nerve-sparing in the setting of unilateral high-risk prostate cancer, which is a challenging clinical scenario without clear guideline recommendations. Martini and colleagues[20] recently developed a χ^2 automated interaction detection (CHAID) model to predict contralateral ssEPE in patients with unilateral high-risk disease. In a model of 705 patients with unilateral high-risk disease, including 88 (12%) with contralateral ssEPE, their model achieved an AUROC of 0·72 and showed higher net benefit compared with the Soeterik MRI model. From our study cohort, 736 patients met the same eligibility criteria, including 315 (43%) with ssEPE. In this subgroup, SEPERA achieved an AUROC of 0·74, compared with 0·73 (logistic regression), 0·74 (Sayyid), 0·72 (Soeterik non-MRI), and 0·69 (Soeterik MRI; data not shown). However, a direct comparison with the CHAID model was not possible since data for MRI index lesion diameter was not available in our dataset.

To the best of our knowledge, our study is the first to apply an algorithmic audit to an AI model in prostate cancer to characterise its risks and biases, which are two important but understudied metrics to evaluate the safety and fairness of prediction models.[10,21] SEPERA performed well despite the heterogeneity in baseline characteristics among the validation cohorts. Furthermore, no differences in performance were observed when stratified by age group, race, year of biopsy, whether MRI findings were used to perform MRI-targeted biopsies, and D'Amico risk group. SEPERA also performed equally well regardless of whether patients underwent their biopsy at academic or community hospitals. These findings provide strong evidence for the generalisability of this model regardless of practice setting.

Predictive errors were also investigated with the algorithmic audit to understand model behaviour. We found that sensitivity remained stable on external validation datasets at the prespecified threshold probability (93% compared with 95% on training data). Most errors were false positives, particularly for older patients with high-risk disease. However, older age and preoperative erectile function might have a greater effect on potency, irrespective of nerve-sparing technique. In a retrospective study of 3126 patients who received bilateral nerve-sparing, erections sufficient for intercourse declined with age (70·3% for those <65 years, 54·0% for those 65–70 years, 49·0% for those 70–75 years, 37·5% for those >75 years; p<0·001).[22] Furthermore, patient preference regarding nerve-sparing might change with older age. Lavery and colleagues[23] showed that when the decision for nerve-sparing was deferred to patient choice, those with lower risks of ssEPE who declined nerve-sparing tended to be older. This finding suggests that although nerve-sparing is an individual choice, this decision might be less relevant in older patients, as maintaining postoperative potency, potentially at the cost of margin status, might be less of a concern. However, no cases with high-grade disease (ISUP grade >2) were included in false negatives. This observation is consistent with a previous systematic review that found that biochemical recurrence, a surrogate marker for prostate cancer progression, was significantly higher in patients with ISUP grade 3–5 disease at positive surgical margins compared with ISUP grade 1–2.[24] The false-positive rate of SEPERA should also be placed into clinical context. Our previous study and others found that when patients were asked to weigh their oncological risk against their potency, they chose to decline nerve-sparing when their risk of ssEPE is between 15% and 30%, which corresponds to accepting a false-positive rate of at least 70%.[12,16] Taken together, we demonstrated the safety of SEPERA by highlighting that the most common errors were not deemed clinically significant.

The generalisability of SEPERA might be attributed to the use of a multi-institutional community-based cohort for model training. By contrast, other nomograms included in this study were derived from single tertiary referral centres. These cohorts might be less representative of the overall prostate cancer population because of the inclusion of high-risk and complex patients in these academic centres. Indeed, retraining SEPERA on an academic cohort (University Health Network) resulted in a greater variation in performance across external validation sets (AUROC 0·73–0·79) compared with training on a community cohort (Trillium Health Partners). These findings align with recent work by Ötleş and colleagues[25] that showed that models trained on regional cohorts outperform those developed using data from tertiary care centres to predict adverse pathological outcomes in patients that have undergone prostatectomy.[25]

In recent years, incorporation of MRI findings into newer ssEPE models has become more common, with the most robust model to date being the Soeterik MRI model (AUROC 0·77–0·83).[16,17] However, this model underperformed in all metrics in our study. We found that MRI had a low sensitivity but high specificity in detecting ssEPE. Similarly, Soeterik and colleagues[16] showed that MRI had a sensitivity of 37%, specificity of 93%, and positive predictive value of 59%. These findings are consistent with a previous meta-analysis that reported a pooled sensitivity of 57%.[26] Although wide variability in MRI use existed across the study cohorts, we also retrained SEPERA on validation cohort 2 (L'Institut Mutualiste Montsouris, Paris) to include both worst PI-RADS score and MRI findings. This MRI SEPERA model performed poorly on patients in the other cohorts with preoperative MRI available (AUROC 0·70–0·74). Overall, use of MRI alone is inadequate to detect ssEPE, particularly focal ssEPE (few extraprostatic cancer glands on 1–2 slides[27]). Interpretation of prostate MRIs is also strongly influenced by radiological expertise, which might limit its applicability in non-tertiary settings. Therefore, although MRI remains an integral part of the pre-prostatectomy workup, its role in guiding nerve-sparing strategy might be limited and cannot replace the use of clinicopathological features.

Our study has limitations that merit discussion. First, although SEPERA achieved an AUROC of 0·77 on the combined validation cohort and is a step forward compared with existing tools, room for further improvement exists. Further studies are also needed to determine whether the incremental performance benefits provided by SEPERA translates to clinically meaningful improvements in oncological and functional outcomes. Second, MRI use was variable and MRI features included in our dataset were limited, which might impact SEPERA's utility in current practice. This use of simple MRI features was done for several reasons: to standardise data capture across multiple institutions over a long study period, to ensure consistency of MRI features with the Soeterik MRI model, and to decrease model complexity to improve applicability. With the widespread use of MRI in prostate cancer assessments, future models would benefit from incorporating more granular MRI features such as tumour capsular contact length, irregular or spiculated margins, bulging prostatic contour, or even deep-learning analysis of the MRI images to improve predictive accuracy.[28,29] Third, our study lacked standardisation of the type of biopsy (ie, systematic only *vs* systemic and MRI targeted), which provider performed the biopsy (ie, urologists *vs* radiologists), and central histopathological review, which might affect the quality of the biopsy itself and pathological reporting. These factors probably lead to more conservative performance estimates. Despite this limitation, we demonstrated that SEPERA performed well across multiple clinically relevant subgroups, including type of biopsy. Furthermore, SEPERA outperformed existing nomograms across several diverse validation cohorts, which strengthens its use in real-world clinical settings. Fourth, inclusion of MRI-targeted biopsies might influence the value of percentage positive cores because of additional sampling of suspicious regions. However, we specifically did a bias assessment to examine this issue and found that among patients who received preoperative MRI, no difference was observed in performance between those who underwent systematic biopsy only versus systematic and MRI-targeted biopsy. These results suggest that heterogeneity in biopsy approaches did not affect performance of SEPERA. Fifth, we did not specify location of ssEPE (ie, base, middle, or apex) or distinguish between focal versus established ssEPE, and whether the former correlates with positive surgical margins is unclear. Sixth, our study only included patients who had transrectal prostate biopsies, thus its performance on patients undergoing transperineal biopsy remains unknown. Because of differences in directionality of the biopsy needle, regions sampled transperineally are different from those sampled through the transrectal approach. As transperineal biopsies are increasingly being adopted worldwide, we are currently developing AI models using both approaches to investigate any performance differences compared with SEPERA, which was trained on exclusively transrectal biopsies. Seventh, our data for race are imperfect as they were either abstracted from clinical notes or determined based on first and last names (appendix p 4). However, the evaluation of racial bias in predictive models is a major unmet need in the medical literature.[30] Therefore, to ensure that SEPERA did not systematically disadvantage against specific patient populations, we believe that it was clinically important to report an estimation of model performance when stratified by race. Finally, our overall cohort was predominantly White, which is comparable to other national cancer databases (appendix p 13). However, we attempted to address this limitation by training SEPERA on the most racially diverse cohort of the group (Trillium Health Partners). Additional work is being done to further evaluate SEPERA on a more diverse group of patients.

For the **fully trained model** see https://doi.org/10.5281/zenodo.7945217

In summary, we developed and externally validated an AI-based Side-specific Extra-Prostatic Extension Risk Assessment tool (SEPERA). This study is also the first to compare the fairness and safety of a prostate cancer model with existing ssEPE models using an algorithmic audit. We showed that SEPERA performs favourably and is less prone to bias compared with existing nomograms on diverse multi-institutional cohorts. Furthermore, we demonstrated that MRI has poor sensitivity in detecting ssEPE and overreliance on this imaging modality might miss focal ssEPE cases. Taken together, we provide strong evidence to support the use of SEPERA to personalise a side-specific nerve-sparing approach during radical prostatectomy. Future work to understand the model's effect on clinical decision making and patient outcomes in real-world practice are underway. Additionally, we plan to apply this model on patients who underwent transperineal biopsies, as this approach is increasingly used to minimise complications compared with the conventional transrectal route.

Contributors

JCCK contributed to the conceptualisation, data curation, formal analysis, investigation, methodology, software, validation, visualisation, writing the original draft, and review and editing of the manuscript. AK, EM, and NT contributed to the data curation and investigation. VC and QM contributed to the investigation. NP, GSK, RJH, NEF, AFi, THvdK, MJ, FP, TS, JRS, AP, RD, YL, RS-S, PM, XC, and ME contributed to the review and editing of the manuscript. CK and AA contributed to the project administration. AEWJ contributed to the methodology, software, and supervision. AFe contributed to the conceptualisation, review and editing of the manuscript, and supervision. ARZ contributed to the conceptualisation, methodology, review and editing of the manuscript, and supervision. JCCK, AK, EM, NT, VC, and QM directly accessed and verified the raw data, and no authors were precluded from accessing the data. AFe and ARZ contributed equally to this work and share senior authorship. All authors have read, share final responsibility for the decision to submit for publication, and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Declaration of interests

We declare no competing interests.

Data sharing

The data dictionary used for this study is available in the appendix (pp 4–5). The final and executable version of the fully trained model, using the final hyperparameters (listed in the appendix p 6) and retrained on both the training and validation cohorts, is available online. The data are not publicly available, since all research or research-related activities that involve an external party might require, at the discretion of the University Health Network (Toronto, Canada), a written research agreement in order to define the obligations and manage the risks.

Acknowledgments

JCCK is supported by the University of Toronto Surgeon Scientist Training Program.

References

1 Nguyen LN, Head L, Witiuk K, et al. The risks and benefits of cavernous neurovascular bundle sparing during radical prostatectomy: a systematic review and meta-analysis. *J Urol* 2017; **198:** 760–69.

2 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.

3 Soeterik TFW, van Melick HHE, Dijksman LM, et al. External validation of the Martini nomogram for prediction of side-specific extraprostatic extension of prostate cancer in patients undergoing robot-assisted radical prostatectomy. *Urol Oncol Semin Orig Investig* 2020; **38:** 372–78.

4 Chen J, Remulla D, Nguyen JH, et al. Current status of artificial intelligence applications in urology and their potential to influence clinical practice. *BJU Int* 2019; **124:** 567–77.

5 Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit Health* 2021; **3:** e195–203.

6 Kwon JM, Cho Y, Jeon KH, et al. A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. *Lancet Digit Health* 2020; **2:** e358–67.

7 Kwong JCC, McLoughlin LC, Haider M, et al. Standardized reporting of machine learning applications in urology: the STREAM-URO framework. *Eur Urol Focus* 2021; **7:** 672–82.

8 Dhiman P, Ma J, Andaur Navarro CL, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med Res Methodol* 2022; **22:** 101.

9 Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Heal* 2021; **3:** e745–50.

10 Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. *Lancet Digit Health* 2022; **4:** e384–97.

11 Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II—binary and time-to-event outcomes. *Stat Med* 2019; **38:** 1276–96.

12 Kwong JCC, Khondker A, Tran C, et al. Explainable artificial intelligence to predict the risk of side-specific extraprostatic extension in pre-prostatectomy patients. *Can Urol Assoc J* 2022; **16:** 213–21.

13 Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2016: 785–94.

14 Lundberg SM, Allen PG, Lee SI. A unified approach to interpreting model predictions. 31st Conference on Neural Information Processing Systems; 2017.

15 Sayyid R, Perlis N, Ahmad A, et al. Development and external validation of a biopsy-derived nomogram to predict risk of ipsilateral extraprostatic extension. *BJU Int* 2017; **120:** 76–82.

16 Soeterik TFW, van Melick HHE, Dijksman LM, et al. Development and external validation of a novel nomogram to predict side-specific extraprostatic extension in patients with prostate cancer undergoing radical prostatectomy. *Eur Urol Oncol* 2022; **5:** 328–37.

17 Veerman H, Heymans MW, van der Poel HG. External validation of a prediction model for side-specific extraprostatic extension of prostate cancer at robot-assisted radical prostatectomy. *Eur Urol Open Sci* 2022; **37:** 50–52.

18 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak* 2006; **26:** 565–74.

19 D'Amico AV, Whittington R, Malkowicz SB, et al. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *JAMA* 1998; **280:** 969–74.

20 Martini A, Soeterik TFW, Haverdings H, et al. An algorithm to personalize nerve sparing in men with unilateral high-risk prostate cancer. *J Urol* 2022; **207:** 350–57.

21 Wawira Gichoya J, McCoy LG, Celi LA, Ghassemi M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Heal Care Informatics* 2021; **28:** e100289.

22 Mandel P, Graefen M, Michl U, Huland H, Tilki D. The effect of age on functional outcomes after radical prostatectomy. *Urol Oncol* 2015; **33:** 203.

23 Lavery HJ, Prall DN, Abaza R. Active patient decision making regarding nerve sparing during radical prostatectomy: a novel approach. *J Urol* 2011; **186:** 487–92.

24 John A, John H, Catterwell R, Selth LA, Callaghan MO. Primary Gleason grade and Gleason grade group at positive surgical margins: a systematic review and meta-analysis. *BJU Int* 2021; **127** (suppl)**:** 13–22.

25 Ötleş E, Denton BT, Qu B, et al. Development and validation of models to predict pathological outcomes of radical prostatectomy in regional and national cohorts. *J Urol* 2022; **207:** 358–66.

26 de Rooij M, Hamoen EHJ, Witjes JA, Barentsz JO, Rovers MM. Accuracy of magnetic resonance imaging for local staging of prostate cancer: a diagnostic meta-analysis. *Eur Urol* 2016; **70:** 233–45.

27 Epstein JI, Carmichael MJ, Pizov G, Walsh PC. Influence of capsular penetration on progression following radical prostatectomy: a study of 196 cases with long-term followup. *J Urol* 1993; **150:** 135–41.

28 Wibmer AG, Kattan MW, Alessandrino F, et al. International multi-site initiative to develop an MRI-inclusive nomogram for side-specific prediction of extraprostatic extension of prostate cancer. *Cancers* 2021; **13:** 2627.

29 Gatti M, Faletti R, Gentile F, et al. mEPE-score: a comprehensive grading system for predicting pathologic extraprostatic extension of prostate cancer at multiparametric magnetic resonance imaging. *Eur Radiol* 2022; **32:** 4942–53.

30 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; **366:** 447–53.