Explainable artificial intelligence to predict the risk of side-specific extraprostatic extension in pre-prostatectomy patients

Jethro C.C. Kwong[1,2]; Adree Khondker[3]; Christopher Tran[3]; Emily Evans[3]; Adrian I. Cozma[4]; Ashkan Javidan[3]; Amna Ali[5]; Munir Jamal[1]; Thomas Short[1]; Frank Papanikolaou[1]; John R. Srigley[6]; Benjamin Fine[5,7,8]; Andrew Feifer[1,5]

[1]Division of Urology, Department of Surgery, University of Toronto, Toronto, ON, Canada; [2]Temerty Centre for AI Research and Education in Medicine, University of Toronto, Toronto, ON, Canada; [3]Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada; [4]Department of Radiation Oncology, University of Toronto, Toronto, ON, Canada; [5]Institute for Better Health, Trillium Health Partners, Mississauga, ON, Canada; [6]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada; [7]Operational Analytics Lab, Trillium Health Partners, Mississauga, ON, Canada; [8]Department of Medical Imaging, University of Toronto, Toronto, ON, Canada

Cite as: 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 January 27; Epub ahead of print. http://dx.doi.org/10.5489/cuaj.7473

Published online January 27, 2022

Corresponding author: Dr. Andrew Feifer, Credit Valley Hospital, Trillium Health Partners, Mississauga, ON, Canada; andrew.feifer@thp.ca

***

Abstract

Introduction: We aimed to develop an explainable machine learning (ML) model to predict side-specific extraprostatic extension (ssEPE) to identify patients who can safely undergo nerve-sparing radical prostatectomy using preoperative clinicopathological variables.
Methods: A retrospective sample of clinicopathological data from 900 prostatic lobes at our institution was used as the training cohort. Primary outcome was the presence of ssEPE. The baseline model for comparison had the highest performance out of current biopsy-derived predictive models for ssEPE. A separate logistic regression (LR) model was built using the same variables as the ML model. All models were externally validated using a testing cohort of 122 lobes from another institution. Models were assessed by area under receiver operating characteristic (AUROC), precision-recall (AUPRC), calibration, and decision curve analysis. Model predictions were explained using Shapley Additive exPlanations. This tool was deployed as a publicly available web application.

Results: Incidence of ssEPE in the training and testing cohorts were 30.7 and 41.8%, respectively. The ML model achieved AUROC 0.81 (LR 0.78, baseline 0.74) and AUPRC 0.69 (LR 0.64, baseline 0.59) on the training cohort. On the testing cohort, the ML model achieved AUROC 0.81 (LR 0.76, baseline 0.75) and AUPRC 0.78 (LR 0.75, baseline 0.70). The ML model was explainable, well-calibrated, and achieved the highest net benefit for clinically relevant cutoffs of 10–30%.
Conclusions: We developed a user-friendly application that enables physicians without prior ML experience to assess ssEPE risk and understand factors driving these predictions to aid surgical planning and patient counselling (*www.ssepe.ml*).

Introduction

Accurate identification of side-specific extraprostatic extension (ssEPE) in pre-prostatectomy patients is essential to ensure a balance between optimal oncological and functional outcomes. Ipsilateral nerve-sparing during radical prostatectomy (RP) is associated with a lower risk of post-operative urinary incontinence and erectile dysfunction[1]. However, nerve-sparing should be considered with caution in patients with an increased risk of ssEPE due to the potential for positive surgical margins[2]. Given the clinical significance of correctly identifying patients with ssEPE, several predictive models have been developed using pre-operative variables such as prostate-specific antigen (PSA), clinical stage, side-specific worst Gleason score, percent positive biopsy cores, and tumour involvement[3].

Additional variables such as percentage of high-grade disease, perineural invasion, and site-specific findings are readily available from biopsy reports but underutilized[4]. With the emergence of artificial intelligence in healthcare, we hypothesize that use of machine learning (ML) methods to incorporate the complete clinicopathological profile may provide robust and personalized ssEPE predictions. This is especially important in resource-limited environments without access to routine pre-operative magnetic resonance imaging (MRI) to potentially improve ssEPE detection[5]. We postulated whether more predictive power can be gained using what is already available.

Furthermore, although ML models have historically been limited due to poor explainability, several approaches have recently been developed under the umbrella of "explainable artificial intelligence" (XAI)[6]. These explainable models not only determine the probability of the outcome, but also highlight the variables that drive these predictions. This helps build trust in the model, by ensuring predictions and explanations are aligned with clinical intuition[7]. XAI has successfully been implemented to better understand hypoxemia risk during anesthesia[8]. To this end, we set out to apply XAI to improve the diagnostic accuracy and our understanding of ssEPE.

Methods

This study was conducted in accordance with the Standardized Reporting of Machine Learning Applications in Urology (STREAM-URO) framework[9]. Herein, features and labels will be used instead of input variables and outcomes, respectively, in accordance with ML terminology. This is a supervised, binary classification problem, in which both features and corresponding labels are known and the ML model is trained to predict the label using available features.

Sample size calculation

Using a significance level of 0.05, we determined that 561 cases (ie: prostatic lobes) were sufficient to provide 80% power to detect a 10% difference in area under the receiver-operating-characteristic curve (AUROC) between our ML model and the reference standard[10], based on a 30% incidence of ssEPE reported in the literature[3].

Data sources

A retrospective cohort of 507 patients (1014 cases) who underwent RP at Credit Valley Hospital, Ontario, between 2010-2020, was used as the training cohort. An external cohort of 99 patients (198 cases) who underwent RP at Mississauga Hospital, Ontario, between 2016-2020, was used as the testing cohort.

Eligibility criteria

Patients were included regardless of type of radical prostatectomy (open or robotic-assisted). All patients underwent transrectal ultrasound (TRUS) guided prostate biopsy, which were performed by radiologists. Biopsy cores were taken from each of 4 standardized sites (base, mid, apex, transition zone). All pathological specimens were reviewed by an expert uropathologist.

Patients were excluded if they previously received radiotherapy or androgen deprivation therapy. Patients with missing data were excluded to avoid the use of synthetic data in the ML model.

Feature and label data collection

Clinical data and all findings available on prostate biopsy reports were collected to investigate whether additional global or site-specific features could improve ssEPE prediction (25 candidate features). Gleason Grade Group was determined based on the grading system by Epstein *et al*[11]. Data was manually extracted from the electronic medical record using a standardized form to ensure consistency. Discrepancies in data extraction were resolved by consensus and a focused second review of the dataset was performed to ensure accuracy and consistency of the data.

The label of interest was the presence of ssEPE in the prostatectomy specimen, defined as tumour that has extended beyond the prostatic capsule in the ipsilateral lobe. A data dictionary

describing each feature and label is provided in Supplementary Table 1. Additional data preparation steps are outlined in Supplementary Figure 1.

Model selection and training

As explainability of the ML model is clinically important, we selected a tree-based model (XGBoost version 1.3.3)[12] for model training as it is less prone to overfitting and is optimized for SHapley Additive exPlanations (SHAP)[13], described in *Model explanations*. This is an ensemble ML model that sequentially builds up decision trees such that each subsequent tree minimizes the predictive errors of prior trees. Stratified tenfold cross-validation was used for model training and hyperparameter tuning using mean AUROC as the scoring metric. Additional details regarding hyperparameter tuning and stratified tenfold cross-validation are provided in Supplementary Table 2.

Reference standard

The model by Sayyid *et al*. was selected as the reference standard, herein referred to as the baseline model[3]. It is a multivariable logistic regression model based on age, PSA, prostate volume, palpable nodule on DRE, hypoechoic nodule on TRUS, side-specific percent positive cores, maximum core involvement, and worst Gleason Grade Group. This model has the highest performance out of current biopsy-derived predictive models for ssEPE that have been externally validated (AUROC = 0.74).

A separate multivariable logistic regression model, herein referred to as the LR model, was built using the same features included in the ML model to measure the effect of model selection on performance.

Model evaluation

Discrimination was assessed by AUROC and area under the precision-recall curve (AUPRC). AUPRC compares sensitivity (recall) and positive predictive values (precision) across various decision thresholds and is more informative than AUROC when evaluating classification performance of imbalanced datasets, such as in this case where there are more negative than positive ssEPE cases. Calibration curves were used to evaluate the accuracy of model risk estimates.

Clinical utility

Clinical utility was determined by decision curve analysis, in which the net benefit is plotted against various decision thresholds for three different treatment strategies: treat all, treat none, and treat only those predicted to have ssEPE by the models[14]. Net benefit of each model was used to calculate the gain in appropriate ipsilateral nerve-sparing per 100 cases compared to a "treat all" strategy.

Model explanations

SHAP provides a unified framework to understanding individual model predictions by fitting a unique linear model to our ML model and calculating corresponding feature weights. The final probability of ssEPE is the sum of these values[13]. Several implementations of SHAP (version 0.39.0) were applied to provide insight into the "black box" ML model. Feature importance rankings were used to identify features with the greatest overall impact on model predictions. Partial dependence plots were used to visualize how a given feature can impact the probability of ssEPE across all values (ie: how does % Gleason pattern 4/5, from 0 to 100%, positively or negatively impact probability of ssEPE?).

Results

Cohort characteristics

After removing cases with missing data, the final training and testing cohorts comprised on 900 and 122 cases, respectively (Figure 1). The clinicopathological characteristics of the study population are summarized in Table 1, with some differences were observed between the two cohorts. The incidence of ssEPE in the training and testing cohorts were 30.7 and 41.8%, respectively. Median time from biopsy to surgery was 3.2 months (IQR: 2.3 to 4.4 months). A comparison of the baseline characteristics between this study cohort and that of Sayyid *et al*. is provided in Supplementary Table 3.

Model specification

From the initial panel of 25 features, 13 features were selected using Boruta feature selection (Supplementary Figure 1A). Two features were removed due high collinearity (Supplementary Figure 1B). The final ML model was trained on 11 features including age, PSA, worst Gleason Grade Group, % Gleason pattern 4/5, perineural invasion, % positive cores, maximum % core involvement, base findings, base % core involvement, mid % core involvement, and transition zone % core involvement. Additional model and hyperparameter specifications are listed in Supplementary Table 2.

Model evaluation

The performance metrics of all models are summarized in Table 2. The baseline model performed comparably to its reported AUROC of 0.74[3]. The ML model achieved the highest AUROC and AUPRC in both cohorts. The ML model was well-calibrated for predicted probabilities between 0-40%, while overestimating risk of ssEPE above 40% probability in the testing cohort (Figure 2A). When stratified based on age, institution, and D'Amico risk classification, the ML model had the highest performance and demonstrated fairness across all subgroups (Supplementary Table 4).

Clinical utility

Threshold probabilities between 10-30% were deemed the most clinically relevant for consideration of nerve-sparing[5]. At this range, all models had similar sensitivities, however the ML model generally achieved the highest specificity, positive, and negative predictive values (Table 3).

On decision curve analysis, the ML model achieved the highest net benefit across these thresholds (Figure 2B). This means that for every 100 cases, 1 (baseline), 8 (LR), and 14 (ML) more patients can safely undergo ipsilateral nerve-sparing at a threshold probability of 15% compared to a "treat all" strategy. When the 15% cut-off was applied to each model on the combined training and testing cohorts, 195 (baseline: 28%), 257 (LR: 37%), and 308 (ML: 44%) out of 695 non-ssEPE cases were below the cut-off and thus nerve-sparing can be safely considered.

Understanding the ML model

Feature importance rankings demonstrated that PSA, maximum % core involvement, and perineural invasion were the three most important features in the ML model to predict ssEPE (Figure 3).

The individual impact of each feature on probability of ssEPE is shown in Figure 4. These plots reveal the complex relationships that were captured by the ML model. An approximate linear association was observed for maximum % core involvement and % Gleason pattern 4/5. PSA exhibits a logarithmic relationship in which the probability of ssEPE rises sharply at low values but remains relatively stable beyond PSA of 10.

Discussion

As novel ML applications continue to be adopted in urology[15], there is growing need to better understand these "black box" models to validate their safety, reliability, and use in appropriate clinical contexts. We demonstrated that addition of a few readily available biopsy features using a ML approach can further improve predictions. By incorporating the complete clinicopathological profile, our model is better able to capture the patient's specific tumour burden and generate a unique ssEPE signature for personalized treatment planning. Our model was adequately powered and outperformed the baseline and LR models with respect to diagnostic accuracy, calibration, and clinical utility. Furthermore, our ML model was comparable to newer, validated models incorporating MRI findings (Martini AUROC: 0.68-0.78[16–18]; Soeterik AUROC: 0.77-0.83[5]) despite using just clinicopathological features (Supplementary Table 5). A particular strength of our model the ability to identify more candidates who can safely undergo nerve-sparing without sacrificing sensitivity.

These improvements can be attributed to the addition of clinically relevant biopsy features and model selection as evidenced by incremental performance gains from the baseline, LR, to ML models. Compared to previous models, this is the first study to incorporate

quantitative % Gleason pattern 4/5, perineural invasion, and site-specific features to predict ssEPE. The prognostic value of the former two features have been well described[19–21]. Inclusion of site-specific features reinforces the role of tumour location as an important predictor of ssEPE[22]. Base % core involvement was the highest ranked site-specific feature and base findings was ranked above worst Gleason Grade Group. These results extend previous findings that a positive basal core is an independent predictor of ssEPE[22,23]. Tumour involvement in the transition zone and mid gland are likely surrogate measures of overall tumour burden[24]. Post-hoc analysis of the combined cohorts revealed that 214/317 transition zone (68%) and 364/575 mid gland (63%) biopsies with positive cores had contiguous tumour involvement in the basal cores ($p<0.01$), which has been shown to be associated with EPE compared isolated positive cores[25]. With respect to model selection, the performance gain from LR to ML may be attributed to the ML model's ability to learn non-linear relationships between features and label. XGBoost have been shown to perform favourably compared to regression-based and other ML models in other clinical applications[8,26].

XAI enables clinicians to assess whether model predictions are aligned with clinical intuition. To demonstrate how to interpret the ML model and highlight the benefits of explainability, consider the case of a patient that did not have ssEPE on histopathological review (Figure 4). The baseline and LR model predicted a 35 and 19% probability of ssEPE, respectively, compared to 9% by the ML model. Using a cut-off of 15%, this patient would have been appropriately recommended a nerve-sparing approach using the ML model. Figure 5A provides an at-a-glance view for an individual patient by highlighting pertinent features that are predominantly driving this prediction. Through this representation, one can appreciate that young age, low % Gleason pattern 4/5, and Grade Group 1 disease in the basal cores outweigh the risk of ssEPE from presence of perineural invasion. Figure 5B presents an in-depth view of the same case by demonstrating the additive effects of each feature included in the ML model. Through this representation, one can identify patient-specific positive and negative factors based on their unique clinicopathological profile.

Finally, our ML model is accessible in that it leverages features that are part of the standard diagnostic workup for localized prostate cancer and does not require additional specialized investigations that are often limited to academic centers. This broadens the applicability of our model to the Canadian population regardless if they are managed at community or academic settings. Our model can be accessed at www.ssepe.ml.

Despite the strengths of this study, several limitations have to be acknowledged. Firstly, MRI features were not included as only a small number of patients in our cohort received pre-operative MRI. However, use of MRI has not been shown to reduce positive surgical margin rates at prostatectomy[27], and has poor sensitivity in detecting microscopic EPE (sensitivity~0.57)[28]. With the adoption of Prostate Imaging Reporting and Data System (PI-RADS) v2, future work will aim to incorporate MRI findings as it becomes routinely used in the pre-operative setting in Canada. Secondly, our model can be further strengthened with additional

external validation using diverse patient cohorts. Thirdly, sampling bias may be introduced as patients with incomplete records were excluded, mostly due to missing site-specific data from prostate biopsies done outside of our institutions. However, only a small proportion (11%) of patients from the training cohort were excluded. Furthermore, our model cannot be applied to patients who received neoadjuvant treatment prior to prostatectomy. From a ML perspective, SHAP is a post-hoc explainability method that is not without limitations, particularly due to issues with consistency and uncertainty[29,30]. Explainability at the individual patient level using XAI is still actively being investigated. Therefore, while XAI can provide us with some insight to better understand the ML model, these explanations should not be accepted as the ground truth. Finally, the reproducibility of ML models has recently been scrutinized due to lack of standardized methodology and reporting of results[31]. To address this, we adhered to a systematic approach based on the STREAM-URO framework[9] to describe the ML methodology used in this study to ensure transparency, reliability, and interpretability of results.

Conclusions

In the present study, we developed and validated a novel ssEPE prediction tool using ML. We illustrated how XAI can be applied to help unlock these "black-box" ML models. To highlight the clinical utility of our model, we demonstrated how our model is able to identify more patients who can safely undergo nerve-sparing prostatectomy compared to current predictive models. We have developed a simple, online tool to enable clinicians without prior ML experience to assess ssEPE risk. These efforts may help provide more personalized counseling and surgical planning for patients.

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(4):760-769. doi:10.1016/j.juro.2017.02.3344
2. Mottet N, Bellmunt J, Bolla M, et al. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur Urol*. 2017;71(4):618-629. doi:10.1016/j.eururo.2016.08.003
3. 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(1):76-82. doi:10.1111/bju.13733
4. Srigley JR, Humphrey PA, Amin MB, et al. Protocol for the examination of specimens from patients with carcinoma of the prostate gland. *Arch Pathol Lab Med*. 2009;133(10):1568-1576. doi:10.1043/1543-2165-133.10.1568
5. 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*. 2020;0(0). doi:10.1016/j.euo.2020.08.008
6. Rudzicz F, Joshi S. Explainable AI for the Operating Theater. In: *Digital Surgery*. Springer International Publishing; 2021:339-350. doi:10.1007/978-3-030-49100-0_25
7. Tonekaboni S, Joshi S, Mccradden MD, Goldenberg A, Ai AG. *What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use*. PMLR; 2019.
8. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2(10):749-760. doi:10.1038/s41551-018-0304-0
9. Kwong JCC, McLoughlin LC, Haider M, et al. Standardized Reporting of Machine Learning Applications in Urology: The STREAM-URO Framework. *Eur Urol Focus*. August 2021. doi:10.1016/j.euf.2021.07.004
10. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148(3):839-843. doi:10.1148/radiology.148.3.6878708
11. Epstein JI, Zelefsky MJ, Sjoberg DD, et al. A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score. *Eur Urol*. 2016;69(3):428-435. doi:10.1016/j.eururo.2015.06.046
12. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol 13-17-August-2016. New York, NY, USA: Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785
13. Lundberg SM, Allen PG, Lee S-I. *A Unified Approach to Interpreting Model Predictions*.; 2017.
14. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. *Med Decis Mak*. 2006;26(6):565-574. doi:10.1177/0272989X06295361
15. 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(4):567-577.

doi:10.1111/bju.14852
16. Martini A, Gupta A, Lewis SC, et al. Development and internal validation of a side-specific, multiparametric magnetic resonance imaging-based nomogram for the prediction of extracapsular extension of prostate cancer. *BJU Int*. 2018;122(6):1025-1033. doi:10.1111/bju.14353
17. Sighinolfi MC, Sandri M, Torricelli P, et al. External validation of a novel side-specific, multiparametric magnetic resonance imaging-based nomogram for the prediction of extracapsular extension of prostate cancer: preliminary outcomes on a series diagnosed with multiparametric magnetic resonance imaging-targeted plus systematic saturation biopsy. *BJU Int*. 2019;124(2):192-194. doi:10.1111/bju.14665
18. 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(5):372-378. doi:10.1016/j.urolonc.2019.12.028
19. Srigley JR, Delahunt B, Samaratunga H, et al. Controversial issues in Gleason and International Society of Urological Pathology (ISUP) prostate cancer grading: proposed recommendations for international implementation. *Pathology*. 2019;51(5):463-473. doi:10.1016/j.pathol.2019.05.001
20. Bostwick DG, Qian J, Bergstralh E, et al. Prediction of capsular perforation and seminal vesicle invasion in prostate cancer. *J Urol*. 1996;155(4):1361-1367.
21. Cozzi G, Rocco BM, Grasso A, et al. Perineural invasion as a predictor of extraprostatic extension of prostate cancer: A systematic review and meta-analysis. *Scand J Urol*. 2013;47(6):443-448. doi:10.3109/21681805.2013.776106
22. Taneja SS, Penson DF, Epelbaum A, Handler T, Lepor H. Does site specific labeling of sextant biopsy cores predict the site of extracapsular extension in radical prostatectomy surgical specimen. *J Urol*. 1999;162(4):1352-1357. doi:10.1016/S0022-5347(05)68284-5
23. Naya Y, Slaton JW, Troncoso P, Okihara K, Babaian RJ. Tumor Length and Location of Cancer on Biopsy Predict for Side Specific Extraprostatic Cancer Extension. *J Urol*. 2004;171(3):1093-1097. doi:10.1097/01.ju.0000103929.91486.29
24. Ishizaki F, Hara N, Koike H, et al. Prediction of pathological and oncological outcomes based on extended prostate biopsy results in patients with prostate cancer receiving radical prostatectomy: a single institution study. *Diagn Pathol*. 2012;7(1):1-8.
25. Kryvenko ON, Diaz M, Meier FA, Ramineni M, Menon M, Gupta NS. Findings in 12-Core Transrectal Ultrasound-Guided Prostate Needle Biopsy That Predict More Advanced Cancer at Prostatectomy. *Am J Clin Pathol*. 2012;137(5):739-746. doi:10.1309/AJCPWIZ9X2DMBEBM
26. Kawakita S, Beaumont JL, Jucaud V, Everly MJ. Personalized prediction of delayed graft function for recipients of deceased donor kidney transplants with machine learning. *Sci Rep*. 2020;10(1):18409. doi:10.1038/s41598-020-75473-z
27. Rud E, Baco E, Klotz D, et al. Does Preoperative Magnetic Resonance Imaging Reduce the Rate of Positive Surgical Margins at Radical Prostatectomy in a Randomised Clinical Trial? *Eur Urol*. 2015;68(3):487-496. doi:10.1016/j.eururo.2015.02.039
28. 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(2):233-245. doi:10.1016/j.eururo.2015.07.029
29. Camburu OM, Giunchiglia E, Foerster J, Lukasiewicz T, Blunsom P. The struggles of feature-based explanations: Shapley Values vs. Minimal Sufficient Subsets. *arXiv*. September 2020.
30. Kumar IE, Venkatasubramanian S, Scheidegger C, Friedler S. Problems with Shapley-value-based explanations as feature importance measures. *arXiv*. February 2020.
31. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell*. 2021;3(3):199-217. doi:10.1038/s42256-021-00307-0

Figures and Tables

Fig. 1. Patient inclusion flowchart.

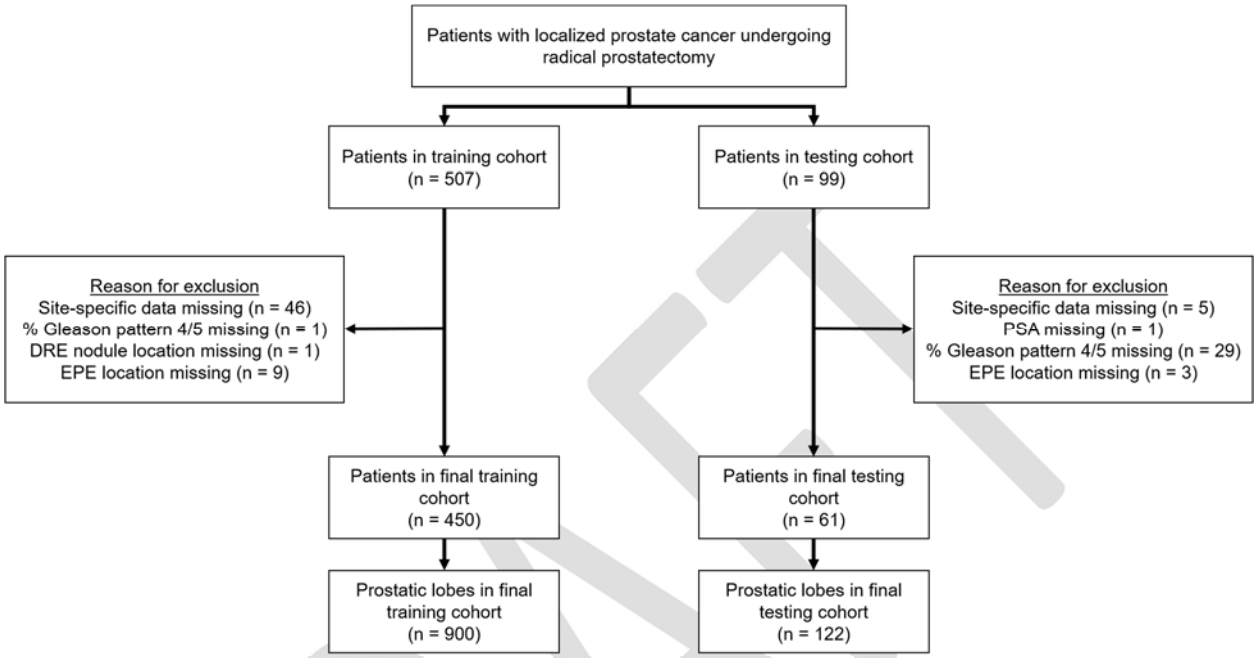
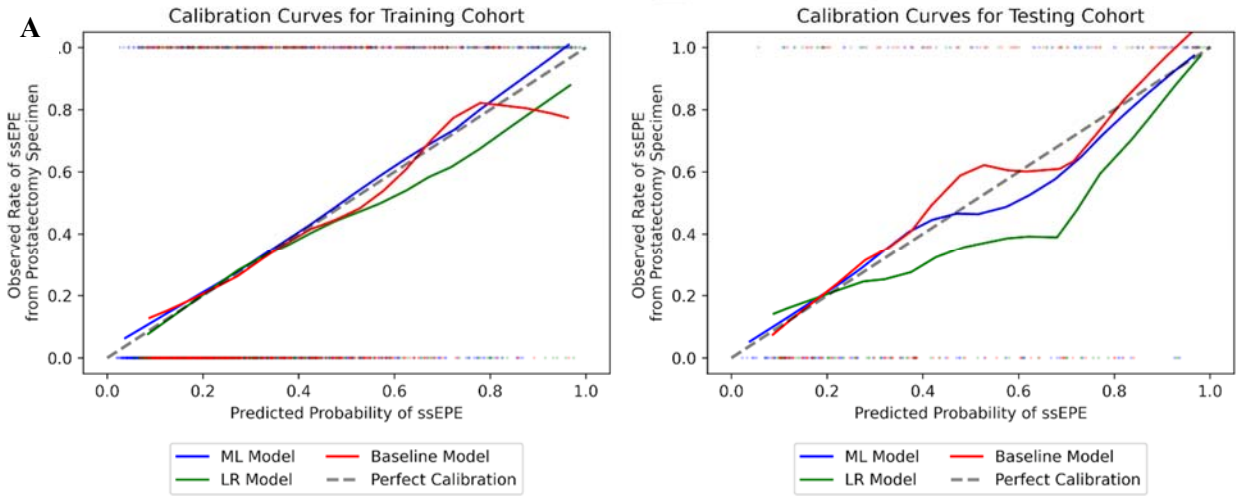
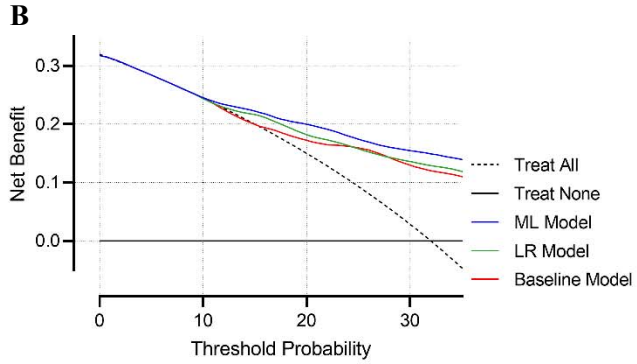


Fig. 2. (A) Calibration curves for the ML (blue), LR (green), and baseline (red) models on stratified tenfold cross-validation of the training cohort and the testing cohort. A perfectly calibrated model corresponds to a 45-degree line. If the calibration curve is above the reference line, it underestimates the risk of ssEPE, which may lead to undertreatment (i.e., risk of positive surgical margins). However, if a calibration curve is below the reference line, it overestimates the risk of ssEPE, which may lead to overtreatment (ie: patient gets unnecessarily treated with a non-nerve sparing approach). ***(B)*** Net benefit of the ML, LR, and baseline models using decision curve analysis on the combined training and testing cohorts. The gain in appropriate ipsilateral nerve-sparing per 100 cases compared to a "treat all" strategy is shown for threshold probabilities from 10–30%.





Threshold Probability	Increase in appropriate ipsilateral nerve-sparing per 100 cases		
	Baseline Model	LR Model	ML Model
10.0	0	0	1
12.5	0	1	6
15.0	1	8	14
17.5	5	13	15
20.0	8	14	20
22.5	14	17	22
25.0	20	19	24
27.5	22	22	26
30.0	24	25	30

Fig. 3. Feature importance rankings of the ML model based on the average impact on probability of ssEPE.

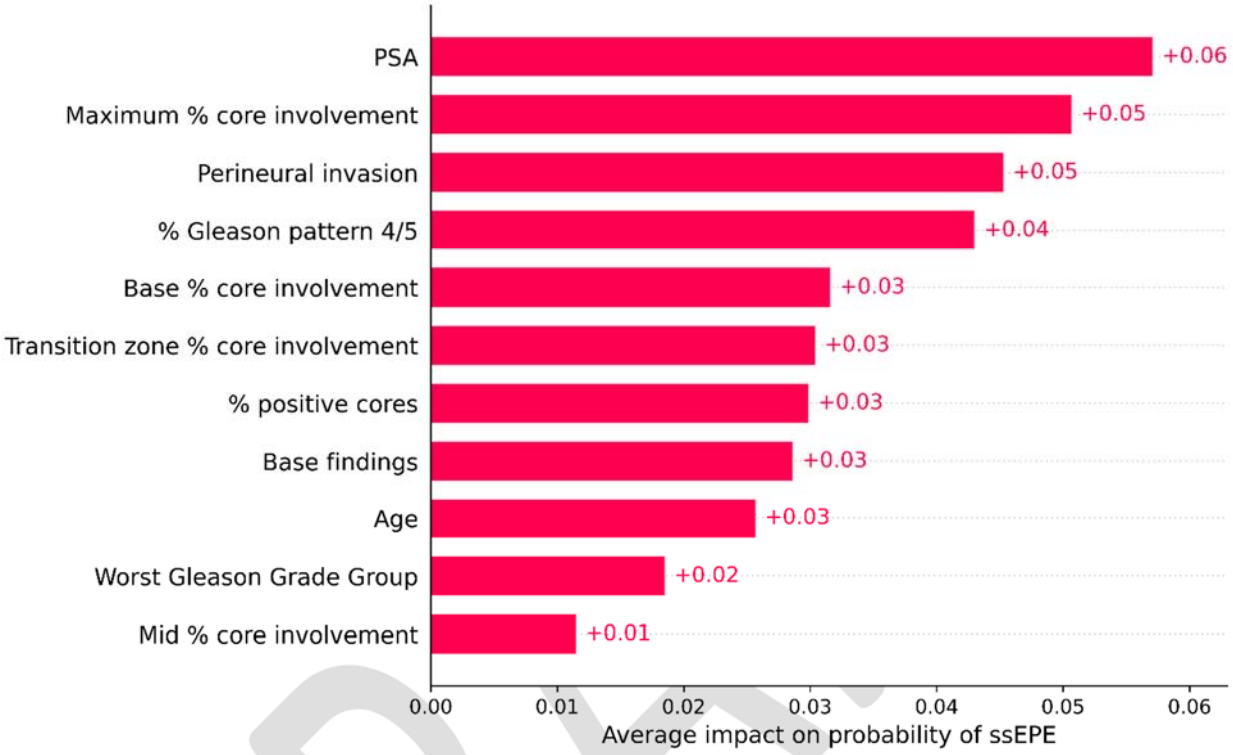


Fig. 4. Partial dependence plots showing the change in probability of ssEPE across all values for each feature. Each data point represents an individual case, while histograms on each plot show the distribution of values for that feature. The overall trend for each plot is depicted by the red line. ASAP, atypical small acinar proliferation; GGG, Gleason Grade Group; HGPIN, high-grade prostatic intraepithelial neoplasia.

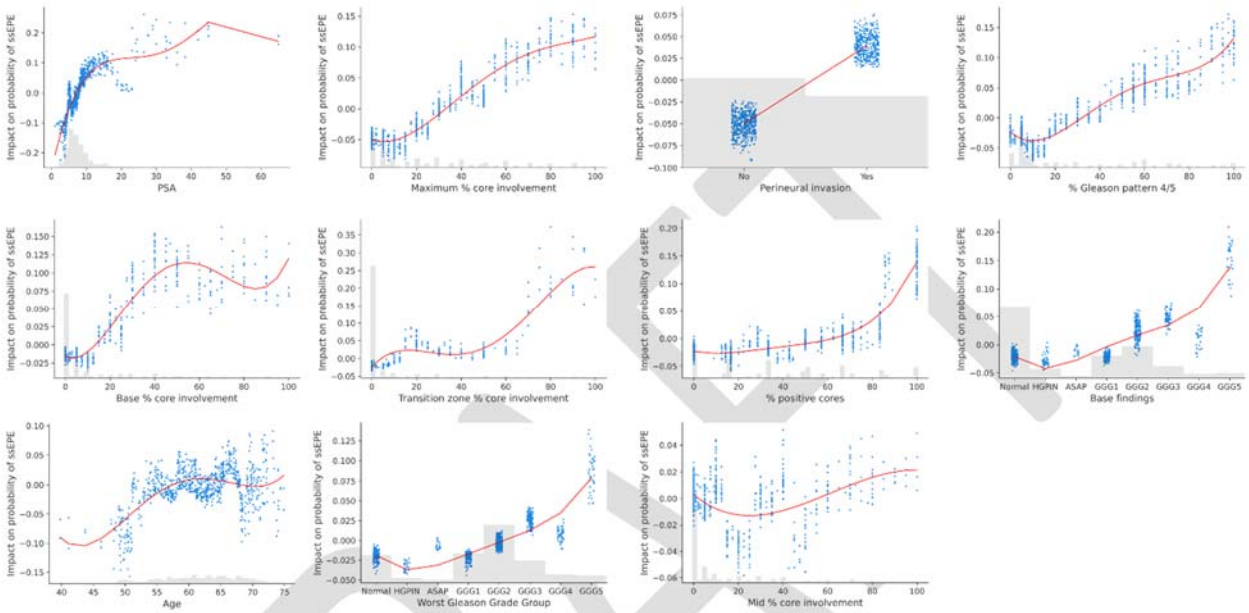


Fig. 5. Probability of ssEPE for an individual case explained using feature contributions. This case was predicted to have a 35 and 19% probability of ssEPE based on the baseline and LR model, respectively, but did not demonstrate ssEPE on pathological review. ***(A)*** This plot highlights the most influential features on the final prediction. Features in red increase the probability of ssEPE (push to the right) while those in blue decrease the probability (push to the left). ***(B)*** Detailed plot showing the cumulative effects of all contributing features. Moving from the bottom to the top of the plot, the effect of each feature is added by increasing order of impact to generate the final probability of 9% for the ML model. Patient-specific feature input values are indicated on the left. For perineural invasion, the value of 1 was coded for "presence of perineural invasion". For base findings, the value of 3 corresponds to Gleason Grade Group 1. For Worst Gleason Grade Group, the value of 4 corresponds to Gleason Grade Group 2.

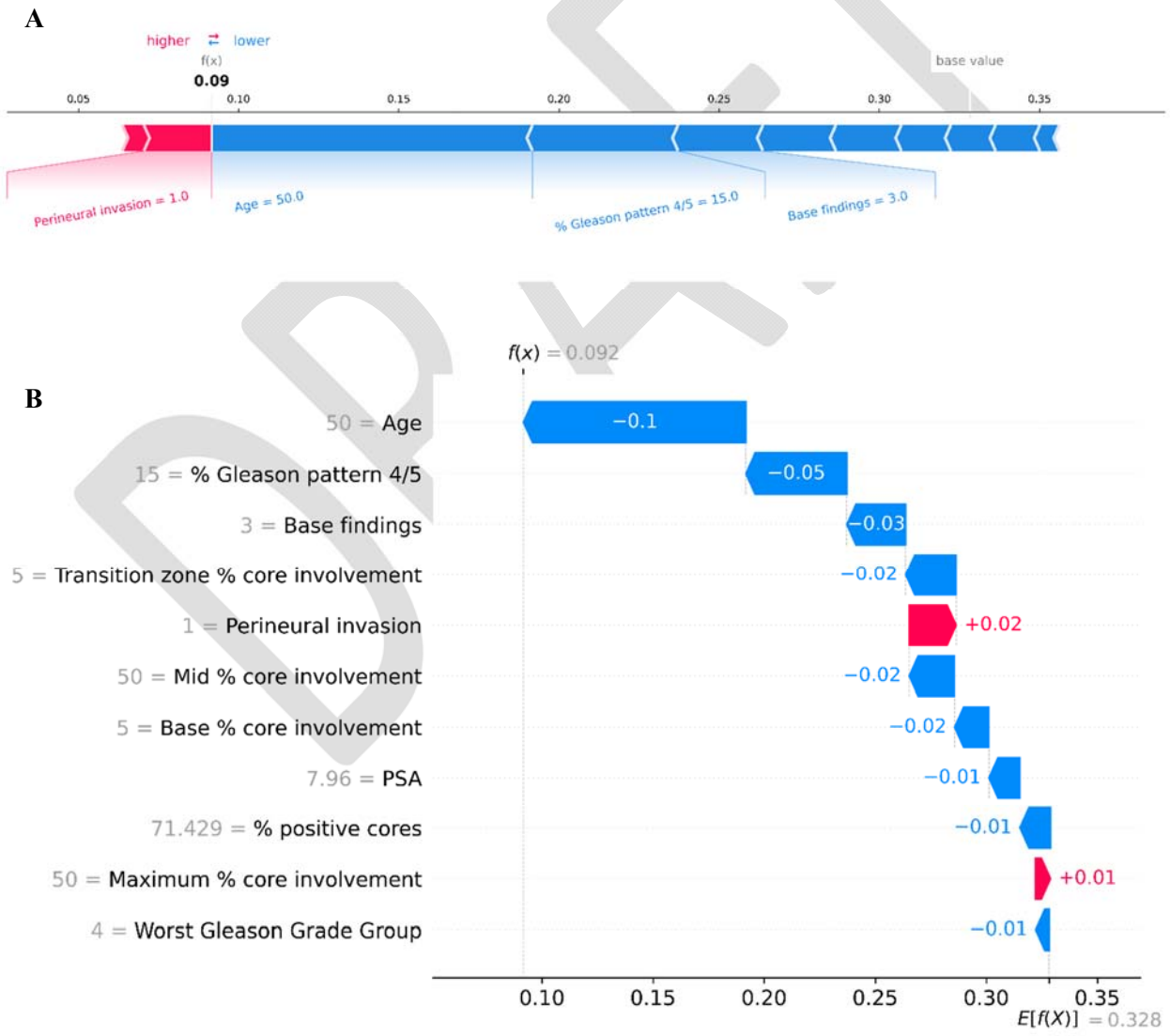


Table 1. Clinicopathological characteristics of the study population			
	Training cohort	**Testing cohort**	**p**
No. of lobes	900	122	
Age, median (IQR)	62 (57–66)	62 (57–65)	0.41
PSA (ng/mL), median (IQR)	7.06 (5.50–9.30)	8.20 (6.00–12.20)	<0.01
% Gleason pattern 4/5, median (IQR)	10.0 (5.0–55.0)	32.5 (10.0–70.0)	<0.01
Perineural invasion, n (%)	400 (44.4)	62 (50.8)	0.18
Prostate volume (mL), median (IQR)	34 (25–44)	35 (27–42)	0.77
Palpable nodule on DRE, n (%)	192 (21.3)	27 (22.1)	0.84
Hypoechoic nodule on TRUS, n (%)	106 (11.8)	13 (10.7)	0.72
% site involvement, median (IQR)	50 (25–75)	50 (25–75)	0.04
% positive cores, median (IQR)	33.3 (14.3–66.7)	42.9 (16.7–83.3)	0.01
Worst Gleason Grade Group, n (%) Normal HGPIN ASAP Grade Group 1 Grade Group 2 Grade Group 3 Grade Group 4 Grade Group 5	 155 (17.2) 26 (2.9) 17 (1.9) 166 (18.4) 326 (36.2) 124 (13.8) 46 (5.1) 40 (4.4)	 20 (16.4) 2 (1.6) 1 (0.8) 13 (10.7) 44 (36.1) 21 (17.2) 15 (12.3) 6 (4.9)	 <0.01
% core involvement at worst Gleason Grade Group, median (IQR)	15.0 (5.0–40.0)	33.8 (5.0–70.0)	<0.01
Maximum % core involvement, median (IQR)	20.0 (5.0–50.0)	40.0 (5.0–75.0)	<0.01
Gleason Grade Group at most involved core, n (%) Normal HGPIN ASAP Grade Group 1 Grade Group 2 Grade Group 3 Grade Group 4 Grade Group 5	 183 (20.3) 11 (1.2) 6 (0.7) 208 (23.1) 320 (35.6) 108 (12.0) 32 (3.6) 32 (3.6)	 23 (18.9) 0 (0) 0 (0) 22 (18.0) 42 (34.4) 21 (17.2) 9 (7.4) 5 (4.1)	 0.03
Base finding, n (%) Normal HGPIN ASAP Grade Group 1 Grade Group 2 Grade Group 3	 407 (45.2) 52 (5.8) 14 (1.6) 127 (14.1) 180 (20.0) 66 (7.3)	 51 (41.8) 5 (4.1) 2 (1.6) 16 (13.1) 27 (22.1) 7 (5.7)	 0.21

Grade Group 4	25 (2.8)	9 (7.4)	
Grade Group 5	29 (3.2)	5 (4.1)	
Base % positive cores, mean (SD)	35.1 (41.9)	47.1 (46.9)	0.02
Base % core involvement, mean (SD)	13.1 (22.3)	24.5 (33.9)	0.01
Mid finding, n (%)			
Normal	354 (39.3)	42 (34.4)	
HGPIN	33 (3.7)	2 (1.6)	
ASAP	16 (1.8)	3 (2.5)	
Grade Group 1	164 (18.2)	18 (14.8)	0.07
Grade Group 2	219 (24.3)	36 (29.5)	
Grade Group 3	75 (8.3)	12 (9.8)	
Grade Group 4	19 (2.1)	7 (5.7)	
Grade Group 5	20 (2.2)	2 (1.6)	
Mid % positive cores, mean (SD)	44.5 (43.7)	52.1 (45.4)	0.08
Mid % core involvement, mean (SD)	15.3 (22.8)	25.3 (31.1)	<0.01
Apex finding, n (%)			
Normal	471 (52.3)	48 (39.3)	
HGPIN	25 (2.8)	7 (5.7)	
ASAP	23 (2.6)	7 (5.7)	
Grade Group 1	150 (16.7)	12 (9.8)	<0.01
Grade Group 2	145 (16.1)	23 (18.9)	
Grade Group 3	54 (6.0)	16 (13.1)	
Grade Group 4	18 (2.0)	7 (5.7)	
Grade Group 5	14 (1.6)	2 (1.6)	
Apex% positive cores, mean (SD)	39.9 (47.6)	47.5 (50.1)	0.12
Apex % core involvement, mean (SD)	13.2 (22.9)	23.1 (32.1)	0.01
Transition zone finding, n (%)			
Normal	609 (67.7)	58 (47.5)	
HGPIN	18 (2.0)	4 (3.3)	
ASAP	8 (0.9)	7 (5.7)	
Grade Group 1	86 (9.6)	12 (9.8)	<0.01
Grade Group 2	116 (12.9)	19 (15.6)	
Grade Group 3	38 (4.2)	12 (9.8)	
Grade Group 4	18 (2.0)	8 (6.6)	
Grade Group 5	7 (0.8)	2 (1.6)	
Transition zone % positive cores, mean (SD)	28.7 (44.8)	42.6 (49.7)	<0.01
Transition zone % core involvement, mean (SD)	8.8 (19.6)	21.4 (31.9)	<0.01
ssEPE, n (%)	276 (30.7)	51 (41.8)	0.01

ASAP: atypical small acinar proliferation; HGPIN: high-grade prostatic intraepithelial neoplasia; TRUS: transrectal ultrasound.

Table 2. Performance metrics of the ML, LR, and baseline models on the training and testing cohorts

		Predictive models			p for pairwise comparisons		
		Baseline	LR	ML	LR vs. baseline	ML vs. baseline	ML vs. LR
Training cohort (stratified tenfold cross-validation)	AUROC (95% CI)	0.74 (0.70–0.77)	0.78 (0.74–0.81)	0.81 (0.77–0.83)	<0.01	<0.01	<0.01
	AUPRC (95% CI)	0.59 (0.53–0.66)	0.64 (0.58–0.70)	0.69 (0.63–0.74)	<0.01	< 0.01	<0.01
Testing cohort	AUROC (95% CI)	0.75 (0.66–0.83)	0.76 (0.67–0.84)	0.81 (0.73–0.88)	0.40	0.01	<0.01
	AUPRC (95% CI)	0.70 (0.59–0.81)	0.75 (0.65–0.85)	0.78 (0.67–0.87)	0.02	<0.01	0.08

The confidence intervals and p-values for pairwise comparisons were calculated using 10 000 bootstrap replications. AUROC: area under the receiver-operating-characteristic curve; AUPRC: area under the precision-recall curve; CI: confidence interval.

Table 3. Performance metrics for the ML, LR, and baseline models per threshold cutoff using the combined training and testing cohorts

Cutoff (%)	Sensitivity			Specificity			Positive predictive value			Negative predictive value		
	Baseline	LR	ML	Baseline	LR	ML	Baseline	LR	ML	Baseline	LR	ML
10.0	0.97	0.98	0.94	0.06	0.13	0.29	0.33	0.35	0.38	0.79	0.92	0.91
12.5	0.93	0.95	0.92	0.19	0.26	0.39	0.35	0.38	0.42	0.84	0.92	0.91
15.0	0.90	0.91	0.91	0.28	0.37	0.44	0.37	0.40	0.43	0.86	0.90	0.91
17.5	0.86	0.87	0.86	0.37	0.47	0.53	0.39	0.44	0.46	0.85	0.89	0.89
20.0	0.82	0.82	0.83	0.47	0.52	0.60	0.42	0.45	0.49	0.85	0.86	0.89
22.5	0.79	0.80	0.82	0.54	0.58	0.65	0.44	0.47	0.53	0.84	0.86	0.88
25.0	0.76	0.76	0.78	0.60	0.63	0.68	0.47	0.50	0.54	0.84	0.85	0.87
27.5	0.74	0.72	0.76	0.66	0.68	0.72	0.51	0.51	0.56	0.84	0.84	0.86
30.0	0.71	0.68	0.72	0.69	0.72	0.74	0.52	0.53	0.57	0.84	0.83	0.85