Contents lists available at ScienceDirect

Current Problems in Diagnostic Radiology

journal homepage: www.cpdrjournal.com

Part II: Effect of different evaluation methods to the application of a computer-aided prostate MRI detection/diagnosis (CADe/CADx) device on reader performance☆,☆☆

Jeffrey H. Maki[a,*], Nayana U Patel[b], Ethan J Ulrich[c], Jasser Dhaouadi[c], Randall W Jones[c]

[a] Department of Radiology, University of Colorado Anschutz Medical Center, 12401 E 17th Ave (MS L954), Aurora, CO 80045, USA
[b] University of New Mexico Department of Radiology, Albuquerque, NM, USA
[c] BOT IMAGE, Inc., Omaha, NE, USA

ARTICLE INFO

Keywords:
Artificial Intelligence
CADe
CADx
Prostate cancer
Prostate cancer detection

ABSTRACT

Introduction: The construction and results of a multiple-reader multiple-case prostate MRI study are described and reported to illustrate recommendations for how to standardize artificial intelligence (AI) prostate studies per the review constituting Part I[1].

Methods: Our previously reported approach was applied to review and report an IRB approved, HIPAA compliant multiple-reader multiple-case clinical study of 150 bi-parametric prostate MRI studies across 9 readers, measuring physician performance both with and without the use of the recently FDA cleared CADe/CADx software ProstatID.

Results: Unassisted reader AUC values ranged from 0.418 – 0.759, with AI assisted AUC values ranging from 0.507 – 0.787. This represented a statistically significant AUC improvement of 0.045 ($\alpha = 0.05$). A free-response ROC (FROC) analysis similarly demonstrated a statistically significant increase in θ from 0.405 to 0.453 ($\alpha = 0.05$). The standalone performance of ProstatID performed across all prostate tissues demonstrated an AUC of 0.929, while the standalone lesion level performance of ProstatID at all biopsied locations achieved an AUC of 0.710.

Conclusion: This study applies and illustrates suggested reporting and standardization methods for prostate AI studies that will make it easier to understand, evaluate and compare between AI studies. Providing radiologists with the ProstatID CADe/CADx software significantly increased diagnostic performance as assessed by both ROC and free-response ROC metrics. Such algorithms have the potential to improve radiologist performance in the detection and localization of clinically significant prostate cancer.

Introduction

A concurrently published companion Part I review explores the current landscape of artificial intelligence (AI) as applied to prostate MRI, including how results are obtained, what they mean, and methodological recommended guidelines for standardizing how prostate AI studies are analyzed and reported.[1] The main goal of this Part II manuscript is to demonstrate the application of the Part I recommendations in the context of a clinical study in an attempt to better standardize how AI studies are constructed and reported, thereby making it more feasible to interpret and compare clinical performance between studies. These recommendations were summarized in Part I as[1]:

- **Biologic Truth**: carefully correlate imaging data with pathology and/or negative follow up data, describe how this correlation is performed and define what is truly a lesion “hit” versus “miss” based upon the granularity of division
  - ○ Specifically define the degree of granularity (or grid size) of the CADe/CADx algorithm in calculating ROC curves
- **Patient/Case Selection**: clinical performance cases should come from an enriched dataset drawn from multiple machine types, with random sampling from MRI-negative cases with negative follow-up, MRI-positive cases with negative follow-up, and MRI-positive cases with positive follow-up

☆ **Date Submitted:** Jan 12, 2024☆☆ **Grant Support:** None

* Corresponding author.
E-mail address: jeffrey.maki@cuanschutz.edu (J.H. Maki).

https://doi.org/10.1067/j.cpradiol.2024.04.003

Available online 21 April 2024

- Clearly define the threshold for clinically significant cancer - typically Gleason score ≥ 7 – or if any other grading of disease severity is employed, this should be so stated
- **Performance Evaluation**:
  - MR interpretation: clearly define how reader MRI interpretation was performed – e.g. PI-RADS, recommendation for biopsy etc.
  - AI assessment type: clearly define whether CADe/CADx standalone assessment (e.g. ROC) is on a case level or a lesion only level, and report both
  - Free Response ROC: conduct a CADe/CADx standalone weighted FROC performance analysis on the ground truth biopsy points
  - Impact of AI on reader performance: conduct a comparison of reader performance, both in terms of Diagnostic Performance and Detection Performance with and without aid of CADe/CADx, with additional recommendations -
    - Employ a statistically relevant number of readers, preferably with varying degrees of experience
    - Reader analysis should be multiple-reader multiple-case (MRMC) at the modified case level, clearly defining the PI-RADS (or other) criteria used for establishing the case-level diagnostic performance of the readers via construction of the ROC curve
    - Conduct a comparative FROC performance of the MRMC readers at the biopsy data points with and without the use of CADe/CADx without the reader's knowledge of the location and status of ground truth lesions to establish detection performance - note this measurement is exclusively at those biopsy points with *a-priori* ground truth data, and not at all points (pixels) within the prostate

Methods

A retrospective multiple-reader (n = 9), multiple-case (n = 150) (MRMC) clinical study was designed and implemented incorporating the above recommendations to illustrate the effect of a CADe/CADx AI algorithm on the performance of its intended users. The primary goal was to illustrate the effect of different evaluation methods (case level and lesion level analysis with ROC and FROC) on the clinical performance of CADe/CADx devices. This was assessed by determining whether providing radiologists with output from ProstatID (BOT IMAGE, Inc. Omaha, NE), a recently FDA-cleared CADe/CADx tool for prostate cancer detection and localization, leads to a statistically significant improvement in performance over the current standard of care using suggested analysis techniques. Secondary goals were to measure the standalone performance of ProstatID and measure the impact on the intended users' detection accuracy, and test if the use of AI would reduce unnecessary biopsies. While data from this study has already been in part published,[2] this work further evaluates these data, particularly in the context of our recommendations on study design and reporting. In addition, more comprehensive details of the ProstateID AI algorithm are presented.

The primary performance endpoint was to measure changes in AUC in a multiple reader, multiple case (MRMC) prostate MRI study employing ProstatID. This was planned as a multiple reader case level analysis without and with use of the ProstatID algorithm, with an additional FROC analysis at the biopsy points and a standalone lesion level analysis of ProstatID. To this means, a power analysis was first performed estimating an experiment with 7 readers and 148 cases being sufficient to detect a difference of 0.05 ROC AUC and achieve a power of 0.80, further based on the assumption that half the cases would be positive and half negative. To be conservative, a total of 9 readers were selected from 6 institutions and with varying experience with prostate MRI (Table 1, median 3 years, range less than 1 year to 10 years) to evaluate 150 retrospective cases of bi-parametric prostate MRI (bpMRI) established prior to vendor algorithm development and training. In keeping with our suggestions, these cases were stratified and sampled from a larger data set of approximately 2000 retrospective cases from 6 contributing sites and 12 machine types (both 1.5 and 3T) so as to end up with a near equal mix of normal, suspicious negative, and suspicious positive cases, with an approximate 40 % positive rate. In all cases suspicious lesions were validated by biopsy results (see Truthing Methods below). These data are summarized in Tables 2 and 3. All sites contributing data for this retrospective study followed their local IRB guidelines.

Table 1
Study reader practice type, number of years' experience interpreting prostate MRI, and change in AUC for ProstatID assisted read.

Reader #	Practice Type	Years Interpreting pMRI	Baseline AUC	Δ AUC
1	Hospital Based	2	0.70	0.04
2	University	3	0.63	0.08
3	Private Practice	4	0.72	0.06
4	Private Practice	1	0.42	0.09
5	Hospital Based	4	0.76	-0.02
6	University	10	0.64	0.10
7	Private Practice	2	0.74	0.04
8	Private Practice	8	0.74	0.00
9	University	2	0.72	0.02

Readers were blinded to patient results and instructed to interpret all 150 bpMRI cases (T2 weighted (T2w), diffusion weighted (DWI), apparent diffusion coefficient (ADC) map) according to PI-RADS v 2.1, using their preferred local DICOM viewer. After a washout period of at least 30 days, the readers were instructed to re-interpret all 150 cases (presented in a different order), this time with the assistance of output from the ProstatID algorithm. This assistance was presented to the readers as an additional registered color-coded image set of predicted malignancy risk overlaid on standard axial T2w imaging that they could utilize and interpret as they wished, similar to that shown in Fig. 1. In the majority of cases the patient's prostate specific antigen (PSA) level was known for both reads, and the readers had access to this information.

Truthing methods

Biological truth

Our reference standard for clinically significant prostate cancer was detection of Gleason grade ≥7 prostate cancer biopsy points on in-bore (n = 57), US/MR fusion (n = 32), and/or cognitive US/MR fusion (n = 17) biopsy. Our reference standard for "negative" were positive MRI for suspicious lesion (by radiologic interpretation) but negative biopsy (Gleason grade <7), acknowledging the unavoidable caveat that biopsy is an imperfect process and it is possible biopsy could miss the intended MR abnormality yielding a false negative. The cases used in this clinical study, however, had image guided biopsy correlation using either in-bore (MRI) or US/MR fusion to keep the targeting error to less than 3 mm, thereby minimizing such potential false negative errors. "Positive" were lesions with positive biopsy confirmation. Our reference standard for "normal patient cohorts" was the consensus case review opinion of four additional blinded radiologists, with at least 15 years of experience in interpreting prostate MRI or an academic fellowship where substantial biological truth data was collected, who agreed that the patient's MRI was not suspicious for csPCa.

Scoring

The scoring criteria used to determine both the physician and CADe/CADx clinical performance was measurement of the physical overlap or concurrence of the sub-volume of the physician's described region of suspicion to the sub-volume of the reference standard to the Truth Table, a volumetric mapping of the location of all suspect lesions (with their

Table 2
Summary of scanner types, field strengths, imaging parameters and patient demographics used in the clinical study.

Manufacturer/Field Strength	Philips 1.5T	Philips 3.0T	Siemens 3.0T	GE 1.5T	GE 3.0T	Overall
Scanner Models	Achieva, Ingenia	Achieva, Ingenia, Intera	Magnetom Vida, Skyra, TrioTim, Verio	Signa HDxt, Optima MR450w	Discovery 750w	
T2W						
Repetition Time (ms)	2975 [2650-6869]	4434 [3752-6435]	4780 [3000-10500]	2975 [2650-6869]	9999 [1500-13599]	4434 [1500-13599]
Echo Time (ms)	125 [120-130]	120	121 [97-123]	96 [91-118]	109 [102-115]	120 [91-130]
Flip Angle (degrees)	90	90	137 [120-160]	90 [90-160]	160 [90-160]	90 [90-160]
Slice thickness (mm)	4.0 [3.0, 4.6]	3.0	3.0 [3.0-3.8]	3.5 [3.5-4.0]	4.0	3.3 [3.0-4.6]
Square matrix size (pix)	512 [512-672]	512 [512-576]	320 [320-640]	512	512	512 [320-672]
In-plane field of view (mm)	200 [160-246]	140 [140-180]	200 [140-220]	200 [180-260]	200 [200-220]	200 [140-260]
In-plane resolution (mm)	0.391 [0.297-0.391]	0.273 [0.273-0.321]	0.573 [0.281-0.688]	0.391 [0.352-0.508]	0.391 [0.391-0.430]	0.391 [0.273-0.688]
DWI						
Repetition Time (ms)	3360 [2700-4847]	6804 [4000-7036]	4800 [3900-7600]	6000 [3515-6000]	4000 [2000-4877]	4800 [2000-7600]
Echo Time (ms)	67 [65-81]	52 [51-84]	62 [62-121]	85 [68-87]	72 [68-74]	67 [51-121]
High b-value (s/mm^2)	1400 [1000-1400]	2000 [750-2000]	1500 [800-2000]	1400	1450 [1150-1450]	1400 [750-2000]
Slice thickness (mm)	4.0 [3.3-4.6]	3.0 [3.0-3.3]	3.5 [3.0-5]	3.0 [3.0-4.0]	4.0 [4.0-4.2]	4.0 [3.0-5.0]
Square matrix size (pix)	160 [128-256]	256 [128-256]	118 [96-280]	256	256	224 [96-280]
In-plane field of view (mm)	216 [180-360]	140 [140-180]	200 [169-240]	240 [240-260]	256 [220-256]	200 [140-360]
In-plane resolution (mm)	1.389 [1.216-1.417]	0.547 [0.547-1.406]	1.695 [0.714-2.255]	0.938 [0.938-1.016]	1.000 [0.859-1.000]	1.287 [0.547-2.255]
Patients						
Total	44	35	42	11	18	150
Age (years)	68.5 [50-82]	65 [49-79]	69 [56-86]	63 [45-83]	65.5 [53-74]	67 [45-86]
With cancer, GS ≥ 7	18 (40.9 %)	16 (45.7 %)	20 (47.6 %)	4 (36.4 %)	9 (50.0 %)	67 (44.7 %)
Total Without cancer	26 (59.1 %)	19 (52.3 %)	22 (52.4 %)	7 (63.6 %)	9 (50.0 %)	83 (55.3 %)
Normal cases	*15 (34.1 %)*	*7 (20.0 %)*	*10 (23.8 %)*	*4 (36.4 %)*	*8 (44.4 %)*	*44 (29.3 %)*
PSA Level (ng/mL)	7.1 [1.3-367.2] 44 cases	6.4 [3.9-10.9] 10 cases	8.0 [2.0-36.6] 37 cases	4.1 [4.0-11.6] 3 cases	6.3 [0.4-20.7] 18 cases	7.2 [0.4-367.2] 112 cases (74.7 %)

Table 3
Biopsy details for the 150 case dataset used for the clinical study, which included a total of 209 pathology proven lesions, 81 of which were true positive (Gleason ≥7) and 128 of which were false positive (Benign or Gleason 6).

Breakdown of 150 Case Dataset – Biopsy Data						
Benign	Gleason 6	Gleason 7	Gleason 8	Gleason 9	Gleason 10	Total
100	28	57	13	9	2	209
True Positive (Gleason ≥7)						81
False Positive (Gleason <7 or Benign)						128

biopsy results hidden from readers and CADe/CADx). This Truth Table for the almost 1000 case dataset used (a subset of which is this 150 case study) consisted of malignant, benign, and normal tissues that were catalogued from carefully curated biopsy locations and clinical reports provided by the sources for each examination. All biopsy sites were *a-priori* chosen based on an expert panels' recommendation (consensus of highly trained academic radiologists specializing in prostate cancer diagnosis), who conservatively biopsied every suspicious point, leading to MRI-guided points that yielded a mix of negative and positive for cancer. The readers and AI were measured against these truth points with both the readers and AI evaluated against all of the lesion sites they independently chose (thus a full "all lesion" analysis as previously described). In the FROC lesion analysis, as per Part I,[1] if the reader or AI under test chose a point other than one of the *a-priori* points with known biological truth, they were not evaluated for this point in the lesion-only analysis because no biological ground truth data existed with which to measure them. This was the same with the modified case level analysis as only the biological ground truth points and their resulting true PI-RADS case score were used.

To obtain accurate reader assessment, the participating physicians were provided a computerized scoring sheet with the numbered patient ID, age, and PSA (if available), as well as columns of pull-down localization choices as described above which identify which sub-region of the prostate the centroid of the lesion was located. They were also instructed to answer whether their interpretation would result in an action, namely biopsy, which is the standard of care for follow-up to lesion classification providing it meets the interpreting physician's threshold. Note that this does vary amongst physicians. Further, each reader was instructed to use a unique row within the scoring sheet to describe each unique lesion for each patient. If there were no lesions or only one lesion was described, only one row would be completed for that patient. These scoring spreadsheets describing their chosen location of the lesion(s) centroid(s) were used for evaluating the accuracy of the physician interpretations to those of the Ground Truth table. This was accomplished using the overlap or concurrence of the physician-chosen sub-volume to that of the Truth Table sub-volume and its ground truth biopsy results; positive or negative for cancer. If the physician did not identify a lesion that was labeled as a true positive in the Truth Table, then the physician was assigned a false negative. Alternatively, if the physician indicated a lesion that was labeled as a true negative in the Truth Table, then the physician was assigned a false positive. This process continued for every lesion of every patient as compared to the Truth Table.

Biopsy decisions

To test if the use of ProstatID would reduce the decisions to biopsy a benign outcome, a mixed effects analysis was used. Fixed effects were the true positive state (outcome) and the use of ProstatID (modality). Random intercepts were included for each reader and each patient case. The response variable to predict was the decision to biopsy.

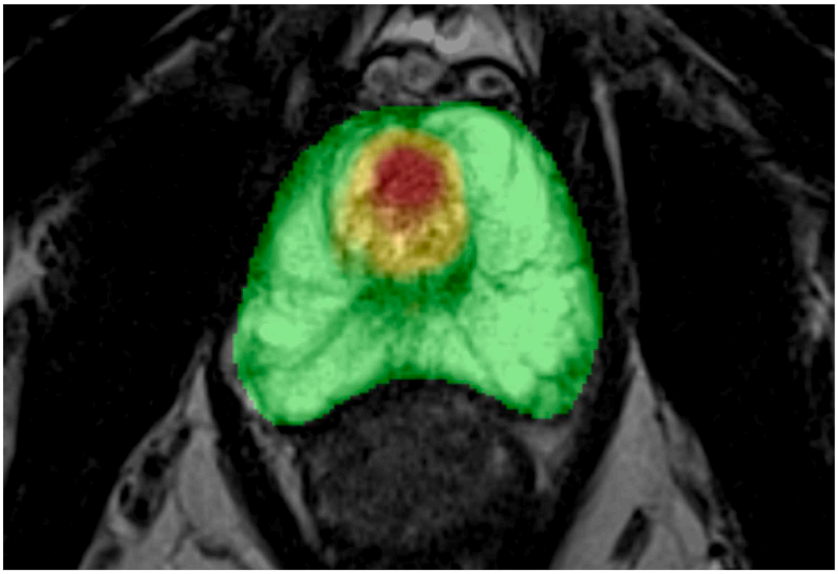

Fig. 1. Example of ProstatID color-coded overlay on a T2w slice through the mid prostate. Colormap per Table 4, with green predicting benign tissue, and the color spectrum toward bright red predicting the highest likelihood of clinically significant cancer. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

AI algorithm

The ProstatID software algorithm consists of six functional steps as per the Architecture Design Chart, which is outlined in Fig. 2 and summarized below. For the sake of brevity this work focuses on lesion detection and classification as well as case-level diagnosis; however, such software requires significantly more functionality in order to accomplish the required clinical components in a seamless manner. For this type of SaaS (Software as a Service) to operate, it requires additional functions including (1) automatic detection of a prostate MRI case at its cloud-based input; (2) testing to ensure that it is in fact a prostate case, that it contains the required series of T2w, DWI, and ADC, that the prostate is correctly centered within the field of view with adequate image quality and meets the minimum resolution requirements (and can send real-time feedback if conditions are not met); (3) case evaluation with detection, classification, prostate volume measurement, and recommended PI-RADs (risk) score; (4) report dissemination including generating a T2W image set with overlayed cancer probability index in a DICOM format, a 3D rendering of the prostate and all lesions, an accurate prostate and lesion volume as well as each lesion’s individual PI-RADS score and a PI-RADS case score, that are all; (5) sent back to the original study within the user’s PACS system; and finally (6) deleting the study from the ProstatID database.

The core of the AI process occurs in Function 3, where segmentation, detection and classification are performed (Fig. 3) prior to making any cancer prediction. Following appropriate input quality testing, ProstatID

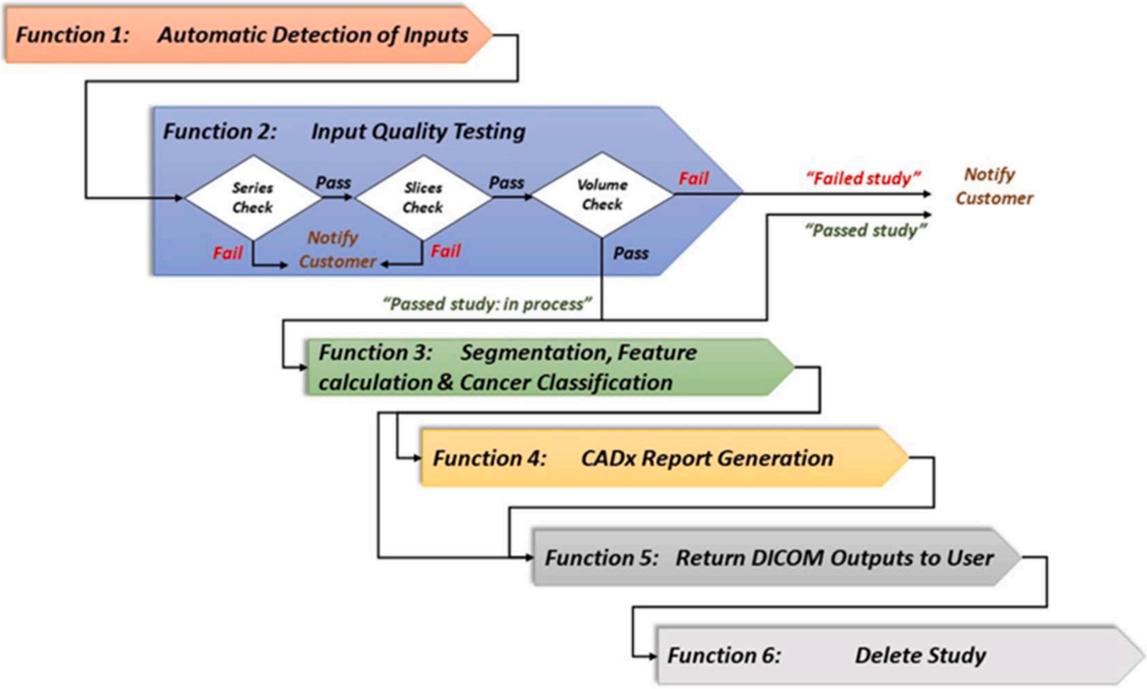
Function 1: Automatic Detection of Inputs
Function 2: Input Quality Testing
Series Check
Pass
Slices Check
Pass
Volume Check
Fail
Fail
Notify Customer
Fail
Pass
“Failed study”
Notify Customer
“Passed study”
“Passed study: in process”
Function 3: Segmentation, Feature calculation & Cancer Classification
Function 4: CADx Report Generation
Function 5: Return DICOM Outputs to User
Function 6: Delete Study

Fig. 2. Diagram of ProstatID workflow.

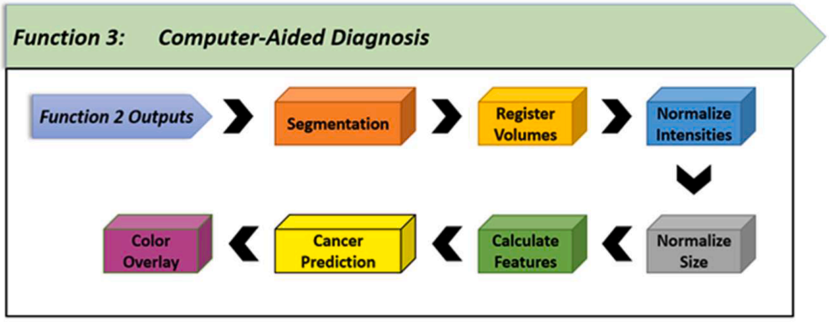

Fig. 3. Diagram breaking Function 3 of ProstatID: segmentation, registration, feature calculation & classification.

first performs image registrations of the following volumes:

- Registration of the ADC volume to the T2w volume using a rigid transformation
- Registration of the DWI volume to the ADC volume using a rigid transformation

This registration step implements the mutual information as a cost metric that is maximized to best align the moving volume with respect to the reference volume.

ProstatID automatically detects prostate anatomy from the T2w MR image of the pelvis. Detection is accomplished by utilizing a 3-D convolutional neural network (CNN) trained to segment the prostate on T2-weighted image volumes. Prostate anatomy is identified by a separate image where all prostate-related voxels are first set to one, and all non-prostate voxels set to zero. The ProstatID software then automatically segments the TZ and CZ as one to differentiate from the PZ. These non-PZ anatomies are identified by setting all related voxels to two.

ProstatID next analyzes the T2-weighted, DWI and ADC MRI data and automatically classifies prostate tissue within the T2-weighted MR image set. Classification is accomplished utilizing an ensemble modelling approach where various base models are used during the process of cancer prediction. The boosted parallel random forest (bpRF – patent-pending) ensemble model (Fig. 4) seeks the wisdom of the crowd and aggregates the prediction of each base model using the majority voting or the weighted average techniques.

In order to make a final and most accurate prediction with less generalization error for the data in which we lack biological ground truth, ProstatID's inference unit acts as a single model and takes into consideration the outcomes of multiple base models including five random forest (RF) models within itself. It was trained on 1538 patient cases, nearly 1000 of those less 150 cases set-aside for testing, with proven cancerous and/or benign diagnoses that were manually annotated/segmented. The classification step operates in a 2-D slice-by-slice fashion, with a set of image features as input to the RF models. These image features are generated from the T2-weighted, DWI, and ADC image data for the regions identified by the segmentation. Tissue is classified, voxel to voxel, on a continuous scale between zero and one, inclusive. A classification of zero indicates a low likelihood of prostate cancer. A classification of one indicates a high likelihood of prostate cancer. RF models were trained for both 1.5T and 3.0T field strengths.

The numerical classification scheme is translated into visual output in the form of a translucent, heatmap-type of colorized probability map, with the probability weighting as indicated in Table 4, with the demonstrated green hue representing 0 % probability of csPCa, a color progression from green to red through yellow indicating increasing probability of csPCa, and the demonstrated red hue 100 % probability of csPCa. This translucent colormap is overlaid onto the axial T2-weighted slices to indicate the relative probability of csPCa within the prostate tissue (Fig. 1). This colorization is intended to assist the physician in following the PI-RADS v2.1 Radiology Interpretation Guide (although in this case the bpMRI pathway as ProstatID does not incorporate the dynamic contrast enhanced (DCE) imaging inclusive in mpMRI) to grade the suspect lesions according to the definitions therein while simultaneously providing AI input into lesion grading. Additionally, the algorithm generates a three-dimensional view of high probability suspect lesions within a translucent volume representation of the prostate as shown in Fig. 5.

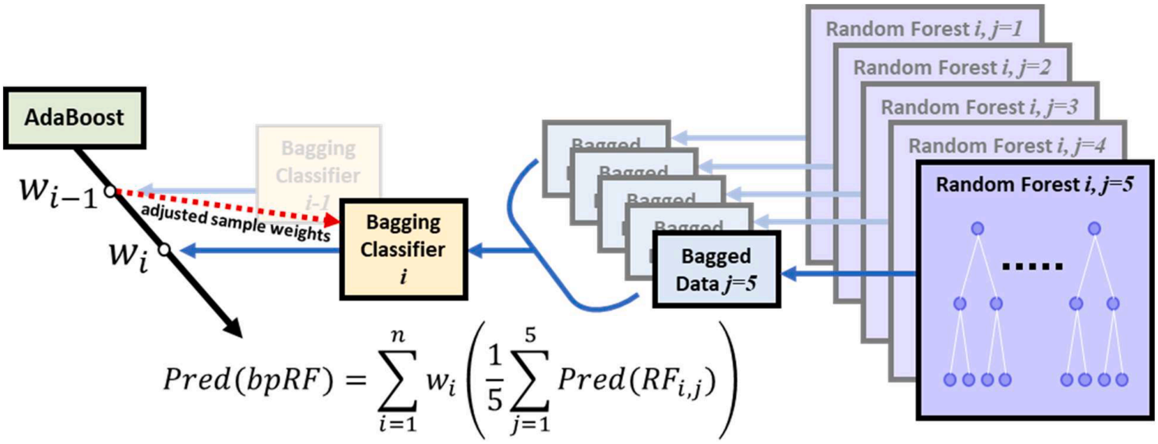

Fig. 4. The boosted parallel random forest (bpRF) model architecture used by ProstatID for this study. bpRF is an ensemble of various base models. The ensemble model seeks the wisdom of the crowd and aggregates the prediction of each base model to make a final prediction with less generalization error. bpRF implements a chain of estimators that starts with an AdaBoost model encapsulating multiple bagging classifiers that are boosted sequentially during training. Each bagging classifier has 5 parallel random forests acting on random slices of data.

Table 4
Colorized Translucent Probability Legend ProstatID algorithm.

Positive Probability (%)	Color Description	Pallet
>62	Red	🟥
>50 & <62	Yellow	🟨
<50	Green	🟩

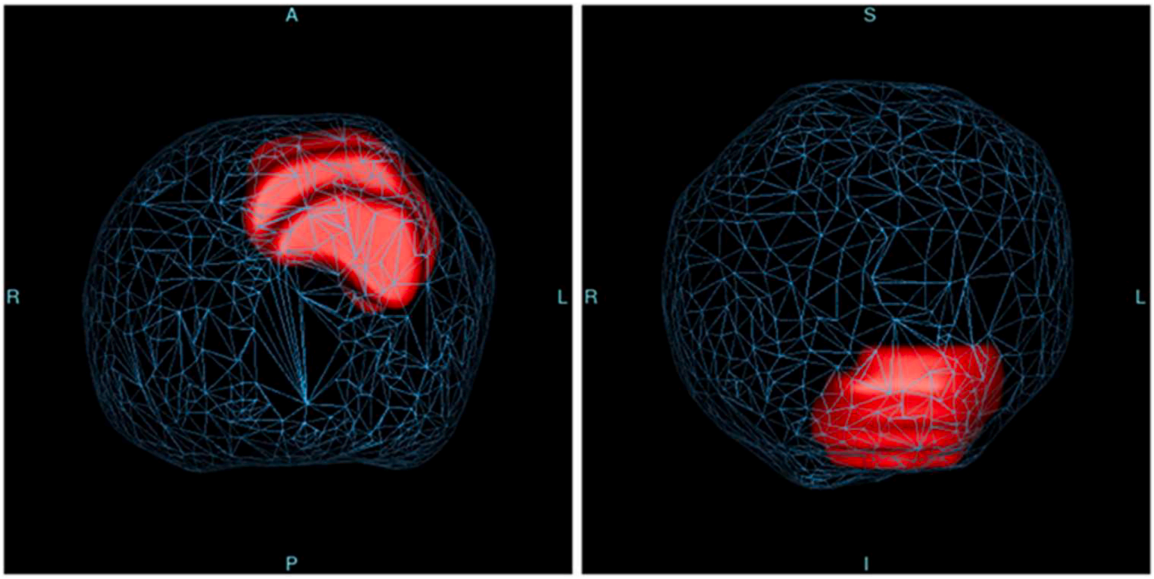

Fig. 5. ProstatID generated 3D transparent outline of the prostate gland with 3D solid tumor outline spatially located within.

Results

Diagnostic performance: reader case level

The reader-averaged ROC curves are shown in Fig. 6 for the unassisted case level read (without CADe/CADx) and the read with CADe/CADx, with unassisted AUC values ranging 0.584 – 0.761 and AI assisted AUC values ranging from 0.637 – 0.799, netting average values of 0.672 and 0.718 respectively as shown in Table 5. This yielded an estimated AI assisted improvement in AUC for rating true positive patients of 0.046, which is statistically significant at the level α = 0.05.

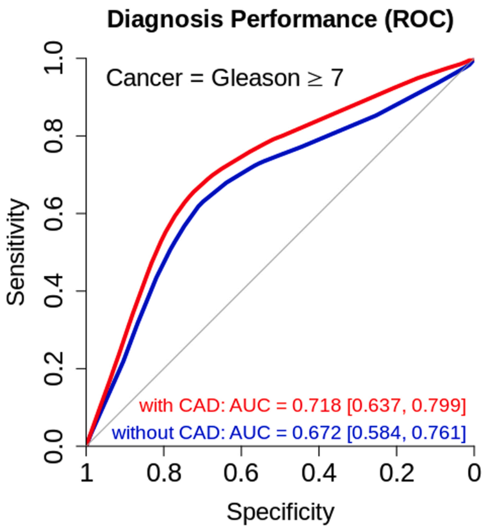

Fig. 6. ROC curve averaged across readers for assigning a case level PI-RADS rating to patients for the unassisted read (without ProstatID) and the read with ProstatID. The difference was statistically significant (α = 0.05).

Table 5
Multiple-reader multiple-case study case level estimate of area under the ROC curve (AUC) for the unassisted read (without CAD) and the read with CAD.

Metric	AUC	95 % CI	p-value
$\text{AUC}_{\text{1st Read}}$ (without CAD)	0.672	[0.584, 0.761]	-
$\text{AUC}_{\text{2nd Read}}$ (with CAD)	0.718	[0.637, 0.799]	-
$\Delta\text{AUC} = \text{AUC}_{\text{2nd Read}} - \text{AUC}_{\text{1st Read}}$	+0.046	[0.010, 0.081]	0.0149

Examining the individual AUC curves for the nine readers, there are clear differences in baseline reader performance (Fig. 7, Table 2), which does not necessarily seem to correlate with years' experience. As can be seen, however, the assistance of AI provided a noticeable improvement in AUC in 7 of 9 readers, with AUC improvements ranging from 0.018 – 0.096. In one reader the AUC was near identical for both, and in another reader there was a slight decrease in AUC with AI assistance (-0.019). Interestingly, both the least (#4) and the most (#6) experienced readers were the two in whom the best AUC improvement were seen (0.09 and 0.10, respectively), with variable but lesser improvement in the other readers.

Lesion-level detection performance without and with CAD

The free-response ROC (FROC) curves for readers at all biopsied sites without and with use of CADe/CADx are shown in Fig. 8. The use of AI yielded an increase in performance, with θ increasing from 0.405 to 0.453. This was statistically significant at the level α = 0.05 (Table 6). As stated above and elaborated in Part I,[1] evaluating reader performance at the lesion level may not in general be practical unless biopsy points are taken at all suspicious regions. Only then would those many points

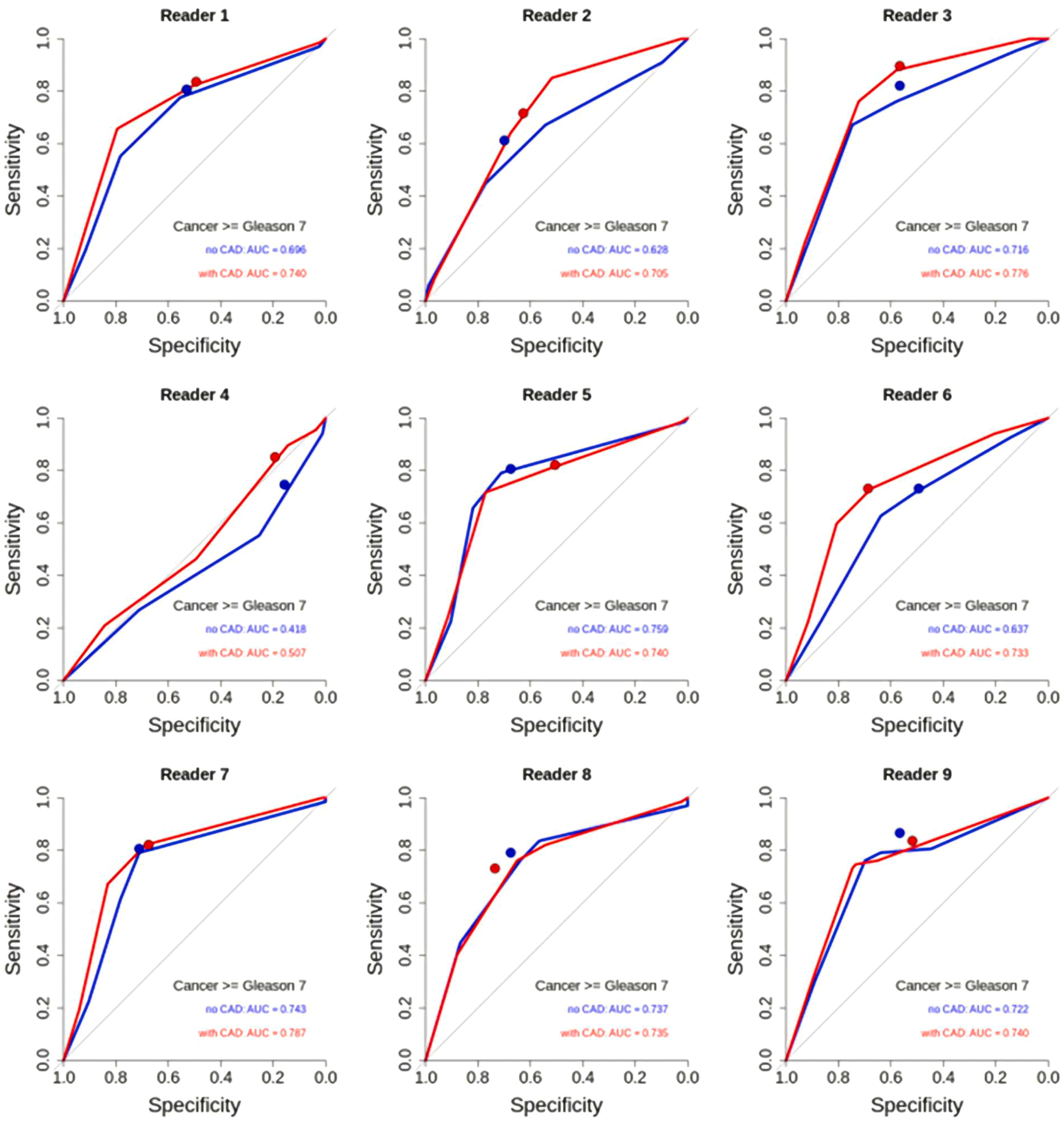

Fig. 7. Individual reader case level ROC curves from the clinical study. Blue represents the unassisted read and red represents the read using ProstatID. The red and blue dots indicate the sensitivity/specificity of the readers' decision to biopsy MR suspicious lesion. AUC values are trapezoidal area under the ROC curve. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

represent an appropriate level of mixed positive and negative ground truth points to which the readers and/or AI could be evaluated.

Diagnostic performance: standalone AI all tissue

The standalone performance of the ProstatID CADe/CADx algorithm across all prostate tissues is shown in Fig. 9. This curve was constructed using predictions from the 3 mm grid method which we concluded was a method that most closely approximates human interpretation for detection and localization of lesions. This result most directly compares with the case level reader results in Figs. 6 and 7. The standalone AUC of 0.929 shows that the AI has an ability to distinguish cancerous tissues from all negative tissues (consisting of normal tissue and benign lesions) superior to all radiologists measured in this study (where best AUC was 0.759).

As was done for the readers, the standalone FROC performance of ProstatID was determined at all biopsy sites and is shown in Fig. 10. When compared to the readers' unassisted read (Fig. 8), ProstatID performed significantly better at detecting and rating cancerous lesions, with a θ of 0.706 vs. 0.405 ($\Delta\theta = +0.301$, statistically significant at the level $\alpha = 0.05$).

Diagnostic performance: standalone AI lesion level

The standalone lesion level performance of the AI at all biopsied locations, retrospectively provided by the panel of experts along with their biopsy results, of the same clinical data set is shown in Fig. 11, with an AUC of 0.710. This demonstrates that the AI has good ability to distinguish cancerous from benign tissues.

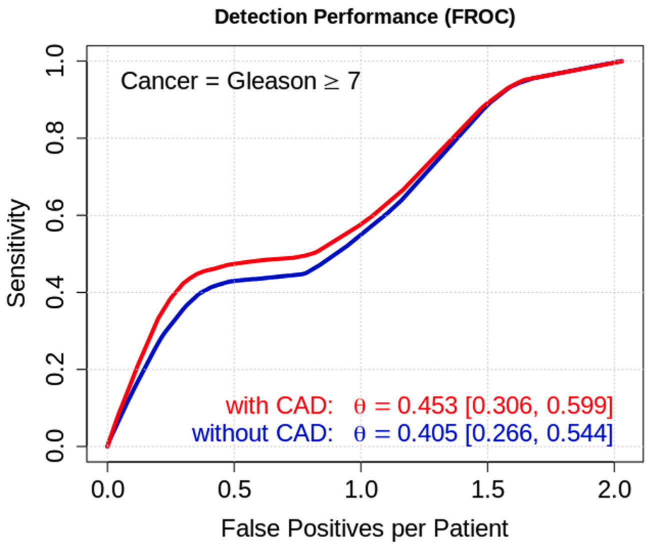

Fig. 8. Free-Response ROC (FROC) curves at all biopsy data points averaged across readers for detecting clinically significant cancerous lesions for the unassisted read (without ProstatID) and the read with ProstatID. The difference was statistically significant (α = 0.05).

Table 6
Free-response ROC (FROC) results at all biopsy points for the unassisted read (without CAD) and the read with CAD.

Metric	θ	95 % CI	p-value
$\theta_{\text{1st Read}}$ (without CAD)	0.405	[0.266, 0.544]	-
$\theta_{\text{2nd Read}}$ (with CAD)	0.453	[0.306, 0.599]	-
$\Delta\ \theta = = \theta_{\text{2nd Read}}$ - $\theta_{\text{1st Read}}$	+0.048	[0.007, 0.088]	0.024

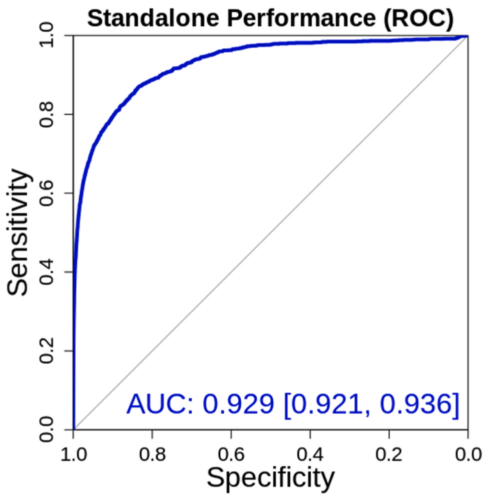

Fig. 9. ROC curve for the ProstatID output evaluated at all pixels (whole prostate, modified case level) using the grid method described.

Reduction of biopsies of benign tissues

The model accuracy was 84.3 % using the mixed effects model for predicting the decision to biopsy. The readers were 1.06 times more likely to biopsy a benign lesion when not using ProstatID, however, the interaction of using ProstatID was not considered significant (p = 0.590). The readers were 1.30 times more likely to biopsy a cancerous lesion when using ProstatID, however, the interaction of using ProstatID was only marginally significant (p = 0.058). Therefore, utilizing ProstatID achieved a 6 % decrease in biopsies of benign tissues and a 30 % increase in biopsies of malignant tissues (detection of csPCa).

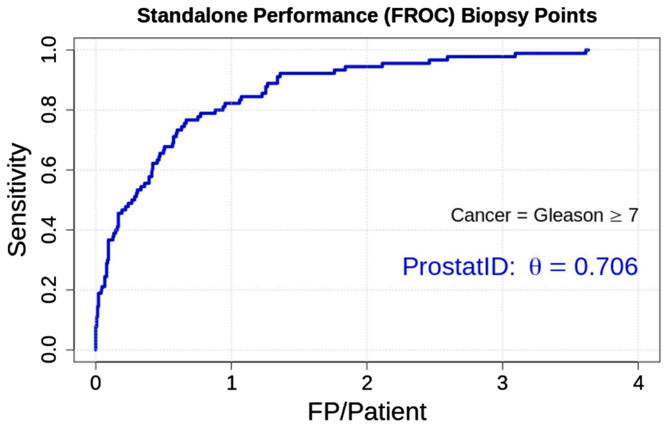

Fig. 10. Corresponding Free-Response ROC (FROC) curves at all biopsy points for the ProstatID algorithm prostate cancer prediction model, comparable to the average FROC of the readers (Fig. 8). The jackknife alternate FROC (JAFROC) performance metric (θ) is shown.
Note: The alternative FROC (AFROC) method of plotting is used so that both curves (Figs. 8 and 10) scale between zero and one on the y-axis. When compared to the readers' unassisted read, ProstatID performed better at detecting and rating clinically significant cancerous lesions (Δθ = +0.301). This difference in performance is statistically significant at the 5 % level (p = 0.029).

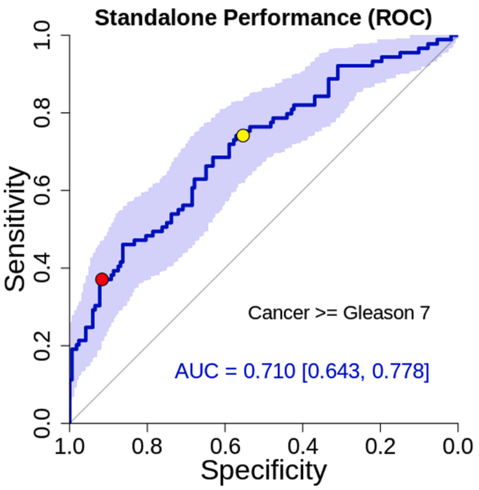

Fig. 11. ROC curve for the ProstatID output evaluated at all biopsy locations (lesion level). The 95 % confidence bounds (shaded) were determined using bootstrapping. The yellow and red dots represent the operating points for the start of the yellow and red color in the AI's color probability map, respectively (Fig. 1, Table 4). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Discussion

The clinical study presented in this Part II manuscript describes a multiple-reader multiple-case (MRMC) design using an enriched dataset of 150 cases from 6 sites, encompassing 12 MR machine types at both 1.5T and 3T, with 9 readers of varying experience from 6 different sites, ultimately evaluating reader diagnostic performance without and with use of a CADe/CADx device. While some of the fundamental results of this study have recently been published elsewhere,[2] a compelling objective of this manuscript was to utilize this MRMC data as a blueprint for implementing the recommendations regarding the construction and reporting of prostate AI studies as proposed in the companion Part I review.[1] While Part I was intended to explore the present landscape of prostate AI and discuss how to evaluate and interpret the performance of such algorithms, this clinical study attempts to put the Part I recommendations into practice. It furthermore expands on some of the clinical data, particularly evaluating AI impact on decision to biopsy and providing more complete details of the AI algorithm. As CADe/CADx technologies increasingly find their way into the reading room, we believe it is essential that end users be equipped with the background to understand the advantages and limitations of these CADe/CADx tools in order to report and compare studies in a rigorous and standardized fashion.

One important, but often confusing principle we emphasize in Part I is the distinction between reporting results on a "case level" or "lesion level" detection basis. The former is more representative of clinical case level patient diagnosis, and for a radiologist interpretation is typically the highest PI-RADS score for that patient; this is how reader performance is typically evaluated in the literature. In this study design, the reader case level diagnosis was evaluated both with and without the reader utilizing CADe/CADx assistance. True lesion level analysis, on the other hand, analyzes only identified lesions where truth exists. This requires *a-priori* knowledge of where the known lesions are, which can be achieved in practice when taking multiple biopsy points from a patient, possibly some targeted and other via systematic core biopsies. The resulting biopsy truth data can then be used to evaluate a reader based upon the methods described above and in Part I. Thus a true and complete lesion level analysis of AI and readers is typically only performed through analysis of all known lesions where truth exists, as we have done using the FROC method. As highlighted in Part I, such an AI lesion level analysis results in a very different and almost certainly less impressive curve (FROC) as compared to the ROC curve analysis. Note also that the grid method of obtaining a ROC analysis of the AI is comparable to the ROC analysis of readers evaluated on a case basis with ground truth biopsy data. We strongly maintain, however, that lesion level analysis is the best and most stringent way for researchers and vendors to report and compare standalone cancer detection results.

Addressing the primary study goal of whether the ProstatID algorithm improved radiologist diagnostic performance, case level analysis was performed for all 9 readers without and with use of the ProstatID algorithm, with an additional FROC analysis at each of the 9 readers selected biopsy points. Without ProstatID, mean ROC for the 9 readers was 0.672. This is somewhat less than seen in two recent PI-RADs v2 meta-analyses, which arrived at values of 0.86-0.87.[3,4] These differences may relate to the known large variability in PI-RADS reads,[5,6] the fact that greater than 40 % of the dataset used for this study was at 1.5T, which is known to be of lower sensitivity than 3T,[3] and the fact that only bi-parametric MR was used. Regardless, case based ROC analysis demonstrated statistically significant improvement in reader performance with the use of ProstatID, with mean ROC increasing to 0.761. Similarly, FROC analysis at the biopsy data points demonstrated a statistically significant improvement with the addition of CADe/CADx.

These are important and exciting results, suggesting such AI algorithms can be easily incorporated into the reading room to improve diagnostic accuracy. Consider also that all readers had no prior experience with ProstatID, as this tool was simply introduced as an additional set of images (cancer probability colormap overlaid on T2 images) that could be utilized as the reader saw fit in their review of the MR images in order to help arrive at a diagnosis. Thus there was no attempt to determine how or when the radiologist decided to let the ProstatID information influence their interpretation, however it is clear that it did so frequently enough to achieve positive statistical significance.

As might be expected, there was a variable degree of improvement in case level AUC with the use of ProstatID across the readers, with 7 of 9 demonstrating improvement, one being unchanged, and one showing a minimal decrease in AUC. This may perhaps relate to how much trust each reader placed in ProstatID, and it would be interesting to know how performance evolves as readers gain more experience with ProstatID. It is interesting to note that both the readers with the least and most experience had the largest increases in AUC, suggesting that algorithms such as ProstatID may be able to improve diagnostic accuracy among across experience levels, one of the expectations for AI.[7] In a similar vein, Litjens [8] performed a sub-analysis based on reader experience, also concluding that both more and less experienced readers achieved similar gains with AI.

Other comparable studies examining whether AI can improve radiologist's reads include a multiple-reader, multiple case study by Winkel et al. [9] evaluating 7 readers using 100 PROSTATEx Challenge cases [10] without, and then with AI assistance. Their AUC demonstrated a statistically significant increase from 0.84 to 0.88 – similar in magnitude to that seen in this study. These authors also saw an increase in inter-reader agreement when incorporating AI. A somewhat different approach to this concept [8] evaluated 7 radiologist's PI-RADS scoring alone versus a "combination" score that was a mathematical combination of PI-RADS score and a continuous likelihood score between 0 and 1 as determined by an AI algorithm for each lesion. This yielded an AUC increase from 0.78 – 0.88. Note that both of these studies were performed exclusively at 3T.

Examining next the secondary goal of ProstatID standalone performance, the all tissue AUC was 0.929, which at first glance appears extremely good. Recall, however, that such an analysis is heavily biased by the typically large volume of normal tissue as compared to the much smaller volume of abnormal tissue, and thus such seemingly high AUC values can be misleading and must be interpreted cautiously. Nevertheless, this provides the most direct comparison with case level reader results, where the best AUC among our readers was 0.759, demonstrating ProstatID has an ability to distinguish benign from cancerous tissues superior to radiologists. Examining the lesion level standalone performance of ProstatID, which we advocate as the best and most stringent way to evaluate AI, AUC was 0.710. This demonstrates that AI has good ability to distinguish cancerous from benign tissues, bearing in mind that the biopsy locations were those areas deemed suspicious enough by expert radiologist Panels to recommend for biopsy; hence all negative biopsies were false positives of those contributing radiologists. Additionally, the expert panel concurred that all tissues NOT biopsied were normal or tissue not sufficiently suspicious to warrant biopsy. It was this data provided by the contributing sites that also provides performance of the contributing radiologists.

It is difficult to compare the ProstatID standalone performance to other published AI algorithms, in part due to the heterogeneity of the datasets and techniques evaluated, and in part because studies often do not report basic facts such as whether reported AUC is on a case level or lesion level. A recent meta-analysis screened 392 citations to arrive at 12 "relevant and pertinent" machine learning studies for prostate cancer,[11] of which only 4 were deep learning and demonstrated a pooled AUC of 0.78 (95 % confidence interval 0.69 – 0.86). Examining these 4 studies more closely, one 364 patient study achieved an AUC of 0.91,[12] however the authors provide little information on the statistical details, and this may very well be a case level analysis. Another study of 140 patients achieved an AUC of 0.726.[13] This was a "classification" study where the lesions were fed into the AI, without algorithm "detection". Given this methodology, this appears to be a lesion level analysis, but again this is not clearly stated. Another approach,[14] performed on an open access training dataset (Cancer Imaging Archive PROSTATEx2$_{\text{train}}$) and also not a detection algorithm demonstrated an AUC of 0.81. Again, this likely reflects a lesion level analysis, however this is not explicitly clear. The final study in this series, another non "detection" algorithm that was part of the PROSTATEx Challenge,[15] achieved an AUC of 0.73, again presumably but not clearly stated lesion level analysis. Of note, none of these studies performed a FROC analysis. This heterogeneity in how the AUC is determined and reported is one of the issues that needs standardization in order to more effectively compare standalone AI performance across different techniques.

The other secondary goal was to determine how ProstatID impacted the incidence of reader request for biopsy. For this determination there was no explicit guidance to the readers, and it was essentially a subjective call as to whether they considered a lesion suspicious enough to recommend biopsy. Such decisions may perhaps also integrate factors such as PSA level and image artifacts, however the call was left to the reader, and to that effect can perhaps be considered a metric of certainty in which the evaluating radiologist trusts their diagnosis. While this is not something we have seen evaluated in similar studies, the use of ProstatID did increase the incidence of readers requesting a biopsy of malignant tissue by 30 %, which nearly achieved statistical significance. Conversely, there was no statistical significance in the 6 % decreased incidence of requesting biopsy of benign tissue. This suggests that difficult to assess or missed regions that ProstatID identifies as suspicious are upgraded in suspicion by the reader, and this may be a large factor contributing to the increase in AUC for reads performed with ProstatID. It is unknown whether increased reader experience with ProstatID would further improve diagnosis and increase the incidence of requesting biopsy of malignant tissue as the algorithm were perhaps more trusted, and this might be an interesting next step to evaluate.

There are several weaknesses of this study. First, only bi-parametric MR data was used, which some studies indicate decreases sensitivity.[7] This largely relates to the added complexity of incorporating contrast enhanced imaging into the AI model, something that is currently in development for the ProstatID algorithm. Not requiring gadolinium contrast, however, does have the advantages of decreased risk of adverse contrast reactions and potential for gadolinium deposition, less cost, and time savings.[7] Second, this was a relatively small study of only 150 cases. This is, however, a typical size for similar studies. There is also the issue of "washout", as the same studies were interpreted twice by all readers. There was a minimum 30 day "washout period" between the two reads, and the cases were presented in a different order, but it is possible that memory of the cases impacted results. Another concern relates to the possibility of a false negative biopsy in an MR suspicious area due to biopsy targeting error. While such false negatives are always possible, we believe that our exclusive use of in bore MR or MR/US fusion biopsy in this study provides as precise as is physically possible localization. Yet another issue is that of proving "negative" tissue. This pitfall plagues all such studies, as truly defining negative is an impossible task. Without complete explant pathology, it can never be known that tissue called negative by the reader or algorithm is truly negative. As a best attempt to circumvent this, our "normal cohort" population was a consensus opinion by an expert panel of 4 radiologists that a non-suspicious study was in fact negative. Next, only a single FDA cleared algorithm was evaluated (ProstatID), without any comparison to other available algorithms. Finally, algorithm gridding was only performed at 3 mm, and as alluded to, there may very well be differences in performance depending on this level of granularity.

Future goals include refinement and continued training of the ProstatID algorithm, including the incorporation of contrast enhanced data into the algorithm. Additional studies on this same dataset, as well as on other curated and perhaps standardized datasets using a refined ProstateID algorithm along with other FDA cleared algorithms will allow for comparison between AI approaches, and provide important information about incorporating such algorithms into clinical workflow.

Conclusion

This multiple case, multiple reader study evaluated reader performance for the diagnosis of clinically significant prostate cancer, both without and with use of the newly FDA cleared CADe/CADx system ProstatID. The addition of this AI tool into radiologist workflow lead to significant improvement in reader performance based on ROC and FROC analysis, as well as increasing the incidence of requesting biopsy of malignant foci by 30 %. This suggests AI augmentation of prostate interpretation improves diagnostic accuracy and will likely play an increasing role in the diagnosis of prostate cancer with MRI. While these results are important and compelling, a fundamental goal of this work was to incorporate and model the Part I recommendations [1] regarding how to construct, analyze and report prostate AI studies so that they can be better and more easily evaluated and compared.

References

1. Maki JH, Patel NU, Ulrich EJ, Dhaouadi J, Jones RW. Part I: Prostate Cancer Detection, Artificial Intelligence for Prostate Cancer and How We Measure Diagnostic Performance: A Comprehensive Review. Curr. Probl. Diagn. Radiol. https://doi.org/10.1067/j.cpradiol.2024.04.002.
2. Anderson MA, Mercaldo S, Chung R, et al. Improving prostate cancer detection with MRI: a multi-reader, multi-case study using Computer-Aided Detection (CAD). *Acad Radiol.* 2022. https://doi.org/10.1016/j.acra.2022.09.009.
3. Zhang L, Tang M, Chen S, et al. A meta-analysis of use of Prostate Imaging Reporting and Data System Version 2 (PI-RADS V2) with multiparametric MR imaging for the detection of prostate cancer. *Eur Radiol.* 2017;27(12):5204–5214. https://doi.org/10.1007/s00330-017-4843-7.
4. Zhen L, Liu X, Yegang C, et al. Accuracy of multiparametric magnetic resonance imaging for diagnosing prostate Cancer: a systematic review and meta-analysis. *BMC Cancer.* 2019;19(1):1244. https://doi.org/10.1186/s12885-019-6434-2.
5. Rosenkrantz AB, Ginocchio LA, Cornfeld D, et al. Interobserver reproducibility of the PI-RADS version 2 lexicon: a multicenter study of six experienced prostate radiologists. *Radiology.* 2016;280(3):793–804. https://doi.org/10.1148/radiol.2016152542.
6. Turkbey B, Rosenkrantz AB, Haider MA, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *Eur Urol.* 2019;76(3):340–351. https://doi.org/10.1016/j.eururo.2019.02.033.
7. Turkbey B, Haider MA. Artificial intelligence for automated cancer detection on prostate MRI: opportunities and ongoing challenges, from the AJR special series on AI applications. *Am J Roentgenol.* 2022;219(2):188–194. https://doi.org/10.2214/ajr.21.26917.
8. Litjens GJS, Barentsz JO, Karssemeijer N, et al. Clinical evaluation of a computer-aided diagnosis system for determining cancer aggressiveness in prostate MRI. *Eur Radiol.* 2015;25(11):3187–3199. https://doi.org/10.1007/s00330-015-3743-y.
9. Winkel DJ, Tong A, Lou B, et al. A novel deep learning based computer-aided diagnosis system improves the accuracy and efficiency of radiologists in reading biparametric magnetic resonance images of the prostate: results of a multireader, multicase study. *Invest Radiol.* 2021;56(10):605–613. https://doi.org/10.1097/rli.0000000000000780.
10. Litjens G, Debats O, Barentsz J, et al. "ProstateX challenge data", the cancer imaging archive. Published 2017. Accessed August 15, 2022. https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=23691656.
11. Cuocolo R, Cipullo MB, Stanzione A, et al. Machine learning for the identification of clinically significant prostate cancer on MRI: a meta-analysis. *Eur Radiol.* 2020;30(12):6877–6887. https://doi.org/10.1007/s00330-020-07027-w.
12. Le MH, Chen J, Wang L, et al. Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. *Phys Med Biol.* 2017;62(16):6497–6514. https://doi.org/10.1088/1361-6560/aa7731.
13. Zhong X, Cao R, Shakeri S, et al. Deep transfer learning-based prostate cancer classification using 3 Tesla multi-parametric MRI. *Abdom Radiol.* 2019;44(6):2030–2039. https://doi.org/10.1007/s00261-018-1824-5.
14. Abraham B, Nair MS. Computer-aided grading of prostate cancer from MRI images using Convolutional Neural Networks. *J Intell Fuzzy Syst.* 2018:1–10. https://doi.org/10.3233/jifs-169913. Preprint(Preprint).
15. Sobecki P, Życka-Malesa D, Mykhalevych I, et al. EMBEC & NBC 2017. In: *Proceedings of the Joint Conference of the European Medical and Biological Engineering Conference (EMBEC) and the Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (NBC)*. 2017:827–830. https://doi.org/10.1007/978-981-10-5122-7_207. June 2017. *IFMBE Proc.* Published online.