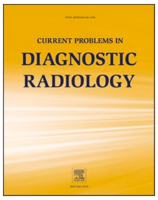
Contents lists available at ScienceDirect

Current Problems in Diagnostic Radiology

journal homepage: www.cpdrjournal.com



CURRENT PROBLEMS IN DIAGNOSTIC RADIOLOGY

Part I: prostate cancer detection, artificial intelligence for prostate cancer and how we measure diagnostic performance: a comprehensive review

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

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

ARTICLE INFO

Keywords:
Artificial Intelligence
CADe
CADx
Prostate Cancer
Study Design

ABSTRACT

MRI has firmly established itself as a mainstay for the detection, staging and surveillance of prostate cancer. Despite its success, prostate MRI continues to suffer from poor inter-reader variability and a low positive predictive value. The recent emergence of Artificial Intelligence (AI) to potentially improve diagnostic performance shows great potential. Understanding and interpreting the AI landscape as well as ever-increasing research literature, however, is difficult. This is in part due to widely varying study design and reporting techniques. This paper aims to address this need by first outlining the different types of AI used for the detection and diagnosis of prostate cancer, next deciphering how data collection methods, statistical analysis metrics (such as ROC and FROC analysis) and end points/outcomes (lesion detection vs. case diagnosis) affect the performance and limit the ability to compare between studies. Finally, this work explores the need for appropriately enriched investigational datasets and proper ground truth, and provides guidance on how to best conduct AI prostate MRI studies. Published in parallel, a clinical study applying this suggested study design was applied to review and report a multiple-reader multiple-case clinical study of 150 bi-parametric prostate MRI studies across nine readers, measuring physician performance both with and without the use of a recently FDA cleared Artificial Intelligence software.[1]

MRI and PI-RADS

Prostate cancer (PCa) is the most common non-cutaneous cancer in men in the United States and the second-leading cause of male cancer deaths. The natural history of this disease, however, is heterogeneous, and there is increasing evidence that many PCa are overtreated.[2] As such, clinical focus has shifted to therapy directed at “clinically significant” (i.e. life-threatening) cancers and active surveillance of more indolent cancers. Multiparametric prostate MRI (mpMRI) — chiefly T2-weighted (T2w), diffusion-weighted (DWI), and high temporal resolution dynamic contrast-enhanced (DCE) sequences — initially developed for loco-regional staging of PCa, has now been recognized as an invaluable tool for tumor detection, localization, characterization, risk stratification, surveillance, assessment of suspected recurrence, and image guidance for biopsy, surgery, focal therapy and radiation therapy of prostate cancer.[3–5] Diagnostically, this has largely manifest as adoption of the international consensus recommendations for Prostate Imaging Reporting and Data System (PI-RADS v2 and v2.1, hereafter referred to as PI-RADS v2), which have proven sensitivity and specificity for detecting clinically significant prostate cancer (csPCa), defined as Gleason score ≥ 7 on pathology/histology, and/or volume ≥ 0.5 mL, and/or extraprostatic extension (EPE).[6,7] As such, PI-RADS v2 has been endorsed by the AUA (American Urological Association)[8] and has gained acceptance by both radiologists and urologists as a preferred method of prostate MRI reporting.[9]

Despite this enthusiasm, there is considerable inter-reader variability in PI-RADS v2 scoring,[10,11] with a substantial learning curve as well as issues with zonal location of lesions.[12,13] A recent study noted that readers missed approximately 16 % of CsPCA lesions at mp MR imaging, while underestimating lesion size in approximately 5 %.[14] Additionally, false positives are commonly encountered. A multicenter study of 3449 men undergoing prostate MRI and MRI-targeted biopsy across 26 centers showed wide variation and an overall low positive predictive value of the PI-RADS v2.1: PPV of 35 % (95 % confidence interval [CI]: 27 %, 43 %) for a PI-RADS score ≥ 3 and 49 % (95 % CI: 40 %, 58 %) for a score ≥ 4.[15] In the PRECISION trial[3] biopsy proven CsPCA was greatest for a

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

https://doi.org/10.1067/j.cpradiol.2024.04.002

Available online 19 April 2024

PI-RADS v2 score of 5 (83 %), followed by a score of 4 (60 %) and score of 3 (12 %). On the other hand, the negative biopsy rate was highest for a PI-RADS v2 score of 3 (67 %), followed by score of 4 (31 %) and score of 5 (6 %).

Artificial intelligence for prostate MRI

Given the challenging nature of correctly identifying and localizing csPCa, as well as the interobserver variability among radiologists, there is clear need to improve our diagnostic paradigm. One such avenue is to utilize artificial intelligence (AI). AI is a generalized term for a variety of techniques performed by machines that make use of prior knowledge, experience, goals and observations to create some sort of desired output, and includes the subsets of Machine Learning and Deep Learning.[16]

As applied to prostate imaging, AI comes in several different forms, the most common of which target improved visualization and workflow enhancement. Multiple vendors offer products that combine AI and other algorithms to segment the prostate and calculate prostate volume and PSA density, register images between different acquisitions and/or planes, fuse different sequences such as diffusion or apparent diffusion coefficient (ADC) with more anatomic (e.g. T2w) images, overlay perfusion curves on other sequences and automatically measure and report user identified regions of concern (e.g. Quantib Prostate, Rotterdam, the Netherlands; MIM Symphony Dx, Cleveland, OH; Philips/DynaCAD, Best, the Netherlands, Ezra Plexo, New York City, NY). Such applications are useful, as accurate lesion segmentation and tumor volume measurement improves radiotherapy outcome, improves fusion biopsy results and are crucial for predicting positive surgical margins, biochemical recurrence, and post prostatectomy survival.[17,18] While certainly important tools, these types of applications do not utilize AI to help detect or classify a lesion.

More advanced AI algorithms are designed to detect, segment and characterize prostate abnormalities, potentially improving interobserver variability in defining lesion borders and identifying smaller satellite lesions, which are almost uniformly missed.[19–21] Per the definitions of Cheng et al. as applied to AI, “classification” refers to assigning an entire image dataset or specific detected lesion either a binary (in our case typically “csPCa” or “non-csPCa”) or multiclass designation.[22] Of note, “characterization” is often used interchangeably with “classification”, although more often refers to a multiclass rather than binary designation (e.g. assigning a PI-RADS 1-5 score is an example of radiologist generated multiclass lesion characterization). “Detection” refers to the identification and localization of the entity of interest, in this case csPCa, with more sophisticated iterations of AI providing true segmentations of the margins of the lesion(s) indicating level of suspicion.[23,24]

It may be obvious that detection and classification are inherently intertwined. In the case of a radiologist reading a prostate MR, lesions are detected based on the perception of groups of imaging voxels, then cognitively processed, typically using classification schemes such as PI-RADS 2.1 to arrive at a combination of location and classification (e.g. PI-RADS 5, highly suspicious for malignancy). Thus the two go hand-in-hand, because detection is based on classification. Turning to AI, a basic pure detection algorithm could simply output a global binary decision or probability of “does this prostate have clinically significant cancer or not?”, however such non-localized (or case level) information is generally less useful. With typical automated lesion detection, the algorithm effectively attempts to voxel-by-voxel classify an image dataset, with the classification being either binary or some form of multiclass designation (e.g. probability scale 0 – 1). If the algorithm is good at detecting cancer and ignoring non-cancerous lesions (high sensitivity and specificity), this inherently means it is good at classifying the underlying voxels, and groupings of similar suspicious voxels can be joined together to define and localize a particular lesion of a certain class or produce an anatomic probability map. Of course, how well either human or machine detection/classification performs can only be determined when there is known biological truth (e.g. correlative pathology), and then typically assessed using metrics from the confusion matrix (e.g. sensitivity, specificity etc.), as each detected lesion is either a true positive or a true negative.

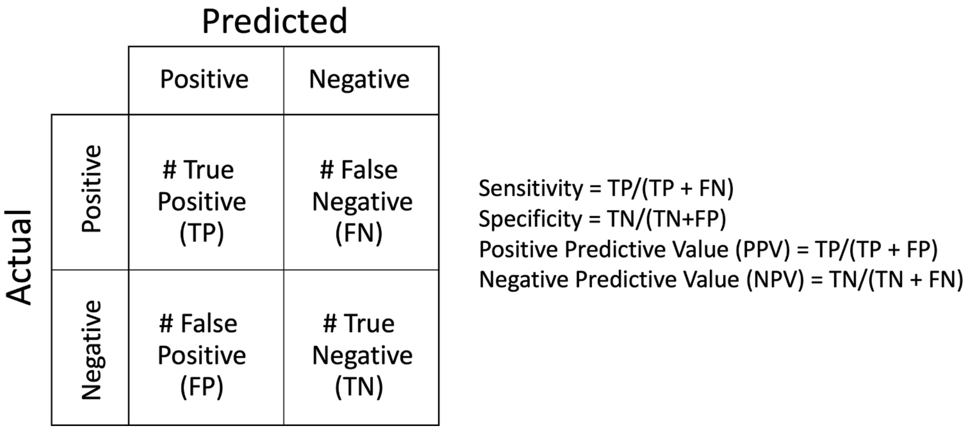
Significant progress has been made on classification and detection using AI, however precise segmentation and associated volume measurement of detected lesions remain problematic, in part because of large interobserver variability, and in part due to a lack of comprehensive correlative imaging and pathologic data.[16] Several recent AI studies using Random Forest or Convolutional Neural Network techniques to detect csPCa have achieved impressive AUC (area under the receiver operator curve) values, ranging from 0.90 – 0.97.[23,25–28] Just what exactly “AUC” is, how it is used and its limitations will be discussed shortly. Ground truth for these studies either used the open source Cancer Imaging Archive prostate imaging with pathologic correlation[29–31] or a combination of experienced radiology readers and biopsy data. How such “ground truth” is established is another important consideration when evaluating algorithm performance. Another diagnostic approach is to use AI detection algorithms in combination with radiologist interpretations (e.g. PI-RADS scoring), with several studies demonstrating that the combination of AI detection algorithms and radiologist interpretation offers significantly improved AUC over radiologist interpretations alone.[21,25,32,33]

AI software, as well as more simple CAD image display software applications, are both considered “Medical Devices”. These are classified by the FDA in increasing order of potential risk as Class I - III, and as such require appropriate FDA clearance. The visualization and workflow focused algorithms as previously described belong to Class II, and are designated “Medical Image Management and Processing Systems” by the FDA Code of Federal Regulations 892.2050, described in part as “… advanced or complex image processing functions for image manipulation, enhancement, or quantification …”. With regards to the more “classification and detection” oriented AI algorithms, the FDA recognizes two designations; CADe (computer-aided detection used in concurrent interpretation) and CADx (computer-aided diagnosis of diseases and their severity). Unless these algorithms have an equivalent predicate, they are typically Class III and must go through what is called the *de novo* certification pathway[34] which involves careful scrutiny by the FDA, including rigorous performance tests to prove their effectiveness.

Measuring performance in prostate cancer detection and diagnosis

Confusion matrix and measurement accuracy

The literature is rife with different metrics for describing the performance of radiologists or AI algorithms in diagnosing prostate cancer, and it is important to understand how these are generated and what their limitations are. The most basic assessment as applied to the binary question of “is there cancer” defines “diagnosis”, which is at the case level. This can be assessed based on the confusion matrix (Fig. 1), a 2 × 2 table with the number of actual positive and actual negative cases on one axis, and the number of predicted positive and predicted negative cases on the other axis. As can be seen, the four boxes then contain the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). From this, the standard statistical measures of sensitivity (or true positive rate), which is the TP divided by all positives = TP/(TP + FN), and specificity (or true negative rate), which is the TN divided by all negatives = TN/(TN+FP), can be calculated. Many studies report sensitivity and specificity results for their readers or their AI algorithms, and in general a “good” reader or AI algorithm has a combination of high sensitivity and high specificity. Other metrics, such as positive and negative predictive value (PPV, NPV) are also often reported in studies, again derived from the confusion matrix, with PPV being TP divided by all those called positive = TP/(TP + FP) and NPV being TN divided by all those called negative = TN/(TN + FN).



		Predicted	
		Positive	Negative
Actual	Positive	# True Positive (TP)	# False Negative (FN)
	Negative	# False Positive (FP)	# True Negative (TN)

Sensitivity = TP/(TP + FN)
Specificity = TN/(TN+FP)
Positive Predictive Value (PPV) = TP/(TP + FP)
Negative Predictive Value (NPV) = TN/(TN + FN)

Fig. 1. Confusion Matrix and the definitions for sensitivity, specificity, positive and negative predictive value.

ROC analyses

Applying such metrics may seem straight-forward, however there are numerous nuances to consider. First, how do we make the binary decision "positive" or "negative"? In fact, PI-RADS does NOT make such a decision as it is basically a five-point scale describing the likelihood of the patient having csPCa. Assume for a moment we call only PIRADS 5 lesions "positive". Provided we have ground truth (e.g. biopsy, explant pathology), we can then calculate sensitivity and specificity. Alternatively, we could choose to call PIRADS $\geq$ 4 lesions positive, yielding a different sensitivity and specificity, and so on for PIRADS $\geq$ 3 lesions and $\geq$ 2 lesions. By doing this, we arrive at four different sensitivities and specificities corresponding to the different PIRADS thresholds. These values can be plotted on a Receiver Operating Characteristic curve (ROC), with the x axis (1 - Specificity) and the y axis Sensitivity, as shown for hypothetical data in Fig. 2. Note that by calling all PIRADS $\geq$ 2 positive our sensitivity is very high, but at the price of low specificity. Conversely by only calling PIRADS 5 lesions positive, our specificity is high, but our sensitivity suffers. The diagonal line is known as the line of "no discrimination", i.e. purely random; points to the left of this line are better than random, and points to the right worse than random, with top left (0,1) being perfect. With AI, there may be a more continuous variable characterizing the probability of prostate cancer, for example a continuous probability ranging from 0 to 1. Under these circumstances, many more points can be generated to fill-in the ROC curve such that it is smoother and can help to choose the optimum threshold given the desired outcome, or compare between different readers/techniques. A useful and often used numerical measurement of performance well suited to comparison is the "area under the (ROC) curve", or AUC, illustrated as the shaded area in Fig. 2, which ranges from 0 to 1. AUC provides a more global picture of performance across differing thresholds. A perfect score would be 1.0, and the literature often describes prostate AI achieving AUC > 0.9; however, we will further discuss how these values are significantly influenced by the exact mechanism by which the ROC/AUCs were measured; meaning that the AUC values reported cannot be directly compared unless the exact methods are also shared and identical.

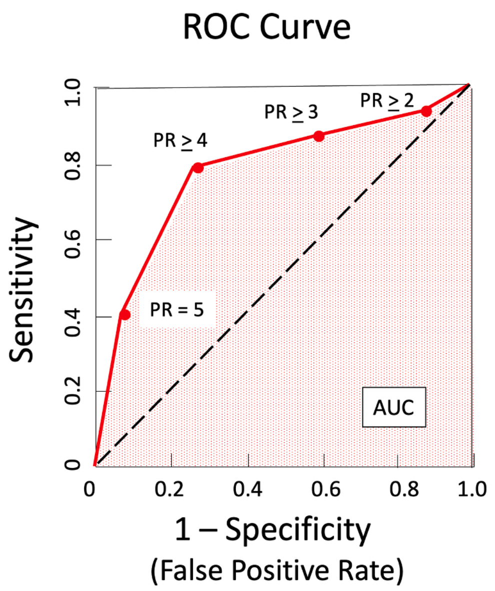


Fig. 2. Hypothetical example Receiver Operating Characteristic (ROC) curve. Area under the curve (AUC) represented by red shading.

PCA case level vs. lesion level diagnosis

Another important consideration when examining how radiologists or algorithms perform has to do with how we define what constitutes a "correct" diagnosis? Considering only a single diagnosis for the case (case level diagnosis), calling a malignant lesion somewhere in the prostate is considered "correct" even if the identified lesion is a false positive and the true malignant lesion is totally missed. On the other hand, we could score each identified lesion individually and use this to determine our performance (lesion level diagnosis). How we chose to do our evaluation has the potential to become even more problematic with AI, which may perform the analysis on a pixel-by-pixel level. Should we somehow try to do the analysis of truth pixel-by-pixel (and is that even possible given what we have for truth?). Do we consider only the highest probability pixel in a "suspicious" region? Or a cluster of higher probabilities of a certain size? These are all different approaches that will lead to different results.

Establishing ground truth and scoring system

All of this introductory information has been provided as background for how to evaluate and analyze a comparative study of radiologists performing PIRADS reads with and without the help of a prostate CADe/CADx system. Given all of the variables discussed, it is clear that one must establish a solid clinical study so as to minimize measurement errors and follow a common standard. We believe the radiology community thus far lacks such guidance that standardizes the methodologies for:

- Measuring a non-continuous diagnostic system such as PI-RADS
- Generating ROC curves with fixed numbers of evaluation points (pixels, 3D grids, PI-RADS sub-regions, etc.)

- Determining how to evaluate reader and software detection of lesions and defining what is truly a “hit” versus “miss” of a lesion based upon the granularity of division of the prostate, or “how you slice it”
- How best to ethically establish and accurately place three-dimensional pathology ground truth points or volumes within the three-dimensional MRI data set
- And perhaps most challenging of all, how to establish whether non-suspicious (unsampled) tissues and patient cases are truly negative for cancer

Although the FDA provides guidance documents on Clinical Performance Assessment of CAD radiologic software, as well as guidance on establishing clinicals studies for Computer-assisted Detection Devices Applied to Radiology Images,[35,36] the detailed methods outlined above are left to the submitting medical device companies. It is therefore highly unlikely that any two radiological CAD devices or peer-reviewed papers discussing the performance of a device can be directly compared because of the significantly different outcomes resulting from the non-standardization of the above methods and procedures.

In a following section we offer alternatives to methods used for evaluation of AI software and suggestions for standardization, with supporting rationale. For instance, there are a multitude of ways in which one can set up an experiment to measure software and physician accuracy in both detection and diagnosis. As an example, simply changing the resolution of three-dimensional localization of specific lesions will impact the ROC curve.

Prostate gland imaging classification

Returning to the concept of a “correct” diagnosis, if we consider the prostate gland to be a single organ as defined by the FDA (CADx devices)[36] it is given a single binary “case level” diagnosis (csPCa or not). This is similar to the intent of PI-RADS whereby the scoring system indicates the likelihood of csPCa within the case score. Alternatively, detection implies localization of some sort, combined with classification or assessment of disease within suspicious lesions (CADe devices) – again as defined by the FDA[36] - and here variations can arise. To evaluate detection accuracy the prostate can be divided into multiple segments, ranging from right vs. left (n=2), to placing a variable sized 3D grid over the prostate (dozens to hundreds), all the way to individually classifying each pixel (thousands). In cases with multiple regions of suspected malignancy, classification might be based on only looking at the single most suspicious lesion such that there is only one correct or incorrect outcome, which again refers to case level diagnosis. Alternatively, it can evaluate any set number of suspected lesions on an individual basis, providing multiple datapoints per case. These datapoints then become the metric for measuring detection accuracy based upon radiologist or AI ability to accurately predict csPCa from the MRI by location.

It follows that how such classification is defined, i.e. how the prostate is sub-divided (or “how you slice it”) and the way in which individual lesions are handled has a large bearing on the results, either as metrics of the confusion matrix such as sensitivity and specificity, or in the AUC or other analysis. As illustrated in Fig. 3, case level diagnosis can be misleading despite apparent good sensitivity and specificity, and could therefore lead to a biopsy of the wrong location. In this example, simply adding minimal localization (right vs. left) dramatically changes the sensitivity and specificity. Hence, evaluation without localization is rather pointless.

Consider now an AI algorithm and two methods of evaluation; the first taking into consideration all points (pixels) in the prostate (“All Tissue”), the second only suspicious lesions for which ground truth biopsy data exists (“Lesions Only)”. As carefully explained in Fig. 4, the ROC curve in this hypothetical case yields an impressive-appearing AUC value of 0.933 for the all tissue analysis, however, the AUC value drops to a most unimpressive 0.512 for the lesions only analysis. This suggests that such a prediction model is unable to distinguish any difference between cancerous and benign lesions, thus providing little benefit to a radiologist. These seemingly discordant results are due to the generally large volume imbalance between normal and benign lesion tissue. So while this hypothetical algorithm is excellent at predicting normal background tissue as negative, of which there is plenty, it is ineffective at predicting benign lesion tissue as negative. This tissue volume imbalance causes the normal tissue performance to overshadow the benign lesion performance. By contrast, the lesion only evaluation more realistically reveals the algorithm’s ability to predict benign lesion tissue as negative. The lesion only evaluation, however, has its own caveat: it cannot account for false positives that may occur in regions of normal tissue. In other words, the reader or AI is only being evaluated at points of suspicion as deemed so by some sort of *a-priori* determination, such as a panel of experts; hence, all other points are ignored. This means that if a reader or AI algorithm is evaluated in this manner, neither would receive a false positive for indicating a suspicious region outside of those known *a-priori*. This simple example demonstrates how different evaluation methods can have a dramatic effect on the conclusion, on even a single image slice from a single case.

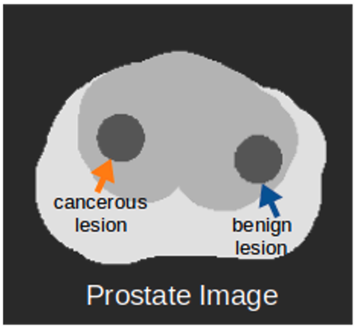


Fig. 3. Example diagram of benign and malignant prostate lesions. Considering a “case level” diagnosis, if a radiologist correctly calls the cancerous lesion positive and the benign lesion negative, this is 100 % sensitivity, 100 % specificity. Conversely, if they call the benign lesion positive and the malignant lesion negative, the case level performance remains 100 % sensitivity, 100 % specificity despite completely incorrect localization. If the prostate were instead divided into left and right segments, the first interpretation again represents 100 % sensitivity and 100 % specificity, while the second interpretation falls to 0 % sensitivity and 0 % specificity.

As an additional nuance, both AI and readers can be evaluated at the lesion level without either being aware of known suspicious lesions that have been biopsied for ground truth; however, in this case the scoring is accomplished by measuring the number of true lesions detected (True Positives) as well as the number of true lesions missed (False Negatives). This can be best characterized by the free response ROC (or FROC) curve, which will be discussed momentarily.

Putting all this into perspective, we firmly believe the best and most stringent approach for evaluating an AI algorithm is a lesion level-based evaluation done at biopsy points, by far the harshest criteria. A related approach, recently used by Lay et al.[23] is a "best-of-both-worlds" evaluation method which systematically divides each slice of the prostate into a grid. A single prediction is calculated for each grid element using a pre-determined method (e.g. mean prediction value within the square), each grid element is compared with ground truth, and an ROC curve is generated. Fig. 5 demonstrates an example of this "grid" method with a 3 mm grid superimposed on a probability map of csPCa. One further point to support the rationale for standardizing a method for measuring ROC/AUC is that if the grid size changes, this nominally affects the results; meaning moving from pixel-by-pixel to larger grid to PI-RADS sub-volumes changes the resulting ROC and its associated AUC. This approach also has a “common sense” attribute of localizing a suspicious lesion within (in this case) a 3 mm x 3 mm x slice thickness volume. Of note, we do not recommend using PI-RADS sub-volumes for AI

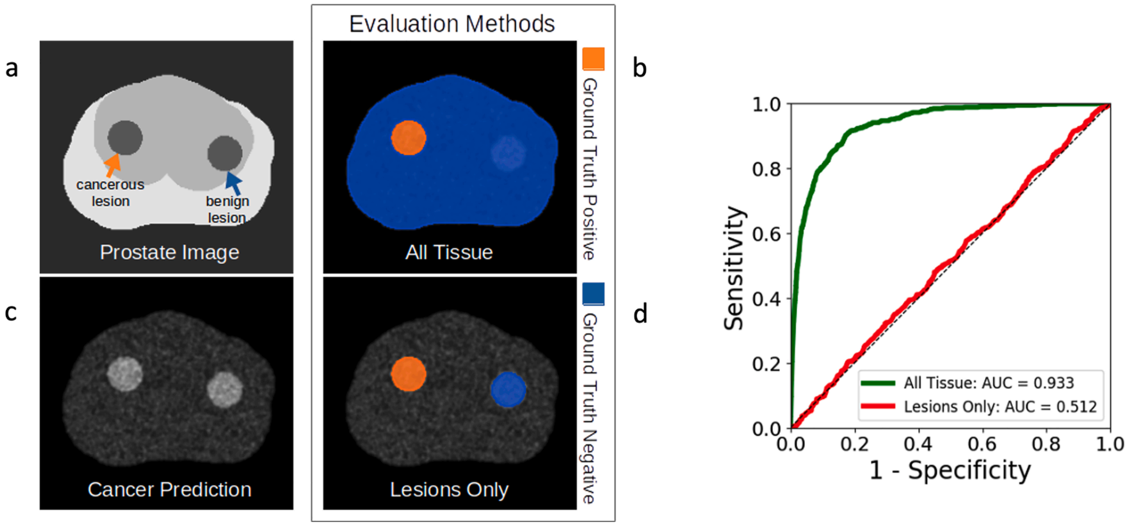


Fig. 4. Depiction of a prostate slice demonstrating two "suspicious" lesions (a), one clinically significant cancer, one benign; both based on ground truth. Hypothetical AI model predicting the likelihood of cancer for each pixel (b), with brighter signal (lighter gray) indicating increased likelihood of cancer (in this example is similar for both lesions), and darker signal in regions representing normal tissue. When the entirety of the prostate is included in the evaluation (all tissue, pixel-by-pixel analysis) (c), the majority of pixels have the correct diagnosis (normal or benign; true negative), with a much smaller number of pixels correctly classified csPCa (orange lesion on left; true positive), and a similar small number of pixels incorrectly classified as csPCa (blue lesion on right, false positive). The resultant (green) (all tissue) ROC curve (e) and its AUC of 0.933 suggests that the cancer prediction model is excellent. If, however, we evaluate only pixels corresponding to suspicious lesions (lesions only) (d), many fewer pixels are evaluated, with approximately 50 % being true positive (orange) and 50 % being false positive (blue). Now only about half of the included voxels are correctly classified, and the (red) (lesions only) ROC curve (e) and its AUC of 0.512 indicate that this prediction model is not able to distinguish any difference between cancerous and benign lesions.

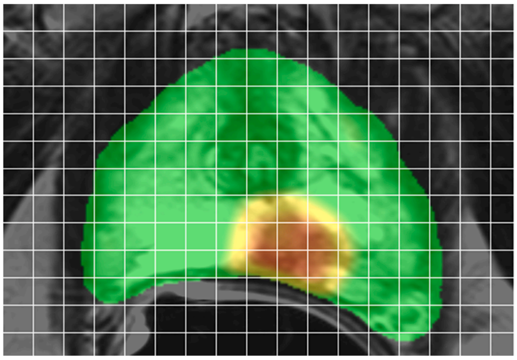


Fig. 5. Example of ProstatID color-coded overlay on a T2w slice through the mid prostate. Color mapping such that green predicts benign tissue, with the color spectrum toward bright red predicting the highest likelihood of clinically significant cancer. Superimposed 3 mm grid.

algorithms because of the many challenges of dividing those sub-volumes as overlayed on each axial slice versus simply overlaying a grid without distinct identification of prostate zonal anatomy.

Prostate gland ground truth

Ground truth defined

In order to score an MRI reader's or AI algorithm's accuracy in detection, one must have a solid ground truth. The FDA loosely defines acceptable ground truth as biological tissue and/or annual follow-up studies of PSA, PSA velocity or MRI.[35] Per scientific consensus, positive ground truth is only defined as those biopsies demonstrating Gleason ≥ 7 histopathology, whereas negative ground truth is defined by those biopsies demonstrating Gleason < 7.[37–39] Furthermore, a third "normal" cohort is defined as those patients considered non-suspicious for clinically significant cancer, such as those with normal PSA and/or digital rectal exam, acceptable PSA velocities and/or negative serial MRIs, and have never had a biopsy.

Thus in order to evaluate reader and/or AI performance, accurately localized ground truth describing whether and where csPCa is present must be known. Perhaps the best "truth" is an explanted prostate gland after prostatectomy, which can be anatomically mapped to correlate with MR imaging.[40–42] But given evolving non-surgical treatments, this is less available, with a larger fraction of pathologic disease confirmation coming from multi-sector trans-rectal US (TRUS) guided biopsy, targeted biopsy or trans-perineal mapping biopsy (TPMB). Regardless of technique, knowing exactly where in the gland the biopsy occurred and exactly how it correlates to the MR image is inexact at best. In general, if a positive biopsy site is believed to be in the same sector of a suspected lesion by MR, that is generally acceptable for "ground truth". Clearly this has limitations, and it is important to recognize these when attempting to quantify radiologist or AI performance at lesion classification.

Proposed guidelines for prostate AI studies

While other more generalized guidelines have been proposed for how to conduct and report medical AI studies[43] a central thesis of this work is to propose a standardized way of conducting and reporting CADe/CADx prostate studies, as no such guidelines exist. Accordingly, and with regard to the nuances of evaluation just discussed, we propose below a detailed standardized methodology geared toward investigators conducting a clinical study to evaluate human readers, an AI algorithm, and human readers assisted by an AI algorithm. Adherence to such guidelines will make performance comparisons between studies more meaningful.

Ground truth

As just discussed, establishing the ground truth for each patient is required to assess the true performance of the readers or an AI algorithm. Thus for MRI-positive patients, we recommend targeted biopsy of suspicious lesions, with an expert then annotating regions corresponding to the biopsied suspicious lesions. This method, however, does have the disadvantage of not confirming tissues outside the biopsied lesions are truly negative. Another acceptable option is the systematic TRUS or trans-perineal biopsy, which by its nature does sample tissues outside the suspicious lesions. Disadvantages of such systematic biopsies are that suspicious lesions may not be adequately sampled, and the exact locations of tissues sampling may be difficult to establish. For MRI-negative patients, we recommend surveillance using PSA or 1-year follow-up MRI to confirm true negative status. To reduce the variation that comes from the subjectivity of MRI interpretation, it is recommended that a panel of experts decide whether a case is MRI-positive or negative. While no established criteria exists for defining such an "expert panel", we suggest this be consensus agreement of at least two radiologists having established and acknowledged expertise in prostate imaging, who are currently interpreting a high volume of prostate MRI exams and are involved with multi-disciplinary teams that review and discuss prostate MR-pathologic correlation. Regardless, the makeup and qualifications of such a panel should be clearly stated.

Patient/case selection

It is important to create and use an "enriched" dataset for analysis, meaning a relatively equal balance of positive and negative cases, randomly sampled from within stratified groups based on a variable such as Gleason score, ISUP grading classification[44] or no csPCa, and containing MRI-negative cases with negative follow-up, MRI-positive cases with negative follow-up, and MRI-positive cases with positive follow-up. The data should ideally come from multiple sites and multiple MR machine types. These factors allow for easier post-analysis of biases and can better assess clinical performance across a spectrum of disease. This dataset needs to be carefully correlated with pathologic data and/or negative follow up data, with a description of how this correlation is performed and defining what is truly a "hit" versus "miss" of a lesion based upon the granularity of division of the prostate.

Performance evaluation

For evaluating the performance of human readers as well as AI, we recommend employing two methods; a modified "case-level" evaluation using PI-RADs scoring for diagnostic performance, and a "lesion-level" evaluation for detection performance.

Diagnostic performance

We recommend a multiple-reader, multiple-case (MRMC) ROC analysis using a modified "case-level" evaluation, meaning one patient case receiving one prediction employing the PI-RADS scoring system wherein each suspect lesion is evaluated and the case PI-RADS score is represented by the highest suspect lesion PI-RADS score. For this to be statistically meaningful, there must be sufficient numbers of cases with the three cohorts of suspicious positive (biopsy), suspicious negative (biopsy) and normal patients, as well as sufficient number of participating physician readers so that the statistical power is adequate.[45] In the context of prostate MRI, a single PI-RADS rating should be assigned by the reader for each case. An ROC curve can be constructed for each reader using the PI-RADS prediction and ground truth (as in Fig. 2). To compare two different reading methods (e.g., a read using standard PI-RADS v2 methods vs. a read assisted by AI), the difference in ROC AUC can be calculated and tested for statistical significance. Although this method is limited from the generalization affect described above using simple case scores, it does employ the ground truth data of all biopsied lesions, and therefore a truthed score sheet and case score with which to compare. More importantly, measuring changes in AUC for reader performance without and with the use of AI will demonstrate the usefulness of the AI.

Detection performance

The ability to detect a lesion is different from the ability to diagnose a patient. Thus, the method for evaluating detection performance should likewise be different. We suggest for this purpose the free-response ROC (FROC) analysis.[46] FROC is an adaptation of ROC for evaluating the performance of both detection and classification in a free-response system. In context of prostate MRI, the task is the localization and classification of cancerous lesions in the image volume. A single case may have no, one, or multiple cancerous lesions. The FROC curve is a plot of sensitivity versus false positives per patient, exemplified in Fig. 6.

Furthermore, a method is required to capture whether or not a reader "detects" a lesion, along with assigning each detection a cancer likelihood prediction. We recommend the simple approach of collecting categorical location descriptors from the readers for each detected lesion, which is then correlated with ground truth. If ground truth for this location is positive and a reader provides this same description while interpreting the case, then that reader has successfully detected the lesion. During an experiment, readers should provide such location descriptors along with their PI-RADS assessment for each described lesion. A FROC curve can then be constructed for each reader based on ground truth, the reader's description of lesion locations, and the reader's PI-RADS prediction. Similar to ROC analysis comparing two different reading methods, any difference in FROC curves can be calculated and tested for statistical significance.[47] Thus, rather than comparing ROC AUC, the weighted alternative FROC (wAFROC) metric (represented by the variable θ) can be used. The wAFROC metric is a measure of detection performance and can be considered analogous to the area under the ROC curve.

Standalone performance of AI

Because an AI algorithm is intended to mimic the abilities of the human analyst, the standalone performance of the AI must also be evaluated. Referring back to the issues that arise from different evaluation methods as previously discussed (Fig. 4), we recommend an AI algorithm be evaluated by two different ROC analyses methods: 1) an all-tissue analysis and 2) a lesion-level analysis. The all-tissue analysis will reveal the ability of the AI algorithm to distinguish cancerous tissues from all other tissues. An algorithm that produces many false positive predictions in normal tissues will have poor performance in such an analysis. The lesion-level analysis, on the other hand, will reveal the ability of the AI to distinguish cancerous tissues from benign tissues, which is the most difficult task. The lesion-level analysis should be done by evaluating the predictions of the AI at all suspicious lesions. Both of these analyses can be performed using either a pixel-by-pixel or grid-

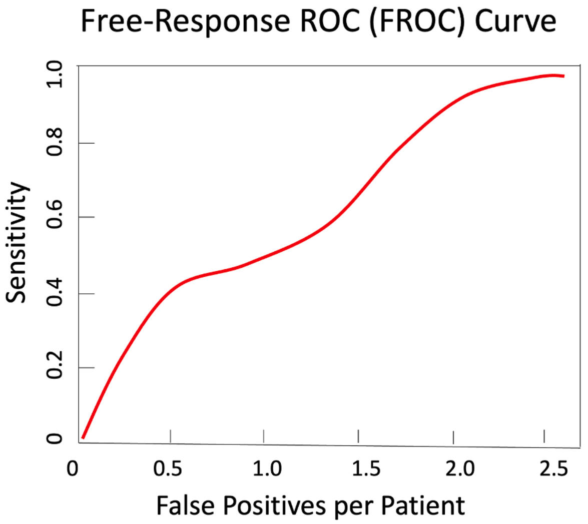


Fig. 6. Hypothetical example Free-Response Receiver Operating Characteristic (FROC) curve. This can be characterized by the weighted alternative FROC (wAFROC) metric θ.

based method, and the specifics of how this was performed should be detailed.

Summary of Proposed Guidelines for Prostate AI Studies

- **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
    - Clearly define the threshold for clinically significant cancer - typically Gleason score $\geq$ 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

Summary

AI applications, including prostate CADe and CADx are finding their way into radiologist's diagnostic workflow. In order to best compare between and utilize these new technologies, it is important to understand how AI results or FDA approved products are reported and analyzed, as there can be considerable variability in study design and implementation, which can in turn greatly impact perceived performance. Furthermore, it is critical that AI researchers use standardized and transparent evaluation methods to allow for more meaningful inter-study comparisons. In this light, this paper addressed many of the more important considerations regarding interpreting, designing and comparing AI studies in order to better prepare the radiologist to put AI algorithms and studies into appropriate context, and provide the AI researcher with a suggested set of standardized research considerations and guidelines.

Grant support

None.

References

1. Maki JH, Patel NU, Ulrich EJ, Dhaouadi J, Jones RW. Part II: Effect of Different Evaluation Methods to the Application of a Computer-Aided Prostate MRI Detection/Diagnosis (CADe/CADx) Device on Reader Performance. *Curr Probl Diagn Radiol.* 2024. https://doi.org/10.1067/j.cpradiol.2024.04.003.
2. Hamdy FC, Donovan JL, Lane JA. 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. *New Engl J Med.* 2016;375(15): 1415–1424. https://doi.org/10.1056/nejmoa1606220.
3. Kasivisvanathan V, Rannikko AS, Borghi M. MRI-targeted or standard biopsy for prostate-cancer diagnosis. *New Engl J Med.* 2018;378(19):1767–1777. https://doi.org/10.1056/nejmoa1801993.
4. Boesen L, Nørgaard N, Løgager V. A prospective comparison of selective multiparametric magnetic resonance imaging fusion-targeted and systematic transrectal ultrasound-guided biopsies for detecting prostate cancer in men undergoing repeated biopsies. *Urol Int.* 2017;99(4):384–391. https://doi.org/10.1159/000477214.
5. Ahmed HU, Bosaily AES, Brown LC. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *Lancet North Am Ed.* 2017;389(10071):815–822. https://doi.org/10.1016/s0140-6736(16)32401-1.
6. ACR. PI-RADS Prostate Imaging – Reporting and Data System 2019 Version 2.1. Published 2019. Accessed August 14, 2022. https://www.acr.org/-/media/ACR/Files/RADS/Pi-RADS/PIRADS-V2-1.pdf.
7. Park KJ, Choi SH, Kim M. Performance of prostate imaging reporting and data system version 2.1 for diagnosis of prostate cancer: a systematic review and meta-analysis. *J Magn Reson Imaging.* 2021;54(1):103–112. https://doi.org/10.1002/jmri.27546.
8. Fulgham PF, Rukstalis DB, Turkbey IB. AUA policy statement on the use of multiparametric magnetic resonance imaging in the diagnosis, staging and management of prostate Cancer. *J Urology.* 2017;198(4):832–838. https://doi.org/10.1016/j.juro.2017.04.101.
9. Spilseth B, Ghai S, Patel NU. A Comparison of radiologists' and urologists' opinions regarding prostate mri reporting: results from a survey of specialty societies. *Am J Roentgenol.* 2017;210(1):101–107. https://doi.org/10.2214/ajr.17.18241.
10. Rosenkrantz AB, Ginocchio LA, Cornfeld D. 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.
11. Turkbey B, Rosenkrantz AB, Haider MA. 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.
12. Stolk TT, de Jong IJ, Kwee TC. False positives in PIRADS (V2) 3, 4, and 5 lesions: relationship with reader experience and zonal location. *Abdom Radiol.* 2019;44(3): 1044–1051. https://doi.org/10.1007/s00261-019-01919-2.
13. Bhayana R, O'Shea A, Anderson MA. PI-RADS versions 2 and 2.1: interobserver agreement and diagnostic performance in peripheral and transition zone lesions among six radiologists. *Am J Roentgenol.* 2021;217(1):141–151. https://doi.org/10.2214/ajr.20.24199.
14. Borofsky S, George AK, Gaur S. What are we missing? false-negative cancers at multiparametric MR imaging of the prostate. *Radiology.* 2018;286(1):186–195. https://doi.org/10.1148/radiol.2017152877.
15. Westphalen AC, McCulloch CE, Anaokar JM. Variability of the positive predictive value of PI-RADS for prostate mri across 26 centers: experience of the society of abdominal radiology prostate cancer disease-focused panel. *Radiology.* 2020;296(1): 76–84. https://doi.org/10.1148/radiol.2020190646.
16. Bardis MD, Houshyar R, Chang PD. Applications of artificial intelligence to prostate multiparametric MRI (mpMRI): current and emerging trends. *Cancers.* 2020;12(5): 1204. https://doi.org/10.3390/cancers12051204.
17. Chung BI, Tarin TV, Ferrari M. Comparison of prostate cancer tumor volume and percent cancer in prediction of biochemical recurrence and cancer specific survival. *Urologic Oncol Seminars Orig Investigations.* 2011;29(3):314–318. https://doi.org/10.1016/j.urolonc.2009.06.017.
18. Lips IM, der Heide UA van, Haustermans K. Single blind randomized Phase III trial to investigate the benefit of a focal lesion ablative microboost in prostate cancer (FLAME-trial): study protocol for a randomized controlled trial. *Trials.* 2011;12(1): 255. https://doi.org/10.1186/1745-6215-12-255. -255.
19. Schie MA van, Dinh CV, Houdt PJ van. Contouring of prostate tumors on multiparametric MRI: Evaluation of clinical delineations in a multicenter radiotherapy trial. *Radiother Oncol.* 2018;128(2):321–326. https://doi.org/10.1016/j.radonc.2018.04.015.
20. Steenbergen P, Haustermans K, Lerut E. Prostate tumor delineation using multiparametric magnetic resonance imaging: Inter-observer variability and pathology validation. *Radiother Oncol.* 2015;115(2):186–190. https://doi.org/10.1016/j.radonc.2015.04.012.
21. Anderson MA, Mercaldo S, Chung R. 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. Published online.
22. Cheng PM, Montagnon E, Yamashita R. Deep learning: an update for radiologists. *Radiographics.* 2021;41(5):1427–1445. https://doi.org/10.1148/rg.2021200210.
23. Lay N, Tsehay Y, Greer MD. Detection of prostate cancer in multiparametric MRI using random forest with instance weighting. *J Med Imaging.* 2017;4(2), 024506. https://doi.org/10.1117/1.jmi.4.2.024506. -024506.
24. Ulrich EJ, Dhaouadi J, Schat R. Comparison of machine learning methods for detection of prostate cancer using bpMRI radiomics features. In: *Proceedings of the 2022 International Society of Magnetic Resonance Imaging.* 2022:2606.

25. Song Y, Zhang Y, Yan X. Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric MRI. *J Magn Reson Imaging*. 2018;48(6):1570–1577. https://doi.org/10.1002/jmri.26047.
26. Sumathipala Y, Lay N, Turkbey B. Prostate cancer detection from multi-institution multiparametric MRIs using deep convolutional neural networks. *J Med Imaging*. 2018;5(4), 044507. https://doi.org/10.1117/1.jmi.5.4.044507.
27. Xu H, Baxter JSH, Akin O. Prostate cancer detection using residual networks. *Int J Comput Ass Rad*. 2019;14(10):1647–1650. https://doi.org/10.1007/s11548-019-01967-5.
28. Lai CC, Wang HK, Wang FN. Autosegmentation of prostate zones and cancer regions from biparametric magnetic resonance images by using deep-learning-based neural networks. *Sensors*. 2021;21(8):2709. https://doi.org/10.3390/s21082709.
29. Litjens G, Debats O, Barentsz J. Computer-aided detection of prostate cancer in MRI. *IEEE T Med Imaging*. 2014;33(5):1083–1092. https://doi.org/10.1109/tmi.2014.2303821.
30. Litjens G, Debats O, Barentsz J. *ProstateX Challenge data", The Cancer Imaging Archive*. et al.; 2017. PublishedAccessed August 15, 2022 https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=23691656
31. Armato SG, Huisman H, Drukker K. PROSTATEx challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *J Med Imaging*. 2018;5(4), 044501. https://doi.org/10.1117/1.jmi.5.4.044501.
32. Wang J, Wu CJ, Bao ML. Machine learning-based analysis of MR radiomics can help to improve the diagnostic performance of PI-RADS v2 in clinically relevant prostate cancer. *Eur Radiol*. 2017;27(10):4082–4090. https://doi.org/10.1007/s00330-017-4800-5.
33. Litjens GJS, Barentsz JO, Karssemeijer N. 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.
34. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digital Med*. 2020;3(1):118. https://doi.org/10.1038/s41746-020-00324-0.
35. FDA. Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Approval (PMA) and Premarket Notification [510(k)] Submissions - Guidance for industry and FDA Staff. Published online 2020. Docket Number FDA-2009-D-0503.
36. FDA. Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions: Guidance for Industry and Food and Drug Administration Staff. Published online July 3, 2012. HHS-0910-2012-F-4294.
37. Kulac I, Haffner MC, Yegnasubramanian S. Should Gleason 6 be labeled as cancer? *Curr Opin Urol*. 2015;25(3):238–245. https://doi.org/10.1097/mou.0000000000000165.
38. Ross HM, Kryvenko ON, Cowan JE. Do Adenocarcinomas of the Prostate With Gleason Score (GS) <= 6 Have the Potential to Metastasize to Lymph Nodes? *Am J Surg Pathology*. 2012;36(9):1346–1352. https://doi.org/10.1097/pas.0b013e3182556dcd.
39. Lepor H, Donan N. Gleason 6 prostate cancer: serious malignancy or toothless lion? *Oncology*. 2014;28(1):16–22.
40. Wildeboer RR, Schalk SG, Demi L. Three-dimensional histopathological reconstruction as a reliable ground truth for prostate cancer studies. *Biomed Phys Eng Express*. 2017;3(3), 035014. https://doi.org/10.1088/2057-1976/aa7073.
41. Rusu M, Shao W, Kunder CA. Registration of presurgical MRI and histopathology images from radical prostatectomy via RAPSODI. *Med Phys*. 2020;47(9):4177–4188. https://doi.org/10.1002/mp.14337.
42. Samavati N, McGrath DM, Lee J. Biomechanical model-based deformable registration of MRI and histopathology for clinical prostatectomy. *J Pathol Inf*. 2012;2(2):10. https://doi.org/10.4103/2153-3539.92035.
43. Mongan J, Moy L, Jr CEK. Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. *Radiology Artif Intell*. 2020;2(2), e200029. https://doi.org/10.1148/ryai.2020200029.
44. Egevad L, Delahunt B, Srigley JR. International society of urological pathology (ISUP) grading of prostate cancer – an ISUP consensus on contemporary grading. *APMIS*. 2016;124(6):433–435. https://doi.org/10.1111/apm.12533.
45. Hillis SL, Obuchowski NA, Berbaum KS. Power estimation for multireader ROC methods an updated and unified approach. *Acad Radiol*. 2011;18(2):129–142. https://doi.org/10.1016/j.acra.2010.09.007.
46. Chakraborty D, Winter L. Free-response methodology: alternate analysis and a new observer-performance experiment. *Radiology*. 1990;174(3 Pt 1):873–881. https://doi.org/10.1148/radiology.174.3.2305073.
47. Chakraborty DP. Recent advances in observer performance methodology: jackknife free-response ROC (JAFROC). *Radiat Prot Dosim*. 2005;114(1-3):26–31. https://doi.org/10.1093/rpd/nch512.