Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI: The PI-CAI Challenge

SUMMARY

Item 1: Title

a) Use the title to convey the essential information on the challenge mission.

Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI

b) Preferable, provide a short acronym of the challenge (if any).

PI-CAI (**P**rostate **I**maging **- C**ancer **AI**)

Item 2: Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Prostate cancer (PCa) is one of the most prevalent cancers in men worldwide. One million men receive a diagnosis and 300,000 die from clinically significant PCa (csPCa) (defined as **ISUP** ≥ 2), each year, worldwide. Multiparametric magnetic resonance imaging (mpMRI) is playing an increasingly important role in the early diagnosis of prostate cancer, and has been recommended by the European Association of Urology (EAU), prior to biopsies (**Mottet et al., 2021**). However, current guidelines for reading prostate mpMRI (i.e. **PI-RADS v2.1**) follow a semi-quantitative assessment that mandates substantial expertise for proper usage. This can lead to low inter-reader agreement (<50%), sub-optimal interpretation and overdiagnosis (**Rosenkrantz et al., 2016**, **Smith et al., 2019**, **Westphalen et al., 2020**).

Unlike the mpMRI protocol, biparametric MRI (bpMRI) does not include dynamic contrast-enhanced imaging (DCE) —thereby reducing costs, eliminating any risk of adverse effects from the use of contrast agents, and shortening examination times (**Turkbey et al., 2019**). Thus, despite providing less diagnostic information than mpMRI (**de Rooij et al., 2020**), bpMRI is more suitable within the scope of high-volume, population-based screening (**Eklund et al., 2021**).

Modern artificial intelligence (AI) algorithms have paved the way for powerful computer-aided detection and diagnosis (CAD) systems that rival human performance in medical image analysis (**Esteva et al., 2017**, **McKinney et al., 2020**). Clinical trials are the gold standard for assessing new medications and interventions in a controlled and comparative manner, and the equivalent for developing AI algorithms are international competitions or "*grand challenges*". Grand challenges can address the lack of trust, scientific evidence and adequate validation among AI solutions (**Leeuwen et al., 2021**), by providing the means to compare algorithms against each other using common training and testing data. Present-day public benchmark of csPCa detection/diagnosis is the **ProstateX challenge** (**Armato et al., 2018**) from 2016-2017, which uses a testing set of 140 mpMRI exams to evaluate and compare AI algorithms. However, its small sample size and weak evaluation format (with publicly available, as opposed to truly "unseen" testing images), limits the ability to reliably draw out definitive conclusions.

The PI-CAI challenge is an all-new grand challenge that aims to validate the diagnostic performance of artificial intelligence and radiologists at csPCa detection/diagnosis in MRI, with histopathology and follow-up (≥ 3 years) as the reference standard, in a retrospective setting. The study hypothesizes that state-of-the-art AI algorithms, trained using thousands of patient exams, are non-inferior to radiologists reading bpMRI. As secondary end-points, it investigates the optimal AI model for csPCa detection/diagnosis, and the effects of DCE imaging and reader experience on diagnostic accuracy and inter-reader variability. However, the study neither validates the utility of AI as an assistive tool for concurrent reading in a prospective setting, nor does it evaluate the role of AI predictions in biopsy management and decision-making tasks.

Key aspects of the PI-CAI study design have been established in conjunction with an international scientific advisory board of 16 experts in prostate AI, radiology and urology —to unify and standardize present-day guidelines, and to ensure meaningful validation of prostate AI towards clinical translation (**Reinke et al., 2021**).

Item 3: Keywords

List the primary keywords that characterize the challenge.

prostate cancer; artificial intelligence; magnetic resonance imaging; radiologists; computer-aided detection and diagnosis

CHALLENGE ORGANIZATION

Item 4: Organizers

a) Provide information on the organizing team (names and affiliations).

Anindo Saha[1], **Jasper J. Twilt**[1], **Joeran S. Bosma**[1], **Bram van Ginneken**[1,2], **Derya Yakar**[3], **Mattijs Elschot**[4,5], **Jeroen Veltman**[6], **Jurgen Fütterer**[1], **Maarten de Rooij**[1], **Henkjan Huisman**[1,4]

[1] Department of Medical Imaging, Radboud University Medical Center, The Netherlands

[2] Fraunhofer Institute for Digital Medicine MEVIS, Germany

[3] Department of Radiology, Nuclear Medicine and Molecular Imaging, University Medical Center Groningen, The Netherlands

[4] Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Norway

[5] Department of Radiology and Nuclear Medicine, St. Olavs Hospital, Trondheim University Hospital, Norway

[6] Department of Radiology, Ziekenhuis Groep Twente, The Netherlands

b) Provide information on the primary contact person.

Anindo Saha: anindya.shaha@radboudumc.nl

Item 5: Lifecycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.
Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

After completion of the 2022 edition of the PI-CAI challenge, while the **Open Development Phase – Validation and Tuning** leaderboard will remain live for continuous submissions, interested teams can only make submissions to the **Open Development Phase – Testing** leaderboard upon request (with supporting documents, e.g. institutional e-mail address, associated publication, etc.). This step is necessary to preserve the integrity of the hidden testing cohort (by avoiding overfitting). It also ensures traceability and verification of all post-challenge solutions that claim to match/outperform prior submissions on the testing dataset. **Closed Testing Phase – Final Ranking** leaderboard will be closed, as it only represents the ranking of AI algorithms which were also trained on the private training datasets. PI-CAI will not be a one-time event. Future iterations may explore the effects of additional information (e.g. using full prostate mpMRI, instead of bpMRI sequences only) and more rigorous testing (e.g. larger number of external testing data centers) on AI performance and generalization.

Item 6: Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

This challenge is not associated with any conference.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

https://grand-challenge.org/

c) Provide the URL for the challenge website (if any).

https://pi-cai.grand-challenge.org/

Item 7: Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

- This challenge only supports the submission of fully automated methods in Docker containers. It is not possible to submit semi-automated or interactive methods.
- All Docker containers submitted to the challenge will be executed in an offline setting (i.e. they will not have access to the internet, and cannot download/upload any resources). All necessary resources (e.g. pre-trained AI model weights) must be encapsulated in the submitted containers apriori.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

- Use of pre-trained AI models on computer vision and/or medical imaging datasets (e.g. **ImageNet**, **Medical Segmentation Decathlon**), and use of any other dataset besides the PI-CAI training datasets, is allowed, only as long as such data and/or models are published under a permissive license (within 3 months of the **Open Development Phase** deadline), and participants clearly state their source and use-case, in each submission.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

- Members of all sponsoring or organizing entities (i.e. **Radboud University Medical Center**, **Ziekenhuis Groep Twente**, **University Medical Center Groningen**, **Norwegian University of Science and Technology**) can freely participate in the challenge, but are not eligible for the final ranking in the **Closed Testing Phase** or any of the prizes.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

- Teams that are responsible for developing the top-performing 3 AI algorithms will receive cash prizes and/or **Amazon Web Services (AWS) credits** (exact details pending ongoing discussions with sponsors).
- All prizes are non-transferrable.
- Members of all winning teams must have their true names and affiliations [university, institute or company (if any); country] displayed accurately on **verified Grand-Challenge profiles**, to be eligible for prizes.

e) Define the policy for result announcement.
Examples:

- Top three performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

- At the end of the **Open Development Phase**, all AI algorithm submissions and their respective performance will be announced publicly via its **leaderboard**.
- At the end of **Closed Testing Phase**, the top-performing 5 AI algorithms (including the winner of the PI-CAI challenge) and their respective performance will be announced publicly via its **leaderboard**.

f) Define the publication policy. In particular, provide details on ...

- … who of the participating teams/the participating teams' members qualifies as author
- … whether the participating teams may publish their own results separately, and (if so)
- … whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

— Upto 3 members from each team, that is responsible for one of the top-performing 5 AI algorithms, will be invited to join the PI-CAI challenge paper, as a consortium author.

— Participants of the PI-CAI challenge, as well as non-participating researchers using the PI-CAI public training dataset, can publish their own results any time, separately. They do not have to adhere to any embargo period. While doing so, they are requested to cite this document (BIAS preregistration form for the PI-CAI challenge), which will be published on Zenodo with a corresponding DOI. Once a study protocol and/or a challenge paper has been published, they are requested to refer to those publication(s) instead.

Item 8: Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

The PI-CAI challenge takes place in two phases:

— ***Open Development Phase*** *(Duration: 6 months)*
Anyone can participate in this phase of the challenge. Interested teams must join the PI-CAI challenge at **https://pi-cai.grand-challenge.org**. Afterwards, they will be provided access to download the **public training dataset**, and in turn, they can start developing and training AI models using their private or public compute resources (e.g. **Google Colaboratory**, **Kaggle**). Each team can submit a single **trained algorithm (in a Docker container)** for evaluation every week (similar to the **AIROGS**, **MIDOG2021** and **CoNIC2022** challenges). During evaluation, algorithms are executed on the grand-challenge.org platform, their performance is estimated on the **hidden validation and tuning cohort**, and team rankings are updated accordingly on a **live, public leaderboard**. Facilitating validation in such a manner, ensures that any image used for evaluation remains truly unseen, and that AI predictions cannot be tampered with. At the end of this phase, each team can choose to submit a single AI algorithm (presumably their top-performing model, but not necessarily their last submission) for evaluation on the **hidden testing cohort**. Based on their performance on this cohort, **all-new rankings will be drawn**.

— ***Closed Testing Phase*** *(Duration: 2 months)*
Teams with the top 5 AI algorithms of PI-CAI will be invited to participate in this phase of the challenge. Participants must prepare Docker containers of their AI algorithms that allow training, and subsequently, inference using the trained weights (similar to the **STOIC2021** and **NODE21** challenges). Organizers will retrain these models with large-scale data (**public + private training datasets**), using their institutional compute resources. Once training is complete, **performance will be re-evaluated on the hidden testing cohort** (with rigorous statistical analyses), and the winners of the PI-CAI challenge will be announced.

Instructions for submitting AI models encapsulated in Docker containers to the Grand Challenge platform: **https://grand-challenge.org/documentation/creating-an-algorithm-container/**

Source code for training a baseline vanilla U-Net (**Ronneberger et al., 2015**), nnU-Net (**Isensee et al., 2021**) or nnDetection (**Baumgartner et al., 2021**) algorithm, and encapsulating it in a Docker container for submission to the PI-CAI challenge: **https://github.com/DIAGNijmegen/picai_baseline**

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

See answer to Item 8(a).

Item 9: Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Tentative timeline:

— 5 May 2022; 6pm CEST: Release of **Public Training and Development Dataset**
— 20 May 2022; 6pm CEST: Accepting AI Algorithms for **Open Development Phase – Validation and Tuning**
— 20 October 2022; 6pm CEST: Closing Submissions for **Open Development Phase – Validation and Tuning**
— 20 October 2022; 6pm CEST: Accepting AI Algorithms for **Open Development Phase – Testing**
— 30 October 2022; 6pm CEST: Closing Submissions for **Open Development Phase – Testing**
— 10 November 2022; 6pm CEST: Accepting AI Algorithms for **Closed Testing Phase**
— 30 November 2022; 6pm CEST: Closing Submissions for **Closed Testing Phase**
— 30 December 2022; 6pm CEST: Winners of the PI-CAI Challenge are Announced Publicly

Item 10: Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval.

The institutional review boards of **Radboud University Medical Center (RUMC)**, **Ziekenhuis Groep Twente (ZGT)**, **University Medical Center Groningen (UMCG)** and **Norwegian University of Science and Technology (NTNU)** have waived the need for informed patient consent, for the retrospective scientific use of anonymized clinical data in this study.

MISSION OF THE CHALLENGE

Item 14: Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Main fields of application: diagnosis, research, risk stratification.

Item 15: Task category(ies)

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Task categories: classification, detection, localization, prediction.

Item 16: Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding gender or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

Target and challenge cohorts of PI-CAI are very similar. Both of them represent the same patient population: men suspected of harboring csPCa, with elevated levels of prostate-specific antigen (≥ 3 ng/mL) or abnormal findings on digital rectal exam, and without a history of treatment or any prior positive histopathology (**ISUP** ≥ 2) findings. Hidden testing and validation cohorts of the challenge include a similar distribution of **PI-RADS** and **ISUP** lesions as the one observed in the multi-center, consecutive 4M cohort (**van der Leest et al., 2019**). Nonetheless, deviations in the challenge cohorts with respect to the target cohort do exist, due to the following factors:

— By sampling one study per patient, we increase the diversity of benign and malignant findings in the hidden testing and validation cohorts. However, in clinical practice or the target cohort, multiple studies from the same patient can be encountered.

— Excluding cases from the training datasets and the hidden testing and validation cohorts, that cannot be annotated due to incomplete imaging, poor scan quality or artifacts (e.g. due to hip prostheses), MRI-invisible lesions and ambiguous diagnostic reports.

— Enriching the pool of 400 testing cases used for the reader study, with additional positives (with respect to clinical routine or the target cohort), in order to improve the statistical power.

— Excluding positive MRI (**PI-RADS** 3-5) cases without any corresponding histopathology reports in the training datasets, whose ground-truth for the presence/absence of csPCa, cannot be definitively established (note, we use follow-up information for such cases in the hidden validation and testing cohorts).

— Absence of patient scans in the training datasets, and the hidden validation and testing cohorts, acquired using MRI vendors besides Siemens Healthineers and Philips Medical Systems (e.g. GE Healthcare, Canon Medical Systems).

— Using a biopsy cohort exclusively (i.e. no negative MRI cases without histopathology) from one of the four data centers (University Medical Center Groningen), for the training datasets, and the hidden validation and testing cohorts.

Item 17: Imaging modality(ies)
Specify the imaging technique(s) applied in the challenge.

Biparametric prostate MRI: Axial, sagittal and coronal T2-weighted imaging (T2W); axial computed high b-value (≥ 1400 s/mm^2) diffusion-weighted imaging (DWI); axial apparent diffusion coefficient maps (ADC). All cases used for the reader study will also include dynamic contrast-enhanced (DCE) sequences.

Item 18: Context information
Provide additional information given along with the images. The information may correspond ...

a) … directly to the image data (e.g. tumor volume).

Context information related to the imaging data, made available for each case:

— **ISUP** or Gleason Grade Group per csPCa lesion, for all patient cases in the training datasets.
— MRI vendor (e.g. Siemens Healthineers) and scanner (e.g. Skyra), for all patient cases in the hidden testing and validation cohorts, and all patient cases in the training datasets.
— Prostate volume (unit: mL), if reported in the corresponding radiology report, for all patient cases in the hidden testing and validation cohorts, and all patient cases in the training datasets.
— Prostate volume (unit: mL), as calculated by **a publicly-available nnU-Net model trained for whole-gland segmentation**, for all patient cases in the hidden testing and validation cohorts, and all patient cases in the training datasets.

b) … to the patient in general (e.g. gender, medical history).

Clinical information related to the patient in general, made available for each case:

— Prostate-specific antigen (PSA) level (unit: ng/mL), if reported in the corresponding radiology report, for all patient cases in the hidden testing and validation cohorts, and all patient cases in the training datasets.
— Prostate-specific antigen density (PSAd) (unit: ng/mL), if reported in the corresponding radiology report, for all patient cases in the hidden testing and validation cohorts, and all patient cases in the training datasets.
— Patient age (unit: years), for all patient cases in the hidden testing and validation cohorts, and all patient cases in the training datasets.

Item 19: Target entity(ies)
a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Abdomen of male patient, shown in prostate MRI scans.

b) Describe the algorithm target, i.e. the structure(s) / subject(s) / object(s) / component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Clinically significant cancerous lesions afflicting the prostate gland.

Item 20: Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (parameter 26), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- *Example 1*: Find liver segmentation algorithm for CT images that processes CT images of a certain size in less than a minute on a certain hardware with an error that reflects inter- rater variability of experts.
- *Example 2*: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter 26).

The goal of prostate AI developed in the PI-CAI challenge is similar to that of **PI-RADS** —which operates on a lesion-level basis (detections with a likelihood or **PI-RADS** score per lesion), and where the patient-level diagnosis is mainly determined by the findings associated with the index lesion. In other words, the properties to be optimized for AI algorithms in this challenge are: lesion-level detection of csPCa in bpMRI; and patient-level diagnosis (or classification) of csPCa in bpMRI (using the predicted detections, as illustrated in Figure 1).

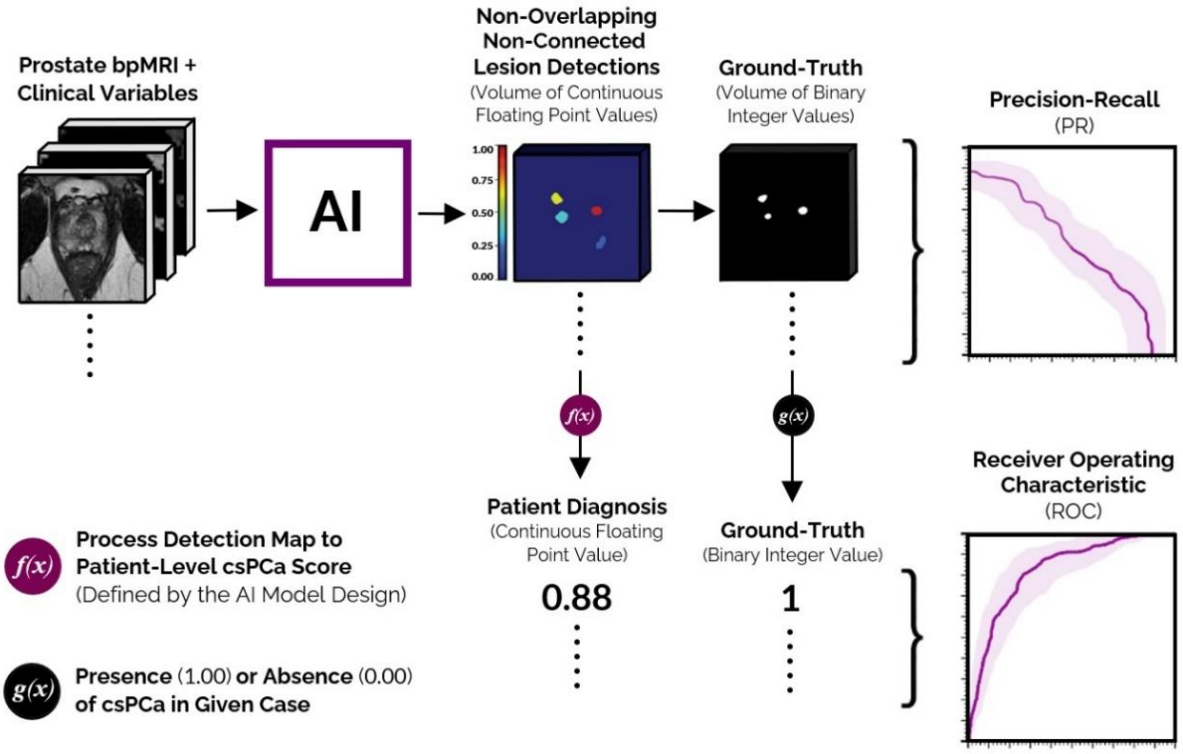

Figure 1 (top) Lesion-level csPCa detection (modeled by '**AI**'): For a given patient case, using the bpMRI exam, predict a 3D detection map of non-overlapping, non-connected csPCa lesions (with the same dimensions and resolution as the T2W image). For each predicted lesion, all voxels must comprise a single floating point value between 0-1, representing that lesion's likelihood of harboring csPCa.

(bottom) Patient-level csPCa diagnosis (modeled by '***f(x)***'): For a given patient case, using the predicted csPCa lesion detection map, compute a single floating point value between 0-1, representing that patient's overall likelihood of harboring csPCa. For instance, ***f(x)*** can simply be a function that takes the maximum of the csPCa lesion detection map, or it can be a more complex heuristic (defined by the AI developer).

CHALLENGE DATA SETS

Item 21: Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

All image acquisitions were obtained using Siemens Healthineers or Philips Medical Systems-based 1.5T or 3T MRI scanners with surface coils.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Multiparametric prostate MRI protocol, as detailed in **Engels et. al, 2020**.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

This retrospective study includes prostate MRI exams, acquired between 2012–2021, at three Dutch centers (**Radboud University Medical Center (RUMC)**, **Ziekenhuis Groep Twente (ZGT)**, **University Medical Center Groningen (UMCG)**), and one Norwegian center (**Norwegian University of Science and Technology (NTNU)**). Data provided from RUMC, also included 329 studies from the **ProstateX challenge**.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

All imaging acquisitions were performed by trained MRI radiographers at RUMC, UMCG, ZGT and NTNU.

Item 22: Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (parameter 21) and may include context information (parameter 18). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Training and test cases both represent prostate bpMRI scans of the abdomen in male patients. All training, validation and testing cases carry expert-derived image-level binary annotations for the presence/absence of csPCa. All validation/testing cases and 86% (1295/1500) of public training cases also carry expert-derived voxel-level lesion delineations of csPCa. All training cases also carry AI-derived voxel-level lesion delineations of csPCa (**Bosma et al., 2022**). We leave it upto the participants to formulate the most effective training strategy for their AI algorithms using some or all of this information.

A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see answers to Item 21) and may include context information (see answers to Item 18). Both training and test cases are annotated in accordance with the reference standard detailed in Item 23.

b) State the total number of training, validation and test cases.

In total, the PI-CAI dataset consists of four data splits with the following use-cases:

— *Public Training and Development Dataset (1500 cases)*
Used by all participants and researchers, to train and develop AI models during the **Open Development Phase**. All data is made available under a non-commercial **CC BY-NC 4.0** license. Includes 329 training and testing cases from the **ProstateX challenge**.

Imaging data is released via: **zenodo.org/record/6517398** (DOI: 10.5281/zenodo.6517398)
Annotations are released and maintained via: **github.com/DIAGNijmegen/picai_labels**

— *Private Training Dataset (7500-9500 cases)*
Used exclusively by the organizers to retrain the top-ranking 5 AI algorithms, with large-scale data, during the **Closed Testing Phase**.

— *Hidden Validation and Tuning Cohort (100 cases)*
Used to facilitate a live, public leaderboard that enables model selection and tuning, during the **Open Development Phase**.

— *Hidden Testing Cohort (1000 cases)*
Used to benchmark AI, radiologists, and test all hypotheses at the end of the **Closed Testing Phase**. Includes internal testing data (unseen cases from seen centers {RUMC, ZGT, UMCG}) and external testing data (unseen cases from an unseen center {NTNU}). A subset of 400 cases from this cohort is used to facilitate the **PI-CAI: Reader Study**.

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

A total of 1100 cases (1000 testing, 100 validation) is used for evaluation, considering both practicality (numbers of cases in our multi-center cohort, for which it would be feasible to acquire at least 3 years of follow-up data) and viability (minimum number of cases need for a reliable AI performance benchmark at present-time and through the coming years). All remaining 9000-11,000 cases in the cohort are used to create the training datasets (1500 public, 7500-9500 private).

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

To meaningfully validate AI for prostate MRI towards clinical translation, it is essential to analyze performance across the complete patient population encountered in clinical routine. For instance, while **radical prostatectomy (RP)** can provide the most comprehensive tissue specimen to facilitate accurate histopathology grading, a cohort of solely RP patients is heavily biased –deviating substantially from the distribution of patients encountered during clinical routine (where the vast majority of men carry benign tissue or low-grade PCa). In other words, a clinically representative validation or testing cohort must also include patients without cancer and with negative MRI, and those who have undergone systematic and/or targeted biopsies (**Schelb et al., 2019**). Various combinations of RP, biopsy and/or negative MRI-based testing cohorts have been investigated across multiple recent studies on automated csPCa diagnosis (**Schelb et al., 2019**, **Mehta et al., 2021**, **Bhattacharya et al., 2021**, **Hosseinzadeh et al., 2021**, **Winkel et al., 2021**, **Saha et al., 2021**, **Netzer et al., 2021**), and such a combination was used for the **ProstateX challenge** as well. Similarly, the hidden testing and validation cohorts of PI-CAI were curated to resemble the real-world distribution of men suspected of harboring csPCa, as closely as possible, while retaining a strong reference standard for the presence/absence of csPCa.

The 4M cohort (**van der Leest et al., 2019**) was used as a point of reference, for the observed patient and lesion-level distributions of csPCa prevalence in a multi-center, consecutive cohort. Exact characteristics and enrichment of the case mixture per split (**Pinsky et al., 2012**), including the distribution of **PI-RADS** and **ISUP** lesions for the PI-CAI datasets, have been considered and recorded. However, this information remains blinded for the hidden validation and testing cohorts, till the study is officially complete. For the private training dataset, these

numbers will be released at a later date (before the start of the **Closed Testing Phase**) pending their full curation. For the public training dataset, these numbers are listed in Table 1.

Table 1 | Data splits for the training datasets, and the hidden validation and testing cohorts.

Data Source	Public Training and Development Set	Private Training Set	Hidden Validation and Tuning Cohort	Hidden Testing Cohort	Total
	RUMC, ZGT, UMCG The Netherlands	**RUMC, ZGT, UMCG** The Netherlands	**RUMC, ZGT, UMCG** The Netherlands	**RUMC, ZGT, UMCG, NTNU** The Netherlands, Norway	
No. of Sites	**11**	**11**	**5**	**6**	**11**
No. of MRI Scanners	**5 S, 2 P**	**6 S, 3 P** †	**6 S, 3 P** †	**6 S, 3 P** †	**6 S, 3 P** †
No. of Patients	**1476**	**8800** †	**100**	**1000**	**11,376** †
No. of Cases	**1500**	**9000** †	**100**	**1000**	**11,600** †
— Benign or Indolent PCa †	**1075**	*{ To-Be-Announced }*	*{ Blinded }*	*{ Blinded }*	–
— csPCa (ISUP ≥ 2)	**425**	*{ To-Be-Announced }*	*{ Blinded }*	*{ Blinded }*	–
Median Age (years)	**67** {IQR: 61–71}	*{ To-Be-Announced }*	*{ Blinded }*	*{ Blinded }*	–
Median PSA (ng/mL)	**8.5** {IQR: 6–13}	*{ To-Be-Announced }*	*{ Blinded }*	*{ Blinded }*	–
Median Prostate Volume (mL)	**57** {IQR: 40–80}	*{ To-Be-Announced }*	*{ Blinded }*	*{ Blinded }*	–
No. of Positive MRI Lesions	**1087**	*{ To-Be-Announced }*	*{ Blinded }*	*{ Blinded }*	–
— PI-RADS 3	**246** (23%)	*{ To-Be-Announced }*	*{ Blinded }*	*{ Blinded }*	–
— PI-RADS 4	**438** (40%)	*{ To-Be-Announced }*	*{ Blinded }*	*{ Blinded }*	–
— PI-RADS 5	**403** (37%)	*{ To-Be-Announced }*	*{ Blinded }*	*{ Blinded }*	–
No. of ISUP-Based Lesions	**775**	*{ To-Be-Announced }*	*{ Blinded }*	*{ Blinded }*	–
— ISUP 1	**310** (40%)	*{ To-Be-Announced }*	*{ Blinded }*	*{ Blinded }*	–
— ISUP 2	**260** (34%)	*{ To-Be-Announced }*	*{ Blinded }*	*{ Blinded }*	–
— ISUP 3	**109** (14%)	*{ To-Be-Announced }*	*{ Blinded }*	*{ Blinded }*	–
— ISUP 4	**41** (5%)	*{ To-Be-Announced }*	*{ Blinded }*	*{ Blinded }*	–
— ISUP 5	**55** (7%)	*{ To-Be-Announced }*	*{ Blinded }*	*{ Blinded }*	–

RUMC: Radboud University Medical Center; **ZGT**: Ziekenhuisgroep Twente; **UMCG**: University Medical Center Groningen; **NTNU**: Norwegian University of Science and Technology; **S**: Siemens Healthineers MRI scanner {Skyra 3T, TrioTim 3T, Prisma 3T, Aera 1.5T, Avanto 1.5T, Espree 1.5T}; **P**: Philips Medical Systems MRI scanner {Ingenia 3T, Achieva 1.5T, Intera 1.5T}

† Tentative numbers.

Item 23: Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

Hidden Testing and Validation Cohort Annotations

For accurate validation of AI and human-reader performance, and in turn, to substantiate any conclusions derived from PI-CAI, a strong reference standard for csPCa is crucial. The PI-CAI reference standard aims to utilize the best possible evidence to define the ground-truth for every case in the validation and testing cohorts, i.e. histologically-confirmed (**ISUP** ≥ 2) positives, and histologically- (**ISUP** ≤ 1) or MRI- (**PI-RADS** ≤ 2) confirmed negatives, with follow-up (≥ 3 years), as detailed below and summarize in Figure 2.

— Patients with negative MRI (i.e. benign or carrying **PI-RADS** 1–2 lesions) generally do not undergo biopsies or RP and lack histologically-confirmed evidence for the absence of csPCa. It is likely that they do not harbor csPCa, but a small percentage (<1% at RUMC; **Venderink et al., 2019**) can still be missed. To alleviate this, upto 40% of the validation and testing cohorts are composed of multi-center patient data from the 4M cohort (**van der Leest et al., 2019**), where all patients with negative MRI had received systematic biopsies and subsequent grading was supervised by an expert uropathologist (> 25 years of experience). In other words, by using data from the 4M cohort, we are able to acquire histopathology evidence for a large fraction of the patient population, that is encountered, but typically not histologically-confirmed during clinical routine.

— Biopsies alone can still be prone to undersampling csPCa, especially in the case of smaller lesions (**Srivastava et al., 2019**). Hence, all negative cases (negative MRI and/or histopathology) in the validation and testing cohorts are be confirmed with follow-up data (e.g. using the national **Dutch Pathology Registry (PALGA)** for centers based in The Netherlands). Negative patient exams found to be positive (via MRI or histopathology) in ≥ 3 years of follow-up, were inspected with an expert radiologists for retrospective signs of potentially missed csPCa. If the presence of csPCa can be definitively confirmed, they are included as positive cases; otherwise,

they are excluded. Negative patient exams with 100% csPCa diagnosis-free survival (DFS) after at least 3 years, are included.

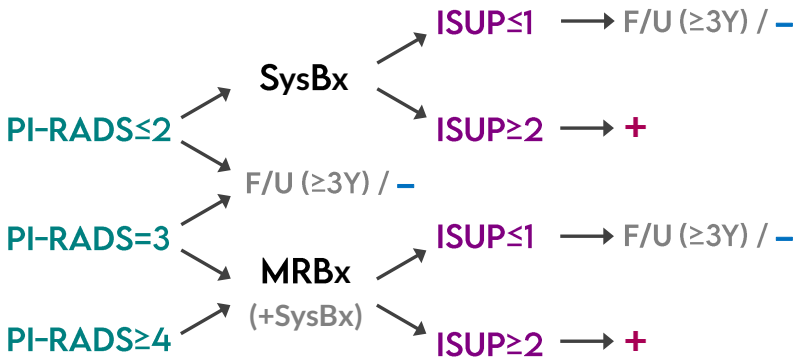

Figure 2. Typical workflow used to establish the ground-truth for each lesion in the hidden validation and testing cohorts. If **systematic biopsies (SysBx)** were performed in addition to **MRI-targeted biopsies (MRBx)**, then **SysBx** findings are only used to upgrade the ISUP score not downgrade. If **RP** is performed, its corresponding findings supersede that of any prior histopathology/radiology findings. Cases for which pathology findings cannot be localized on MRI (e.g. MRI-invisible lesions, **SysBx** diagnostic reports with ambiguous or missing location information) are excluded.

Training Dataset Annotations
Patient cases used for the training datasets of PI-CAI are annotated with the same reference standard as used for the **ProstateX challenge**, i.e. histologically-confirmed (**ISUP** ≥ 2) positives, and histologically- (**ISUP** ≤ 1) or MRI- (**PI-RADS** ≤ 2) confirmed negatives, without follow-up.

Annotators
For all cases, voxel-level csPCa lesion annotations are delineated and/or patient-level csPCa outcomes are recorded, by one of 10 trained investigators or 1 radiology resident, under supervision of one of 3 expert radiologists, at RUMC, UMCG or NTNU. For all training cases, automated AI-derived delineations of csPCa lesions (**Bosma et al., 2022**) have also been made available.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

Each annotation is derived using MRI scans, diagnostic reports (radiology, pathology) and whole-mount prostatectomy specimen (if applicable). Lesion delineations are created using **ITK-SNAP v3.80**. Due to the standardized precautionary measures (e.g. minimal temporal difference between acquisitions, administration of antispasmodic agents to reduce bowel motility, use of rectal catheter to minimize distension) taken in the imaging protocol (**Engels et al., 2020**), we typically observe negligible patient motion across different MRI sequences. Nonetheless, any patient exam in the validation or testing cohorts, with substantial misalignment between its sequences, is manually registered (rigid transformation; with six degrees of freedom for 3D translation and rotation) using **ITK-SNAP v3.80**, before their annotations are made.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

See answer to Item 23(a).

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

Multiple annotations were not encountered for any validation or testing cases. Multiple annotations (AI, human expert) for the same training cases were not merged, but rather provided as-is to all participants.

e) In an analogous manner, describe and quantify other relevant sources of error.

Contours of all csPCa lesion delineations are susceptible to annotation errors, due to the following factors:

- Exact spatial extent of csPCa lesions cannot be clearly estimated on MRI, sometimes even while using whole-mount histopathology as reference.
- Inter-reader variability among annotators.

Note, that we primarily investigate image-level classification and lesion-level detection (using a lenient hit criterion, as detailed in Item 26(a)) performance in the PI-CAI challenge. We do not evaluate segmentation or the exact spatiant extent of csPCa, predicted by radiologists or AI. Thus, annotation uncertainty along lesion boundaries, due to the factors listed above, have negligible impact (if any) on the outcomes of this study.

ASSESSMENT METHODS

Item 26: Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (parameter 20). State which metric(s) were used to compute the ranking(s) (if any).

- *Example 1*: Dice Similarity Coefficient (DSC) and run-time
- *Example 2*: Area under curve (AUC)

Source code to compute all performance metrics discussed in Item 26: Metrics, Item 27: Ranking Method and Item 28: Statistical Analyses, have been provided at: **https://github.com/DIAGNijmegen/picai_eval**

Primary Performance Metrics

Key performance metrics used to evaluate AI and radiologists, have been summarized in Figure 3.

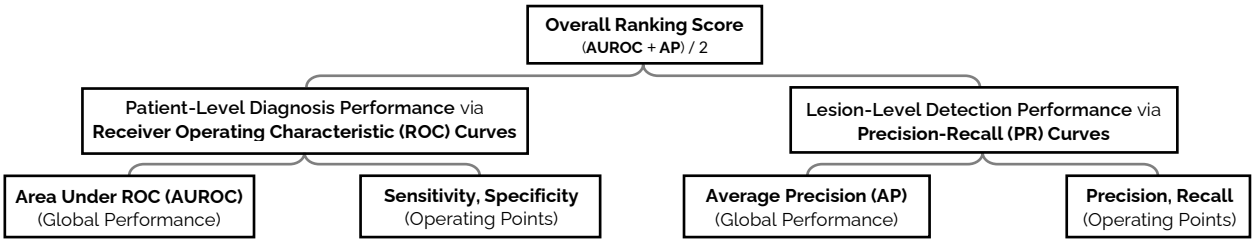

Figure 3. Summary of key performance metrics used to evaluate AI and radiologists in PI-CAI. Overall ranking score is only used to evaluate different AI algorithms w.r.t each other and facilitate the validation and testing leaderboards. When comparing AI performance to that of radiologists at specific operating points, AI is thresholded to match the sensitivity, specificity, precision or recall of radiologists' PI-RADS operating points (as recommended in Penzkofer et al., 2022, Padhani et al., 2019, Schelb et al., 2019).

Secondary Performance Metrics

Intersection over Union (IoU) is used for spatial congruence analysis of AI detections (but not for validation or testing, given that IoU cannot accurately evaluate detection or diagnosis performance (Reinke et al., 2020)). Similarly, Free-Response Receiver Operating Characteristic (FROC) curves are also used for secondary analysis of AI detection performance (as recommended in Penzkofer et al., 2022).

Comparison to Radiologists

Each of the top-ranking AI algorithms will be compared against radiologists from the reader study, in multiple ways (as shown in Figure 4). Additionally, AI will be compared against the radiologists who prospectively

performed the actual clinical read for each patient case in the hidden testing cohort, using the historical findings noted in their radiology reports.

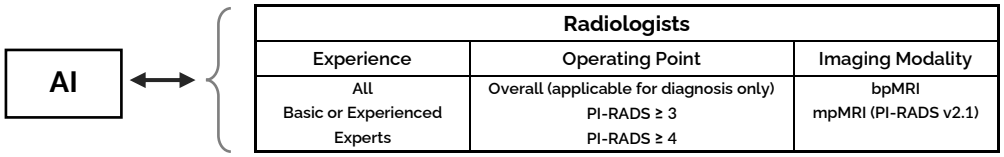

Figure 4. Radiologists from the reader study can be stratified into multiple possible configurations on the basis of experience, operating point and imaging modality (e.g. a panel of expert radiologists reading bpMRI at an operating point of **PI-RADS** ≥ 3). Considering these factors, each AI model will be compared against 18 possible configurations of radiologists at patient diagnosis, and 12 possible configurations of radiologists at lesion detection. For each comparison, we primarily use the performance metrics listed in Figure 3.

Hit Criterion

A "hit criterion" is a condition that must be satisfied for each predicted lesion to count as a *hit* or true positive. For csPCa detection in recent prostate-AI literature, hit criteria have been typically fulfilled on the basis of localizing predictions to a specific region (as defined by sector maps), by achieving a minimum degree of prediction-ground truth overlap, or by localizing predictions within a maximum distance from the ground-truth.

For the 3D detections predicted by AI, we opt for a hit criterion based on object overlap:

— *True Positives*
For a predicted csPCa lesion detection to be counted as a true positive, it must share a minimum overlap of 0.10 IoU in 3D with the ground-truth annotation. Such a threshold value, is in agreement with other lesion detection studies from literature (**Duran et al. 2022**, **Baumgartner et al., 2021**, **Saha et al., 2021**, **Hosseinzadeh et al., 2021**, **McKinney et al., 2020**, **Jaeger et al., 2019**).

— *False Positives*
Predictions with no/insufficient overlap count towards false positives, irregardless of their size or location.

— *Edge Cases*
When there are multiple predicted lesions with sufficient overlap (≥ 0.10 IoU), only the prediction with the largest overlap is counted, while all other overlapping predictions are discarded. Predictions with sufficient overlap that are subsequently discarded in such a manner, do not count towards false positives to account for split-merge scenarios.

For the point-coordinate annotations predicted by radiologists, we opt for a hit criterion based on distance:

— *True Positives*
For a predicted csPCa lesion coordinate to be counted as a true positive, it must reside inside or within ≤ 5 mm of the csPCa annotation boundary (as done in **Cao et al., 2019**). Such a margin is considered to account for smaller csPCa lesions, where the ground-truth annotation spans 1-2 slices, and radiologists' point predictions can register a miss from marginal deviations (despite correct cognitive localization).

— *False Positives*
Predictions that are not inside or within ≤ 5 mm of the csPCa annotation boundary, count towards false positives.

— *Edge Cases*
In the case of multiple point-coordinate annotations inside or within ≤ 5 mm, only the closests prediction is counted. Predictions within the minimum distance that are subsequently discarded in such a manner, do not count towards false positives to account for split-merge scenarios.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

See answers to Item 20 and Item 26(a).

Item 27: Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

See answers to Item 20 and Item 26(a).

b) Describe the method(s) used to manage submissions with missing results on test cases.

Submissions with missing results will be disqualified, and not presented on any leaderboard.

c) Justify why the described ranking scheme(s) was/were used.

For patient pathway planning, both patient-level risk stratification and lesion-level detection are instrumental. The two are linked, and equally important for the diagnostic process. Therefore, we opted for equal weighting between AUROC and AP (see Figure 3).

Item 28: Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

Statistical tests are performed in the final arc of the PI-CAI challenge, i.e. after selecting the top-ranking 5 AI algorithms, and estimating the performance of all radiologists from the reader study. Each test is facilitated using AI and/or radiologists' predictions on the hidden testing cohort. More details and the source code for the exact implementation of the statistical tests reported in this study: **https://github.com/DIAGNijmegen/picai_eval**

Each AI algorithm is trained on the same training dataset and evaluated on the same testing dataset, multiple times (5-10x), and all of these independently trained instances are used in each statistical test. By doing so, we account for the performance variance resulting from the stochastic optimization of machine/deep learning models (due to which, the same AI architecture, trained on the same data, for the same number of training steps, typically can exhibit different performance each time (**Bosma et al., 2022**)). Our goal is to avoid basing any conclusions off of one arbitrary training run (which may prove "lucky" or "unlucky" for a given AI algorithm), and to promote reproducibility.

By sampling one case per patient for the hidden testing and validation cohorts, we can increase the diversity of malignant and benign findings. But this shifts the cohort distribution from what is typically observed in clinical routine: where multiple studies from the same patient can be encountered in 1000 consecutive cases. Similarly, enriching the case mixture with additional positives can facilitate meaningful analysis, when there is a limited number of testing cases and clinically significant disease with low prevalence (**McKinney et al., 2020**, **Pinsky et al., 2012**). But once again, it shifts the distribution from what is typically observed in clinical routine. Hence, inverse probability weighting (**Mansournia et al., 2016**) is incorporated in each statistical test, to minimize such bias stemming from preferential patient sampling techniques and the use of non-consecutive cohorts.

In total, four types of comparisons are statistically evaluated. First two comparisons address the primary hypotheses of the PI-CAI challenge, i.e. the stand-alone performance of state-of-the-art AI algorithms, w.r.t. that of radiologists:

— *AI vs Radiologists from Clinical Routine*

Comparison: Between each of the top-ranking 5 AI algorithms, and the historical reads made by radiologists during clinical routine.

Statistical Question: What is the probability that a given trained AI algorithm outperforms radiologists from clinical routine, when the AI algorithm is trained on the complete public + private training datasets, and evaluated on all 1000 cases from the hidden testing cohort, while accounting for the performance variance stemming from different cases and the AI algorithm's training method?

Statistical Test: Paired bootstrapping (as applied in **Ruamviboonsuk et al., 2022**, **McKinney et al., 2020**, **Rodriguez-Ruiz et al., 2019**), using predictions from a given operating point. Here, the operating point is that of radiologists (**PI-RADS** ≥ 3 or **PI-RADS** ≥ 4) from clinical routine, and trained AI algorithms are thresholded at matched sensitivity/specificity (for patient diagnosis) or recall/precision (for lesion detection). In each of 1M replications, $\sim \mathcal{U}(0, N)$ cases are sampled with replacement, and used to calculate the test statistic. Iterations that sample only one class are rejected. Test statistic is the rank of historical reads made by radiologists, with respect to the predictions made by trained AI algorithms, where the rank is determined by the conjugate performance metric.

— *AI vs Radiologists from Reader Study*

Comparison: Between each of the top-ranking 5 AI algorithms, and a given panel of radiologists from the reader study (refer to Figure 4 for all possible panel configurations).

Statistical Question: What is the probability that a given AI algorithm outperforms the average reader from a given panel of radiologists, when the AI algorithm is trained on the complete public + private training datasets and evaluated on 400 cases from the hidden testing cohort, while accounting for the performance variance stemming from different readers, different cases and the AI algorithm's training method?

Statistical Tests: Multi-reader multi-case (MRMC) analysis (as applied in **McKinney et al., 2020**, **Rodriguez-Ruiz et al., 2019**, **Bejnordi et al., 2017**) using the publicly available **iMRMC v4.0.3 software** (Division of Imaging, Diagnostics, and Software Reliability, FDA/CDRH/OSEL) (**Gallas et al., 2009**), and permutation tests (as applied in **Ruamviboonsuk et al., 2022**, **Bulten et al., 2022**, **McKinney et al., 2020**).

Using MRMC analysis, a non-inferiority test (with a non-inferiority margin of 0.05) is used to compare overall patient-level diagnosis performance (using patient-level predictions from AI, and patient-level suspicion scores from radiologists). For radiologists, average AUROC is computed using the diagonal average (which is area preserving (**Chen et al., 2014**)). Non-inferiority is concluded if the AUROC difference between AI and readers is greater than 0 and the lower bound of the 95% confidence interval is greater than the non-inferiority margin (**Rodriguez-Ruiz et al., 2019**). By utilizing MRMC analysis of variance, results can be generalized to new readers, new non-diseased cases and new diseased cases.

Permutation tests are used to statistically compare lesion-level detection and patient-level diagnosis performance at **PI-RADS** operating points. Here, in each of 1M replications, performance metrics (reader performance w.r.t. AI performance at reader's operating point) for the hidden testing cohort are shuffled across methods (AI, radiologists) and their instances (independently trained samples of AI algorithm, different readers).

Next two comparisons address the secondary endpoints of this study, i.e. determining the best overall AI algorithm for csPCa detection and diagnosis in bpMRI (and in turn, the winner of PI-CAI grand challenge), and investigating the effects of DCE imaging, reader experience and imaging modality on the diagnostic accuracy of radiologists:

— *AI vs AI*

Comparison: Between every pair of AI algorithms among the top-ranking 5 AI algorithms.

Statistical Question: What is the probability that one AI algorithm outperforms another, when both are trained on the complete public + private training datasets, and evaluated on all 1000 cases from the hidden testing cohort, while accounting for the performance variance stemming from different cases and each AI algorithm's training method?

Statistical Test: Permutation tests (as applied in **Ruamviboonsuk et al., 2022**, **Bulten et al., 2022**, **McKinney et**

al., 2020), where in each of 1M replications, performance metrics (ranking score, AP or AUROC) for the hidden testing cohort are shuffled across methods (different AI algorithms) and their instances (independently trained samples of each method).

— *Radiologists vs Radiologists from Reader Study*

Comparison: Between different panels of radiologists from the reader study (refer to Figure 4 for all possible panel configurations).

Statistical Question: What is the probability that an average radiologist from a given panel of radiologists, outperforms an average radiologist from another, while accounting for the performance variance stemming from different readers and different cases?

Statistical Tests: Multi-reader multi-case (MRMC) analysis (as applied in **McKinney et al., 2020**, **Rodriguez-Ruiz et al., 2019**, **Bejnordi et al., 2017**) using the publicly available **iMRMC v4.0.3 software** (Division of Imaging, Diagnostics, and Software Reliability, FDA/CDRH/OSEL) (**Gallas et al., 2009**), and permutation tests (as applied in **Ruamviboonsuk et al., 2022**, **Bulten et al., 2022**, **McKinney et al., 2020**).

Using MRMC analysis, a non-inferiority test (with a non-inferiority margin of 0.05) is used to compare overall patient-level diagnosis performance (using patient-level suspicion scores from radiologists). Average AUROC is computed using the diagonal average (which is area preserving (**Chen et al., 2014**)). Non-inferiority is concluded if the AUROC difference between the two separate panels of readers is greater than 0 and the lower bound of the 95% confidence interval is greater than the non-inferiority margin (**Rodriguez-Ruiz et al., 2019**). By utilizing MRMC analysis of variance, results can be generalized to new readers, new non-diseased cases and new diseased cases.

Permutation tests are used to statistically compare lesion-level detection and patient-level diagnosis performance at **PI-RADS** operating points. Here, in each of 1M replications, performance metrics (agreement with histopathology/follow-up reference standard, at **PI-RADS** operating points) for the hidden testing cohort are shuffled across methods (different panels of radiologists) and their instances (different readers).

Power Analysis

A prospective or a priori power analysis is performed to estimate the required sample size and optimal study design for the PI-CAI: Reader Study —such that it is able to substantiate the results of a non-inferiority test between AI and radiologists at patient-level diagnosis of csPCa (see answers to Item 28). While the power analysis is performed using the iMRMC sizing module, it is important to note that this software reserves power analysis only for superiority tests. Non-inferiority tests (as intended for the PI-CAI challenge) typically require larger sample sizes than superiority studies, to achieve the same statistical power (**Vavken et al., 2011**). Therefore, the implications of the following power analysis (which assumes a superiority test) are limited.

The iMRMC sizing module uses the F-test from **Hillis et al., 2011**, where statistical power is calculated by utilizing MRMC components of variance for a given number of cases, its class distribution, the number of readers and the effect size. We compute the same, in accordance with the workflow of **Gallas et al., 2019a**, and as reported in the VIPER study (Supplementary Materials, **Gallas et al., 2019b**).

Since the performance of AI models submitted to the PI-CAI challenge, and radiologists enlisted for the reader study, are not known apriori, we use 15 independently trained instances of our institutional state-of-the-art AI for csPCa detection in bpMRI (**Bosma et al., 2022**), an expert consensus of radiologists and an external cohort of 300 cases (96 cases with csPCa; 204 cases with indolent PCa or benign tissue) from the 4M study (**van der Leest et al., 2019**), as pilot data. The AI algorithm has an average AUROC of 0.898 (95% CI: 0.862, 0.934). Permutations of case-level **PI-RADS** scores from the expert consensus is used to simulate a distribution of 45 readers. Their **PI-RADS** scores are subsequently used to sample ROC curves and calculate reader AUROCs accordingly. For the PI-CAI challenge, these are estimated from the patient-level suspicion scores (see answers to Item 29), which allow for more accurate ROC sampling using a larger number of operating points. Average reader AUROC is 0.812 (95% CI: 0.778, 0.846), which is in concordance with the reported heterogeneity of prostate MRI assessments in literature (**Westphalen et al., 2020**, **Smith et al., 2019**, **Rosenkrantz et al., 2016**, **Garcia-Reyes et al., 2015**).

Within iMRMC, different split-plots, class enrichment strategies and total cohort sizes are simulated by changing the number of normal cases, diseased cases, split groups and readers. Table 2 presents the power estimations and total standard errors for a superiority or non-equivalence test using the aforementioned pilot data, with a

significance level of 0.05, effect size of 0.05, and various possible study designs (all of which, limit the number of cases read by each reader to 100). For reference, in our intended non-inferiority test between the stand-alone csPCa diagnosis performance of AI w.r.t. radiologists, we target a power of at least 0.80 (with a non-inferiority margin of 0.05).

Table 2 | Power estimation for possible study designs (w.r.t. number cases, class enrichment, number of readers).

Total Cases (N_C)	Percentage of Negatives (N_0) : Positives (N_1)	Number of Readers (N_R)	Number of Split Groups (N_G)	Cases / Reader (N_C/N_R)	Total Standard Error (S.E.)	Power
200	70 : 30	20	2	100	0.0229	0.59
200	70 : 30	40	2	100	0.0206	0.68
200	70 : 30	60	2	100	0.0198	0.71
200	60 : 40	20	2	100	0.0217	0.63
200	60 : 40	40	2	100	0.0194	0.73
200	60 : 40	60	2	100	0.0186	0.77
400	70 : 30	20	4	100	0.0191	0.75
400	70 : 30	40	4	100	0.0162	0.87
400	70 : 30	60	4	100	0.0152	0.91
400	60 : 40	20	4	100	0.0182	0.78
400	60 : 40	40	4	100	0.0159	0.90
400	60 : 40	60	4	100	0.0143	0.94

Within 20-60 readers, 200-400 cases and 60-70% positives in the case mixture, results indicate that a greater power is achieved primarily from an increase in the total cohort size and the number of readers, while smaller increases in power are observed for a higher enrichment of positive cases in the case mixture. Furthermore, a larger number of split groups seemingly have relatively low impact on the total standard error and power (similar to the observations reported in **Gallas et al., 2019**, **Chen et al., 2018**). Outcomes of this power analysis were used to inform the proposed reader study design (see answers to Item 29).

b) Justify why the described statistical method(s) was/were used.

See Item 28(a).

Item 29: Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- **inter-algorithm variability,**
- common problems/biases of the submitted methods, or
- ranking variability

Reader Study: Tasks and Interface

Parallel to the AI study, the objective of the reader study is evaluate the performance of the average or typical prostate radiologist at patient-level diagnosis and lesion-level detection of csPCa. However, in contrast to the AI study, the reader study is conducted using both bpMRI (enabling head-to-head comparisons against AI trained on bpMRI), and mpMRI (enabling comparisons between AI and current clinical practice, i.e. **PI-RADS** v2.1), using a two-stage annotation workflow (as illustrated in Figure 5):

— *Reading bpMRI*

Radiologists receive the same information as AI (i.e. prostate bpMRI exams and their clinical variables). They identify and set a point-coordinate inside each suspected csPCa lesion (if present; with a maximum of 4 lesions annotations permitted per case), and a corresponding **PI-RADS** score (ranging from 3–5) for its severity. They also provide an overall patient-level likelihood score for the presence of csPCa between 0–100, with incremental steps of 1 (similar to **Winkel et al. 2021**, **Jacobs et al., 2021**, **McKinney et al., 2020**, **Rodriguez-Ruiz et al., 2019**). Such a score allows us to estimate the level of suspicion for healthy cases and those with **PI-RADS** 1–2 lesions, where no lesion-level annotations will be made. Additionally, it also allows us

compare radiologists' overall diagnostic performance against that of AI using a larger range of operating points (instead of the four **PI-RADS** points that can be assumed, by considering the index lesion per case for patient-level diagnosis). Once completed, their answers are saved and cannot be revisited any further.

— *Reading mpMRI*
Immediately afterwards, radiologists are provided access to the full mpMRI (including the DCE sequence) from the same study. In other words, they can review the patient exam in compliance with **PI-RADS v2.1**. They can update their marks (point-coordinate annotation and **PI-RADS** score per lesion, if any) in light of the additional information, or keep them as they were. As radiologists cannot revisit any case, bpMRI assessments cannot be adjusted with knowledge obtained from mpMRI.

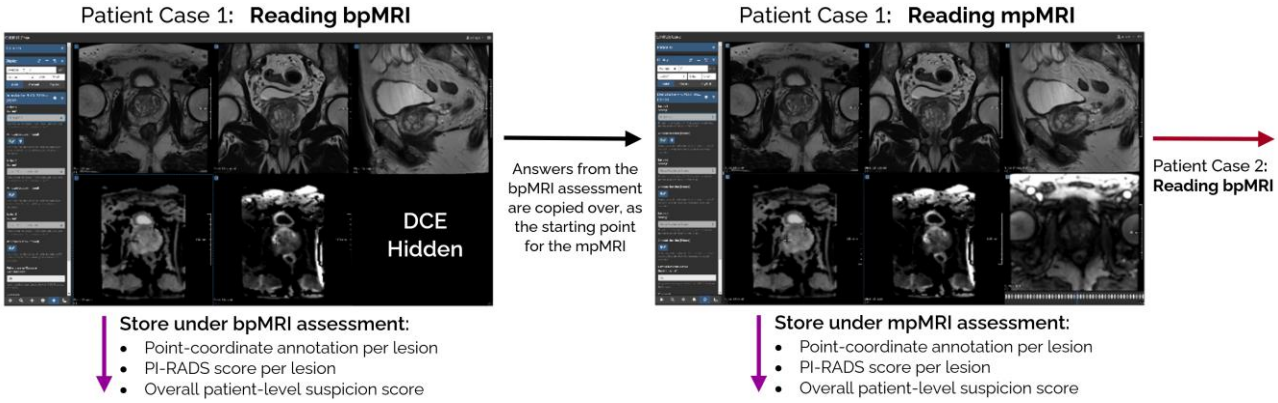

Figure 5. Reading interface and two-stage annotation workflow aimed at facilitating bpMRI and mpMRI assessments, sequentially. At the end of each bpMRI assessment, answers are copied over as the starting point for the mpMRI assessment of the same case. Hence, readers are only required to make changes (if needed) during the mpMRI assessment. They are not required to remember their previous scores or re-evaluate the complete exam from scratch (an alternative methodology, that would ideally require a washout period).

Reader Study: Study Design
The PI-CAI: Reader Study is hosted on **grandchallenge.org/reader-studies** (which supports prostate MRI viewers and annotation workflows) using a subset of 400 cases (40%) from the hidden testing cohort. Primary costs of the study are reader time and effort. As reading all 400 cases is too labor-intensive, we opt for a split-plot design (**Obuchowski et al., 2009**, **Obuchowski et al., 2012**, **Chen et al., 2018**). By doing so, we preserve a sample size of 400 cases, while the individual workload is reduced to a maximum of 100 cases (as illustrated in Figure 6). Prior to the start of this study, readers are provided a detailed guide on the annotation workflow, including all tools made available on the platform (e.g. navigation, zooming, windowing, measuring scale). They are also provided access to a practice session with 6 example cases (from the **ProstateX challenge**), to get familiarized with both the reading interface and the expected workflow. Afterwards, each reader uses their grand-challenge.org account to access their instance of the reader study (with all 100 allotted cases). Cases are made available sequentially and cannot be revisited post-assessment. Readers are expected to complete their assessments in 3-5 months.

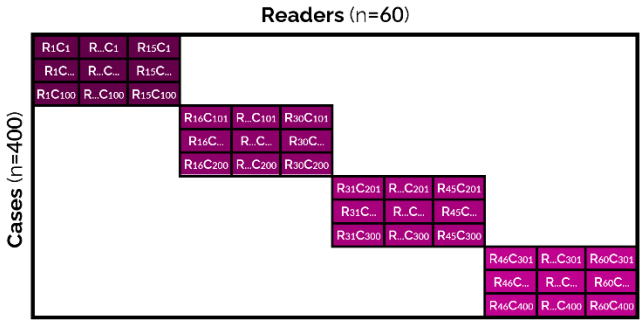

Figure 6. Tentative study design for the distribution of readers and cases in a 4x4 split-plot configuration. All 60 readers and 400 cases are divided into 4 blocks, in a stratified manner, that takes reader and case distributions into account to minimize any potential differences between separate blocks. Each block of readers reads their own set of cases. As this study design is reader-dependent, it is susceptible to changes based on the final outcomes of reader recruitment.

Reader Study: Registered Radiologists

We include any clinician who reads and reports prostate MRI in clinical practice, and is aware of the **PI-RADS v2.1** guidelines. Readers among all experience levels have been nominated or invited to participate in the reader study. Reader experience is categorized according to the **2020 ESUR/ESUI consensus statements**, and by their years of experience in reading and reporting prostate MRI. Prior to performance analysis, the names of all readers are pseudonymized, such that any individual performance cannot be traced back to its respective identity by anyone, but key researchers leading the PI-CAI challenge. As of 6th April 2022, in total, 63 readers have registered for the reader study (as summarized in Figure 7).

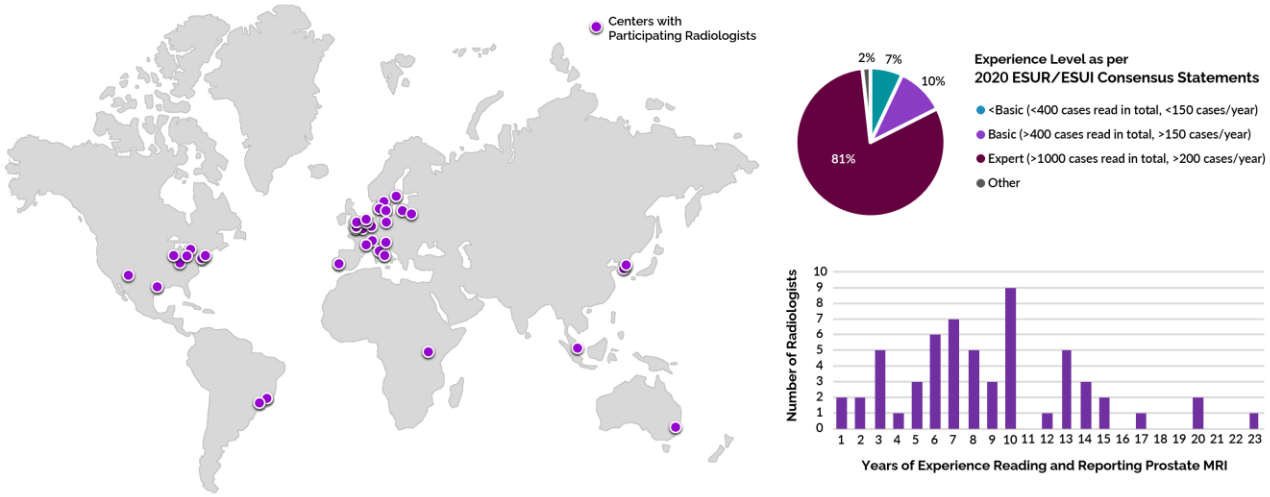

Figure 7. Distribution of 63 radiologists participating in the PI-CAI: Reader Study, representing 42 centers across 18 countries. Reader experience varies between 1 and 23 years (median 8 years), where 78% (49) of readers can be categorized as "expert" based on ESUR/ESUI guidelines (**de Rooij et al., 2020**).