Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study

Anindo Saha, Joeran S Bosma*, Jasper J Twilt*, Bram van Ginneken, Anders Bjartell, Anwar R Padhani, David Bonekamp, Geert Villeirs, Georg Salomon, Gianluca Giannarini, Jayashree Kalpathy-Cramer, Jelle Barentsz, Klaus H Maier-Hein, Mirabela Rusu, Olivier Rouvière, Roderick van den Bergh, Valeria Panebianco, Veeru Kasivisvanathan, Nancy A Obuchowski, Derya Yakar, Mattijs Elschot, Jeroen Veltman, Jurgen J Fütterer, Maarten de Rooij†, Henkjan Huisman†, on behalf of the PI-CAI consortium‡*

Summary

Background Artificial intelligence (AI) systems can potentially aid the diagnostic pathway of prostate cancer by alleviating the increasing workload, preventing overdiagnosis, and reducing the dependence on experienced radiologists. We aimed to investigate the performance of AI systems at detecting clinically significant prostate cancer on MRI in comparison with radiologists using the Prostate Imaging—Reporting and Data System version 2.1 (PI-RADS 2.1) and the standard of care in multidisciplinary routine practice at scale.

Methods In this international, paired, non-inferiority, confirmatory study, we trained and externally validated an AI system (developed within an international consortium) for detecting Gleason grade group 2 or greater cancers using a retrospective cohort of 10 207 MRI examinations from 9129 patients. Of these examinations, 9207 cases from three centres (11 sites) based in the Netherlands were used for training and tuning, and 1000 cases from four centres (12 sites) based in the Netherlands and Norway were used for testing. In parallel, we facilitated a multireader, multicase observer study with 62 radiologists (45 centres in 20 countries; median 7 [IQR 5–10] years of experience in reading prostate MRI) using PI-RADS (2.1) on 400 paired MRI examinations from the testing cohort. Primary endpoints were the sensitivity, specificity, and the area under the receiver operating characteristic curve (AUROC) of the AI system in comparison with that of all readers using PI-RADS (2.1) and in comparison with that of the historical radiology readings made during multidisciplinary routine practice (ie, the standard of care with the aid of patient history and peer consultation). Histopathology and at least 3 years (median 5 [IQR 4–6] years) of follow-up were used to establish the reference standard. The statistical analysis plan was prespecified with a primary hypothesis of non-inferiority (considering a margin of 0·05) and a secondary hypothesis of superiority towards the AI system, if non-inferiority was confirmed. This study was registered at ClinicalTrials.gov, NCT05489341.

Findings Of the 10 207 examinations included from Jan 1, 2012, through Dec 31, 2021, 2440 cases had histologically confirmed Gleason grade group 2 or greater prostate cancer. In the subset of 400 testing cases in which the AI system was compared with the radiologists participating in the reader study, the AI system showed a statistically superior and non-inferior AUROC of 0·91 (95% CI 0·87–0·94; p<0·0001), in comparison to the pool of 62 radiologists with an AUROC of 0·86 (0·83–0·89), with a lower boundary of the two-sided 95% Wald CI for the difference in AUROC of 0·02. At the mean PI-RADS 3 or greater operating point of all readers, the AI system detected 6·8% more cases with Gleason grade group 2 or greater cancers at the same specificity (57·7%, 95% CI 51·6–63·3), or 50·4% fewer false-positive results and 20·0% fewer cases with Gleason grade group 1 cancers at the same sensitivity (89·4%, 95% CI 85·3–92·9). In all 1000 testing cases where the AI system was compared with the radiology readings made during multidisciplinary practice, non-inferiority was not confirmed, as the AI system showed lower specificity (68·9% [95% CI 65·3–72·4] *vs* 69·0% [65·5–72·5]) at the same sensitivity (96·1%, 94·0–98·2) as the PI-RADS 3 or greater operating point. The lower boundary of the two-sided 95% Wald CI for the difference in specificity (–0·04) was greater than the non-inferiority margin (–0·05) and a p value below the significance threshold was reached (p<0·001).

Interpretation An AI system was superior to radiologists using PI-RADS (2.1), on average, at detecting clinically significant prostate cancer and comparable to the standard of care. Such a system shows the potential to be a supportive tool within a primary diagnostic setting, with several associated benefits for patients and radiologists. Prospective validation is needed to test clinical applicability of this system.

Funding Health~Holland and EU Horizon 2020.

Lancet Oncol 2024

Published Online June 11, 2024 https://doi.org/10.1016/S1470-2045(24)00220-1

See Online/Comment https://doi.org/10.1016/S1470-2045(24)00284-5

*Joint first authors

†Contributed equally

‡For the complete list of the PI-CAI consortium members see the appendix (pp 1–2)

Diagnostic Image Analysis Group (A Saha MSc, J S Bosma MSc, Prof B van Ginneken PhD, Prof H Huisman PhD) **and Minimally Invasive Image-Guided Intervention Center** (A Saha, J J Twilt MSc, Prof J J Fütterer MD), **Department of Medical Imaging** (M de Rooij MD), **Radboud University Medical Center, Nijmegen, Netherlands; Department of Urology, Skåne University Hospital, Malmö, Sweden** (Prof A Bjartell MD); **Division of Translational Cancer Research, Lund University Cancer Centre, Lund, Sweden** (Prof A Bjartell); **Paul Strickland Scanner Centre, Mount Vernon Cancer Centre, London, UK** (Prof A R Padhani MD); **Division of Radiology, Deutsches Krebsforschungszentrum Heidelberg, Heidelberg, Germany** (Prof D Bonekamp MD); **Department of Diagnostic Sciences, Ghent University Hospital, Ghent, Belgium** (Prof G Villeirs MD); **Martini Clinic, Prostate Cancer Center, University Medical Centre Hamburg-Eppendorf, Hamburg, Germany** (Prof G Salomon MD); **Urology Unit, Santa Maria della Misericordia University Hospital, Udine, Italy**

(Prof G Giannarini MD); **Division of Artificial Medical Intelligence in Ophthalmology, University of Colorado, Aurora, CO, USA** (Prof J Kalpathy-Cramer PhD); **Department of Medical Imaging, Andros Clinics, Arnhem, Netherlands** (Prof J Barentsz MD); **Division of Medical Image Computing, Deutsches Krebsforschungszentrum Heidelberg, Heidelberg, Germany** (Prof K H Maier-Hein PhD); **Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany** (Prof K H Maier-Hein); **Departments of Radiology, Urology and Biomedical Data Science, Stanford University, Stanford, CA, USA** (M Rusu PhD); **Department of Urinary and Vascular Imaging, Hôpital Edouard Herriot, Hospices Civils de Lyon, Lyon, France** (Prof O Rouvière MD); **Faculté de Médecine Lyon-Est, Université de Lyon, Lyon, France** (Prof O Rouvière); **Department of Urology, Erasmus Medical Center, Rotterdam, Netherlands** (R van den Bergh MD); **Department of Radiological Sciences, Oncology and Pathology, Sapienza University of Rome, Rome, Italy** (Prof V Panebianco MD); **Division of Surgery and Interventional Sciences, University College London and University College London Hospital, London, UK** (V Kasivisvanathan MD); **Department of Quantitative Health Sciences and Department of Diagnostic Radiology, Cleveland Clinic Foundation, Cleveland OH, USA** (Prof N A Obuchowski PhD); **Department of Radiology, University Medical Center Groningen, Netherlands** (D Yakar MD); **Department of Radiology, Netherlands Cancer Institute, Amsterdam, Netherlands** (D Yakar); **Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Tronheim, Norway** (M Elschot PhD, Prof H Huisman); **Department of Radiology and Nuclear Medicine, St Olavs Hospital, Trondheim University Hospital,**

Research in context

Evidence before this study

We searched MEDLINE for studies published in English between May 30, 2012, and May 30, 2022, using the search terms "prostate cancer" and "MRI" and "diagnosis" or "detection" and "artificial intelligence" or "machine learning" or "deep learning" or "radiomics", with an emphasis on studies that validated performance in comparison to radiologists using paired data. We also reviewed reference lists of eligible texts. We observed that most studies were limited to single centres, small sample sizes (fewer than 1000 cases and fewer than five radiologists), and poorly defined statistics with high variability in their reported outcomes, which hindered the ability to draw any definitive conclusions. We identified no studies with confirmatory designs and no studies that provided access to both their developed algorithms and datasets, restricting transparency and reproducibility.

Added value of this study

To our knowledge, this is the first study to show evidence for viability of an artificial intelligence (AI) system for prostate cancer detection compared with the average radiologist, at scale, with a confirmatory study design, and under the multidisciplinary oversight of international experts across the patient pathway. We provided level 2b evidence that an AI system could alleviate overdiagnosis and potentially omit unnecessary biopsies within a primary diagnostic setting, with 50·4% fewer false-positive results and 20% fewer indolent cancer detections than a pool of 62 radiologists. We have publicly released our source code for analysis, the trained AI system, a subset of our training dataset (with the means to access the full training dataset), and the means for independent researchers to benchmark their algorithms across the same sequestered testing cohort in a fully masked, standardised manner to promote reproducibility, transparency, and facilitate future research in this domain.

Implications of all the available evidence

We provided evidence that AI systems, when adequately trained and validated for a target population with thousands of patient cases, could potentially support the diagnostic pathway of prostate cancer management. A clinical trial is required to determine if such a system translates to improvements in workflow efficiency, health-care equity, and patient outcomes.

Introduction

Prostate cancer is a genomically diverse disease with a broad spectrum of outcomes. Multiple trials have demonstrated that indolent (clinically insignificant) prostate cancer has a high prevalence and low cancer-specific mortality (1%).[1,2] However, aggressive (clinically significant) prostate cancer leads to advanced-stage disease, resulting in more than 375 000 deaths worldwide in 2020.[3]

MRI has an increasingly important role in the diagnostic pathway for prostate cancer and has been recommended before biopsies by clinical guidelines in Europe, UK, and the USA.[4–6] Radiologists follow the Prostate Imaging—Reporting and Data System (PI-RADS), which is a standardised approach to interpret prostate MRI examinations.[7] MRI-driven workflows can reduce unnecessary biopsies, but remain susceptible to low specificity and high inter-reader variability.[8–11]

Artificial intelligence (AI) models have matched expert clinicians in medical image analysis across several specialties, including prostate and breast cancer.[12–14] AI-assisted image interpretation can address the rising demand in medical imaging worldwide.[15–17] However, limited scientific evidence on efficacy impedes the widescale adoption of AI systems for prostate cancer diagnosis.[16,18]

We hypothesised that state-of-the-art AI models, trained using thousands of patient examinations, are non-inferior to radiologists when detecting clinically significant prostate cancer using MRI. To test this hypothesis, we designed an international, comparative study, the Prostate Imaging—Cancer Artificial Intelligence (PI-CAI) challenge. In this study, we investigated an AI system that was independently developed, trained, and externally tested for the detection of clinically significant prostate cancers using a large multicentre cohort. We compared this system to results from radiologists participating in an international reader study and radiology readings from multidisciplinary routine practice.

Methods

Study design and participants

In this international, paired, non-inferiority, confirmatory study, we combined two substudies. Algorithm developers designed AI models using 10 207 MRI cases (a cohort of 9129 patients; appendix pp 3–7). In parallel, 62 radiologists (from 45 centres in 20 countries; appendix p 21) participated in a multireader, multicase observer study. Algorithm developers and radiologists were invited to participate through referrals, outreach programmes of clinical and technical societies, presentations at conferences, and through an open call on the grand-challenge.org platform.

Preregistration and the outcomes of this study have been reported in compliance with the Biomedical Image Analysis Challenges (BIAS) guidelines, the Standards for Reporting of Diagnostic Accuracy Studies—Artificial Intelligence (STARD-AI), and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines.[19–21] The study was approved by the institutional or regional review board of each

participating centre (Prostaat Centrum Noord-Nederland: IRB 2018–597; St Olav's Hospital, Trondheim University Hospital: REK 2017/576; Radboud University Medical Center: CMO 2016–3045; and Ziekenhuisgroep Twente: ZGT 23–37). Informed consent was exempted given the retrospective scientific use of deidentified patient data.

This study included 9129 patients from four European tertiary care centres (Radboud University Medical Center; Ziekenhuisgroep Twente; Prostaat Centrum Noord-Nederland; and St. Olav's Hospital, Trondheim University Hospital). Patient data from Jan 1, 2012, to Dec 31, 2021, were retrospectively collected and de-identified by the centres individually through their institutional electronic health records. All patients were male (adults aged ≥18 years; median 66 years [IQR 61–70]) and suspected of having prostate cancer. Patients were included if they had an abnormal digital rectal examination (associated with palpable lumps, enlargements, and areas of hardness or significant pain) or at least 3 ng/mL prostate-specific antigen (in adherence with the recommendations of the European Association of Urology), or both.[9] Among them, 610 patients who were biopsy-naive underwent extended prostate biopsy protocols (ie, additional biopsies were done) through participation in clinical trials.[9,22,23] Ethnicity, race, and gender data were not recorded or considered for diagnostic decision-making in prostate cancer management during routine practice at the participating centres and this information could not be retrospectively collected (via electronic health records) or included for the purpose of this study. Patients with a history of prostate-specific treatment or at least Gleason grade group 2 findings (at the time of imaging) were excluded. Examinations with incomplete reporting or diagnostically insufficient image quality (ie, severe metal artefacts [eg, from catheters or hip prostheses], susceptibility artefacts [eg, induced by rectal gas], or motion artefacts within the prostate gland and its immediate periphery) that impeded an accurate diagnostic interpretation were excluded. More details on patients included or excluded from analysis on a per-centre basis are in the appendix (pp 3–6).

Procedures

MRI images were acquired with various commercial 1.5 Tesla or 3 Tesla scanners (Siemens Healthineers, Erlangen, Germany; Philips Medical Systems, Eindhoven, Netherlands). Images were read during clinical routine by at least one of 18 radiologists who were practicing at the participating centres between Jan 1, 2012, and Dec 31, 2021 (1–21 years of experience in reading prostate MRI). Readings were performed in compliance with PI-RADS (1.0, 2.0, or 2.1). Lesions were given a PI-RADS score between 1 and 5 to stratify the risk of prostate cancer (with higher scores indicating higher suspicion for clinically significant cancer). Patient history and peer consultation were available to aid diagnosis. Patients with positive MRI findings (ie, in whom an area with a score of PI-RADS ≥3 was identified) underwent biopsies. In the absence of abnormal areas on MRI (ie, a negative result with a maximum PI-RADS score of 1 or 2), patients were not offered a biopsy or underwent systematic biopsies exclusively. Deviations from this protocol, if any, took place due to scientific interventions, patient-specific factors, or changing clinical guidelines (appendix p 8).

Biopsies were done by urologists or radiologists, or by trained medical students, researchers, or technicians under the supervision of urologists or radiologists (depending on local practice). Two to four cores were obtained for each MRI-targeted lesion. Medial or lateral cores were obtained from each sextant of the prostate gland (six to 16 cores, in total) during systematic biopsies. Biopsy specimens were graded during clinical routine using whole-slide imaging or microscopic examination by at least one of 18 pathologists who were practicing at the participating centres between Jan 1, 2012 and Dec 31, 2021 (1–25 years of experience in reading prostate histopathology). Immunohistochemistry was available to aid tumour identification and grading. Readings were reported using Gleason scores in compliance with the International Society of Urological Pathology guidelines.[24]

Within the scope of this study, clinically insignificant cancer was defined as Gleason grade group 1 (Gleason score 6; low risk) and clinically significant cancer was defined as Gleason grade group 2–5 (Gleason score 7–10; intermediate to very high risk). If the grade group of a lesion was discordant between findings from two different biopsy methods, then the higher grade was applied and the lower grade was attributed to sampling error. In the case of patients who underwent prostatectomy, grade groups determined from whole-mount specimens were applied. We considered a follow-up period of 3 years or more (median 5 [IQR 4–6] years) to confirm the absence of clinically significant prostate cancer in all patients included for testing as negative cases. Patient outcomes were retrospectively tracked using institutional electronic health records and national registries. Details on the measures taken to control for biases and address missing outcome data are in the appendix (p 8).

Development of the AI system

The PI-CAI challenge was hosted on the grand-challenge.org platform, where it will be continually hosted for at least 5 years (ie, from May 5, 2022 until at least May 5, 2027). AI developers worldwide could opt-in, download an annotated public dataset of 1500 MRI cases, and train AI models for clinically significant prostate cancer detection at biparametric MRI. There were no restrictions on developer participation. For every examination, AI models were required to complete two tasks: localise and classify each lesion with clinically significant cancer (if any) using a 0–100 likelihood score and classify the overall case using a 0–100 likelihood

Trondheim, Norway (M Elschot); **Department of Radiology, Ziekenhuisgroep Twente, Hengelo, Netherlands** (J Veltman MD); **Department of Multi-Modality Medical Imaging, Technical Medical Centre, University of Twente, Enschede, Netherlands** (J Veltman)

Correspondence to: Anindo Saha, Diagnostic Image Analysis Group, Department of Medical Imaging, Radboud University Medical Center, Nijmegen 6525 GA, Netherlands **anindya.shaha@radboudumc.nl**

See **Online** for appendix

score for clinically significant cancer diagnosis. To this end, AI models could use imaging data and several metadata associated with the examination (ie, patient age, prostate-specific antigen level, prostate volume, and MRI scanner name) to inform their predictions. Developers could request an independent evaluation of their trained AI models on a held-out tuning cohort of 100 cases, periodically. At the end of the development cycle, each team could submit a single trained AI model for validation across a hidden testing cohort of 1000 cases in a remote, offline, fully masked setting. The hidden testing cohort included prostate MRI examinations from 1000 patients across four centres, including 197 cases from an external unseen centre. Histopathology and a follow-up period of at least 3 years were used to establish the reference standard. We independently retrained the five AI models, with the highest diagnostic performance as of Nov 28, 2022, at the central coordinating centre using 9107 cases (including a sequestered dataset of 7607 cases and the public dataset of 1500 cases). Once trained, these models were ensembled with equal weighting into a single AI system. Details on the guidelines for algorithm development, intermediary ranking schemes, and algorithm design are in the appendix (pp 9–20).

Reader study

In a multireader, multicase observer study that was hosted on the grand-challenge.org platform from Aug 8, 2022, to Feb 21, 2023, 62 radiologists (45 centres in 20 countries) read 400 multiparametric MRI examinations that were randomly sampled from the testing cohort. All readers were practising PI-RADS (2.1) in clinical routine (median 7 [IQR 5–10] years of experience in reading prostate MRI) and did not have a history of practising at one of the participating centres. 46 (74%) readers were categorised via self-reporting as experts based on the 2020 European Society of Urogenital Radiology and European Association of Urology: Section of Urological Imaging consensus statements.[25] We adopted a split-plot design, where readers and cases were randomly distributed into four blocks of 100 cases each. Each case was read in two sequential rounds, without a washout period in between. First, the same set of biparametric imaging and metadata as used for testing the AI system were made available to readers. Readers were asked to rate the case using PI-RADS (2.1) scores and an overall 0–100 likelihood score for clinically significant cancer diagnosis. Next, multiparametric imaging for the same case was shown. Readers could use this additional information to update their findings, if necessary. Readers did not have access to patient history or peer consultation. Readers were also not allowed to revisit ratings or cases, with the exception of non-compliant readings, which were revised in a single round (from March 27 to May 29, 2023), after a washout period of 5 weeks. Within the context of this study, only multiparametric MRI readings were considered for analysis (appendix pp 21–26).

Statistical analysis

The main outcomes of this study were the diagnostic performance of the AI system in comparison with that of the 62 readers and the historical radiology readings made during clinical practice. Our primary hypothesis was the non-inferiority of stand-alone AI diagnosis with respect to PI-RADS (2.1) and the standard of care. When confirmed, we tested a secondary hypothesis for the superiority of the AI system. We evaluated the diagnostic performance of the AI system and radiologists, according to their case-level predictions of clinically significant cancer. When comparing the AI system to the pool of 62 radiologists participating in the reader study, we defined the test statistic as the difference in the area under the receiver operating characteristic curve (AUROC) metric. When comparing the AI system to the radiology reads made during multidisciplinary practice, we defined the test statistic as the difference in specificity at the same sensitivity as the PI-RADS 3 or greater threshold (according to standard diagnostic criteria). Non-inferiority was concluded if the test statistic was greater than zero and the lower boundary of its two-sided 95% Wald CI was greater than –0·05. If non-inferiority was concluded, then the superiority of the AI system over radiologists was assessed and concluded if the lower boundary of the two-sided 95% CI for the test statistic was greater than zero. Details of the statistical analysis plan and the power analysis used for sample size deduction are in the appendix (pp 27–30). The statistical analysis plan was prespecified and independently reviewed by an expert biostatistician (NO). In a post-hoc analysis, we revisited patient history, imaging, and histopathology outcomes for all MRI examinations that were marked as false positives or false negatives in the same manner by all radiologists participating in the reader study and the AI system. Any discordant findings were adjudicated by MdR (11 years of experience in reading prostate MRI), with respect to the reference standard. Significance thresholds for p values were corrected for multiplicity via Holm's method in a hierarchical adaptive order. Statistical analyses were conducted using Python (3.10) and R (4.2.2). This study was conducted in compliance with the institutional data monitoring committees of the participating centres. This study was registered with ClinicalTrials.gov, NCT05489341.

Role of the funding source

The funders of this study had no role in study design, data collection, data analysis, data interpretation, writing of the report, or any aspect pertinent to the study.

Results

Between June 12, and Nov 28, 2022, a total of 839 individuals (from 53 countries) opted-in to the

	Public training and development set			Sequestered training set			Hidden tuning cohort			Hidden testing cohort				Total
	RUMC	ZGT	PCNN	RUMC	ZGT	PCNN	RUMC	ZGT	PCNN	RUMC	ZGT	PCNN	STOH	
Sites	2	1	8	2	1	8	2	1	3	2	1	4	1	12
Patients	792	346	338	5010	1413	312	40	30	30	336	262	205	197	9129
Median age, years	65 (60–69)	67 (62–72)	68 (63–72)	66 (61–70)	68 (62–72)	67 (62–71)	64 (58–70)	66 (61–71)	66 (60–74)	66 (60–70)	64 (59–68)	67 (63–72)	65 (60–69)	66 (61–70)
Median prostate-specific antigen, ng/mL	9 (6–14)	7 (5–11)	9 (6–12)	8 (5–11)	7 (5–11)	9 (6–12)	8 (5–11)	8 (6–11)	9 (6–14)	7 (5–10)	7 (5–9)	9 (6–13)	7 (5–12)	8 (5–11)
Median prostate volume, mL	63 (45–88)	49 (36–70)	50 (35–70)	67 (47–96)	51 (37–74)	45 (33–65)	64 (46–91)	46 (35–54)	42 (30–65)	63 (46–87)	54 (41–73)	48 (35–63)	50 (36–72)	61 (43–88)
MRI scanners														
Manufacturer	Siemens Healthineers	Siemens Healthineers	Siemens Healthineers, Philips Medical Systems	Siemens Healthineers	Siemens Healthineers	Siemens Healthineers, Philips Medical Systems	Siemens Healthineers	Siemens Healthineers	Siemens Healthineers, Philips Medical Systems	Siemens Healthineers	Siemens Healthineers	Siemens Healthineers, Philips Medical Systems	Siemens Healthineers	Siemens Healthineers, Philips Medical Systems
Model	Avanto, Prisma, Skyra, TrioTim	Skyra	Achieva, Aera, Avanto, Ingenia, Prisma, Skyra	Avanto, Prisma, Skyra, TrioTim	Skyra	Achieva, Aera, Avanto, Espree, Ingenia, Intera, Prisma, Skyra	Prisma, Skyra	Skyra	Achieva, Aera, Avanto, Ingenia, Prisma, Skyra	Prisma, Skyra	Skyra	Achieva, Aera, Avanto, Ingenia, Prisma, Skyra	Skyra	Achieva, Aera, Avanto, Espree, Ingenia, Intera, Prisma, Skyra, TrioTim
Field strength, Tesla	1.5, 3	3	1.5, 3	1.5, 3	3	1.5, 3	3	3	1.5, 3	3	3	1.5, 3	3	1.5, 3
Coil type	PPA	PPA	PPA	PPA	PPA	PPA	PPA	PPA	PPA	PPA	PPA	PPA	PPA	PPA
Cases	800	350	350	5778	1501	328	40	30	30	336	262	205	197	10 207
Benign or indolent prostate cancer*	564 (71%)	270 (77%)	241 (69%)	4624 (80%)	1127 (75%)	214 (65%)	24 (60%)	18 (60%)	17 (57%)	245 (73%)	183 (70%)	121 (59%)	119 (60%)	7767 (76%)
Clinically significant prostate cancer (Gleason grade group ≥2)	236 (30%)	80 (23%)	109 (31%)	1154 (20%)	374 (25%)	114 (35%)	16 (40%)	12 (40%)	13 (43%)	91 (27%)	79 (30%)	84 (41%)	78 (40%)	2440 (24%)
Positive MRI lesions	614	186	287	2650	760	263	21	25	33	187	188	176	128	5518
PI-RADS 3	149 (24%)	32 (17%)	65 (23%)	348 (13%)	81 (11%)	58 (22%)	4 (19%)	3 (12%)	7 (21%)	38 (20%)	32 (17%)	27 (15%)	29 (23%)	873 (16%)
PI-RADS 4	226 (37%)	71 (38%)	141 (49%)	1335 (50%)	338 (44%)	138 (52%)	10 (48%)	7 (28%)	17 (52%)	82 (44%)	83 (44%)	97 (55%)	33 (25%)	2578 (47%)
PI-RADS 5	239 (39%)	83 (45%)	81 (28%)	967 (36%)	341 (45%)	67 (26%)	7 (33%)	15 (60%)	9 (27%)	67 (36%)	73 (39%)	52 (30%)	66 (52%)	2067 (37%)

(Table continues on next page)

	Public training and development set			Sequestered training set			Hidden tuning cohort			Hidden testing cohort			Total
	RUMC	ZGT	PCNN	RUMC	ZGT	PCNN	RUMC	ZGT	PCNN	RUMC	ZGT	PCNN	STOH	
(Continued from previous page)														
ISUP-based lesions	411	163	202	2164	772	222	25	25	23	207	165	151	129	4659
Gleason grade group 1	150 (36%)	74 (45%)	87 (43%)	837 (39%)	325 (42%)	100 (45%)	6 (24%)	13 (52%)	8 (35%)	93 (45%)	77 (47%)	61 (40%)	23 (18%)	1854 (40%)
Gleason grade group 2	136 (33%)	46 (28%)	78 (39%)	642 (30%)	256 (33%)	77 (35%)	8 (32%)	7 (28%)	8 (35%)	60 (29%)	60 (36%)	49 (32%)	39 (30%)	1466 (31%)
Gleason grade group 3	64 (16%)	21 (13%)	24 (12%)	285 (13%)	106 (14%)	30 (14%)	5 (20%)	1 (4%)	4 (17%)	30 (14%)	10 (6%)	27 (18%)	39 (30%)	646 (14%)
Gleason grade group 4	28 (7%)	6 (4%)	7 (3%)	183 (8%)	33 (4%)	7 (3%)	2 (8%)	1 (4%)	2 (8%)	5 (2%)	4 (3%)	7 (5%)	13 (10%)	298 (6%)
Gleason grade group 5	33 (8%)	16 (10%)	6 (3%)	217 (10%)	52 (7%)	8 (4%)	4 (16%)	3 (12%)	1 (4%)	19 (9%)	14 (8%)	7 (5%)	15 (12%)	395 (8%)

Data are n, n (%), or median (IQR), unless otherwise specified. ISUP=International Society of Urological Pathology. PCNN=Prostaat Centrum Noord-Nederland. PPA=pelvic phased-array surface coil. PI-RADS=Prostate Imaging Reporting and Data System. PSA=prostate-specific antigen. RUMC=Radboud University Medical Center. STOH=St Olav's Hospital, Trondheim University Hospital. ZGT=Ziekenhuisgroep Twente. *Based on negative MRI (PI-RADS ≤2) or negative histopathology (Gleason grade group ≤1). A follow-up period of ≥3 years (median 5 [IQR 4–6] years) was used to confirm the reference standard for each patient in the hidden tuning and testing cohorts.

Table: **Summary of patient distribution across the training datasets and the hidden tuning and testing cohorts**

development of the AI system and 293 AI algorithms were submitted—of which, the top five highest performing algorithms were deep learning models developed by teams primarily based at the University of Sydney (Sydney, NSW, Australia), University of Science and Technology (Hefei, China), Guerbet Research (Villepinte, France), Istanbul Technical University (Istanbul, Türkiye), and Stanford University (Stanford, CA, USA; appendix pp 12–20). We trained and tested the resultant AI system using 10 207 MRI examinations (median age 66 [IQR 61–70] years; median prostate-specific antigen level 8 [5–11] ng/mL), where 2440 cases with histologically confirmed, clinically significant prostate cancer were observed (table).

In the subset of 400 cases from the testing cohort that was used to facilitate the reader study, the AI system showed an AUROC of 0·91 (95% CI 0·87–0·94). The AI system passed the prespecified criteria for non-inferiority (with a lower boundary of the two-sided 95% Wald CI for the difference in AUROC of 0·02), and furthermore, showed superior case-level diagnosis (p<0·0001) compared with the pool of 62 radiologists with an AUROC of 0·86 (95% CI 0·83–0·89; figure). In comparison with the mean PI-RADS of 3 or greater operating point of all readers, the AI system detected 6·8% (nine of 133) more clinically significant cancers at the same specificity (57·7%, 95% CI 51·6–63·3). The system resulted in 50·4% (57 of 113) fewer false-positive results and detected 20·0% (eight of 40) fewer Gleason grade group 1 cancers at the same sensitivity as radiologists (89·4%, 95% CI 85·3–92·9; appendix pp 32–33). On average, radiologists had a positive predictive value of 53·2% (95% CI 47·0–59·3) and a negative predictive value of 90·2% (85·2–94·1) at the mean PI-RADS 3 or greater operating point. The AI system showed a positive predictive value of 68·0% (60·5–74·6) and a negative predictive value of 93·8% (90·6–96·7), when the threshold was adjusted to match the same sensitivity (89·4%) as this operating point (appendix pp 31–32).

In all 1000 cases from the testing cohort, the AI system had an AUROC of 0·93 (95% CI 0·91–0·94). The system did not show non-inferiority, given its lower specificity (68·9%, 65·3–72·4) compared with the radiology reads made during multidisciplinary practice (69·0%, 65·5–72·5), when the threshold was adjusted to match the same sensitivity (96·1%, 94·0–98·2) as the PI-RADS 3 or greater operating point. However, the lower boundary of the two-sided 95% Wald CI for the difference in specificity (–0·04) was within the non-inferiority margin (–0·05) and a p value below the significance threshold was obtained (p<0·001; figure). Radiology reads had a positive predictive value of 60·6% (56·5–64·8) and a negative predictive value of 97·3% (95·8–98·7) at the PI-RADS 3 or greater operating point. The AI system showed a positive predictive value of 60·5% (95% CI 56·4–64·7) and the same negative predictive value of

97·3% (95·8–98·7) when the threshold was adjusted to match the same sensitivity (96·1%) as this operating point (appendix pp 31–32).

In a post-hoc analysis, we observed that no clinically significant cancer was missed by all radiologists and the AI system (ie, no examination was marked as a false negative by all groups). However, in 3·5% of patients (14 of 400), all radiologists in a given reader block and the AI system at the PI-RADS 3 or greater threshold detected the same false positive findings for which the reference standard diagnoses were negative. During routine practice, 13 of these patients had been reported with PI-RADS 3 or greater lesions. Subsequently, 12 individuals were diagnosed with Gleason grade group 1 cancer using systematic and MRI-targeted biopsies (eight) or prostatectomy specimens (four). Two remaining individuals did not exhibit any signs of prostate cancer, based on biopsy specimens and follow-up analysis. A review of these 14 discordant findings was repeated in conjunction with site investigators and an expert radiologist (MdR; 11 years of experience in reading prostate MRI) at the central coordinating centre, who did not participate in the reader study or the original historical reads. Abnormal MRI findings were attributed to indolent cancer and non-cancerous confounders (eg, granulomatous prostatitis). We concluded that such cases warrant active surveillance or follow-up and were rightfully recommended for biopsies during practice (in concurrence with the retrospective diagnoses made by readers and the AI system).

Discussion

To our knowledge, PI-CAI is the first international diagnostic accuracy study to assess radiologists and a stand-alone AI system in detecting clinically significant prostate cancer on MRI at scale. The PI-CAI challenge showed that a state-of-the-art AI system was superior in discriminating patients with clinically significant prostate cancer at biparametric MRI compared with the mean of 62 radiologists using PI-RADS (2.1) within an international reader study. When comparing the AI system to the standard of care in routine practice, the AI system was not found to be non-inferior; however, the observed performance gap was 0·1% in specificity at the same sensitivity. We hypothesise that this difference in performance between the radiologists participating in the reader study and the radiologists reporting in practice was due to those reporting in practice having access to patient history (including previous prostate-specific antigen levels and imaging and biopsy outcomes), peer consultation (or multidisciplinary team meetings), and protocol familiarity. We recommend that future studies investigate multimodal prostate-AI systems that factor in continuous health data across the complete patient pathway to improve performance further.[26,27]

We presented level 2b evidence that an AI system might safely omit unnecessary biopsies within a primary diagnostic setting. In a cohort of 400 patients, the AI system generated 50·4% fewer false-positives and detected 20.0% fewer indolent cancers (associated with several benefits for the patient [eg, avoiding post-biopsy haematoma or infection, discomfort, and anxiety]), and detected the same number of clinically significant cancers as the pool of 62 radiologists at their PI-RADS 3 or greater operating point. Predictive values observed for the AI system were high (89·5% sensitivity at 79·1% specificity; 93·8% negative predictive value at an estimated 33% prevalence) in comparison with that of radiologists at multiparametric MRI in the PROMIS trial (88% sensitivity at 45% specificity; 76% negative

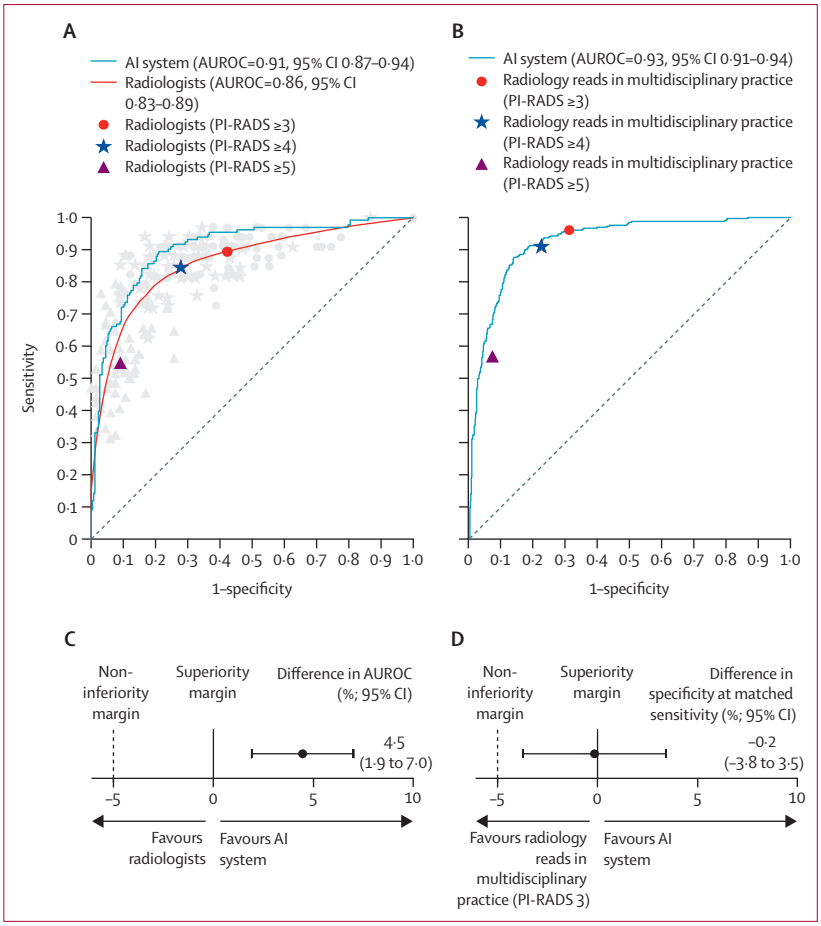

Figure: **Performance of the AI system at clinically significant prostate cancer diagnosis in the hidden testing cohort**

(A) Receiver operating characteristic curves of the AI system and the pool of 62 radiologists, considering the subset of 400 testing cases used to facilitate the reader study. Light grey circle, star, and triangle markers indicate the PI-RADS operating points of each individual radiologist. The diagonal dashed line represents the receiver operating characteristic curve for a random classifier with an AUROC of 0·50. (B) Receiver operating characteristic curve of the AI system and the PI-RADS operating points of the radiology reads made during multidisciplinary routine practice, considering all 1000 testing cases. The diagonal dashed line represents the receiver operating characteristic curve for a random classifier with an AUROC of 0·50. (C) Difference in the AUROC metric between the AI system and the pool of 62 radiologists, considering the subset of 400 testing cases used to facilitate the reader study. (D) Difference in specificity when the threshold of the AI system was adjusted to match the same sensitivity (96·1%) as the PI-RADS 3 or greater operating point of the radiology reads made during multidisciplinary routine practice, considering all 1000 testing cases. AI=artificial intelligence. AUROC=area under the receiver operating characteristic curve. PI-RADS=Prostate Imaging Reporting and Data System.

predictive value at an estimated 53% prevalence), and as reported in two separate meta-analyses of 42 studies (90·8% negative predictive value) and 3857 patients (96% sensitivity at 29% specificity).[28–30] We advise caution when interpreting such findings across studies, owing to their different populations, comparators, outcomes, and study designs. We recommend further investigation before deploying to practice (eg, simulating paired reading configurations for risk management).

Our study has some limitations. First, the dataset was retrospectively curated over several years and multiple sites. This resulted in a mix of consecutive patients and convenience samples. Second, radiologists participating in the reader study provided their analysis for retrospective data (ie, their diagnosis would not have influenced patient outcomes) through a controlled, online reading environment that might have differed significantly from their native workstation. Third, biopsy planning and histological verification for each case were guided by the original radiology readings and not by the prospective readings made during the reader study or the predictions made by the AI system. Fourth, this study is limited by differential verification bias (ie, where all patient examinations are verified but multiple standards [eg, biopsies, prostatectomies, and follow-up] are combined to establish the presence or absence of significant cancer). Fifth, this study did not record patient data on ethnicity, and 93·4% of all included MRI examinations were acquired from one MRI manufacturer. Whether the results will be reproducible under other circumstances is not known.

We observed that a deep learning-based AI system, which was trained using thousands of biparametric MRI examinations, was superior at discriminating Gleason grade group 2 or greater prostate cancer compared with the pool of 62 radiologists using PI-RADS (2.1) on average. Prospective validation (eg, as stated in the CHANGE trial[26]) is required to test clinical applicability.

Contributors

AS, JSB, JJT, BvG, MdR, JJF, and HH conceptualised and designed the study, with input from investigators at the participating centres (DY, ME, JV) and the scientific advisory board (AB, ARP, DB, GV, GS, GG, JKC, KHM-H, NAO, OR, RvdB, VP, VK). JSB, JJT, and AS did the statistical analysis, with input from NAO. JSB, JJT, and AS directly accessed and verified the underlying data reported in the manuscript. AS, JSB, JJT, MdR, and HH were involved in data interpretation. AS wrote the first draft of the report with input from JSB, JJT, BvG, MdR, JJF, and HH. All authors revised the report and provided important intellectual content. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

NAO provides statistical consultation to Siemens Healthineers, Takeda, and Qure, and serves as a committee member of the Eastern Cooperative Oncology Group–American College of Radiology Imaging Network, the Tomosynthesis Mammographic Imaging Screening Trial, and the National Cancer Institute's Clinical Imaging Steering Committee (Bethesda, MD, USA). AB has been a consultant and advisor for Astellas and Bayer; board membership, officer, and trustee for Glactone Pharma, and LIDDS Pharma; has received lecture honoraria for Accord, Astellas, AstraZeneca, Bayer, Ipsen, Janssen, and Merck; has participated in trials run by Astellas, Ferring, and Janssen; and holds stock in Glactone Pharma, LIDDS Pharma, Noviga, and WntResearch. BvG holds stocks in and is a founder of Thirona. JKC has received research funding from GE Healthcare and Genentech and is the co-inventor of software that has been licensed to Siloam Vision. JKC has equity ownership in Siloam Vision. GS has been an advisory board member of Exact Imaging and Angiogenesis and has received lecture honorarium from Hitachi. OR has received funding for travel expenses from Philips Medical Systems. ARP has received research funding from Siemens Healthineers, holds stocks in Lucida Medical, and has received lecture honoraria for Siemens Healthineers and Bayer. HH has received research funding from Siemens Healthineers and Canon Medical Systems. GV has been a clinical advisory board member of AGFA Healthcare. VK has received lecture honoraria on prostate cancer diagnosis from the European Association of Urology and Singapore Urology Association and has received research funding from Prostate Cancer UK and the John Black Charitable Foundation. DB has received lecture honorarium from Bayer Vital and holds stocks in NVIDIA, Microsoft, and MSCI-World ETF. RvdB has been an advisory board member for Janssen; has received lecture honoraria from Amgen, Astellas, Ipsen, Janssen, and MSD; has received research support from Astellas and Janssen; and has participated in trials run by Janssen. All other authors declare no competing interests.

Data sharing

We provide access to the full dataset of 10 207 MRI examinations curated during this study in different ways (as listed on https://pi-cai.grand-challenge.org/). A deidentified, annotated subset of 1500 cases from the training dataset, our complete codebase for analysis, and the trained artificial intelligence system that was investigated in this study, have been made publicly available for scientific research. We have provided the means for independent researchers and developers to access the full sequestered training dataset of 9107 cases or benchmark their algorithms across the sequestered testing cohort of 1000 cases, in a fully masked, standardised manner with investigator support, on a per-application basis. Our study protocol and statistical analysis plan are also available online.

Acknowledgments

This study was funded by the European Commission (EU Horizon 2020: ProCAncer-I project) and Health~Holland (LSHM20103). This study was endorsed by the European Association of Urology, the European Society of Urogenital Radiology, the Medical Image Computing and Computer Assisted Intervention Society, and the Medical Imaging with Deep Learning Foundation. We thank the staff at Amazon SageMaker for their technical support in facilitating the study. MR is in receipt of funding from the National Cancer Institute (Bethesda, MD, USA) of the National Institutes of Health (R37CA260346).

References

1 Hamdy FC, Donovan JL, Lane JA, et al. 15-Year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. *N Engl J Med* 2023; **388:** 1547–58.
2 Godtman RA, Holmberg E, Khatami A, Stranne J, Hugosson J. Outcome following active surveillance of men with screen-detected prostate cancer. Results from the Göteborg randomised population-based prostate cancer screening trial. *Eur Urol* 2013; **63:** 101–07.
3 Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021; **71:** 209–49.
4 Mottet N, van den Bergh RCN, Briers E, et al. EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer—2020 update. *Eur Urol* 2021; **79:** 243–62.
5 National Institute for Health and Care Excellence. NICE guidance—prostate cancer: diagnosis and management. *BJU Int* 2019; **124:** 9–26.
6 Eastham JA, Auffenberg GB, Barocas DA, et al. Clinically localized prostate cancer: AUA/ASTRO guideline part I: introduction, risk assessment, staging and risk-based management. *J Urol* 2022; **208:** 10–18.
7 Turkbey B, Rosenkrantz AB, Haider MA, et al. Prostate Imaging Reporting and Data System version 2.1: 2019 update of Prostate Imaging Reporting and Data System version 2. *Eur Urol* 2019; **76:** 340–51.

8 Ahdoot M, Wilbur AR, Reese SE, et al. MRI-targeted, systematic, and combined biopsy for prostate cancer diagnosis. *N Engl J Med* 2020; **382:** 917–28.

9 van der Leest M, Cornel E, Israël B, et al. Head-to-head comparison of transrectal ultrasound-guided prostate biopsy versus multiparametric prostate resonance imaging with subsequent magnetic resonance-guided biopsy in biopsy-naïve men with elevated prostate-specific antigen: a large prospective multicenter clinical study. *Eur Urol* 2019; **75:** 570–78.

10 Rouvière O, Puech P, Renard-Penna R, et al. Use of prostate systematic and targeted biopsy on the basis of multiparametric MRI in biopsy-naive patients (MRI-FIRST): a prospective, multicentre, paired diagnostic study. *Lancet Oncol* 2019; **20:** 100–09.

11 Westphalen AC, McCulloch CE, Anaokar JM, et al. 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:** 76–84.

12 Milea D, Najjar RP, Zhubo J, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. *N Engl J Med* 2020; **382:** 1687–95.

13 Bulten W, Kartasalo K, Chen PC, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat Med* 2022; **28:** 154–63.

14 Lång K, Josefsson V, Larsson AM, et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol* 2023; **24:** 936–44.

15 James ND, Tannock I, N'Dow J, et al. The *Lancet* Commission on prostate cancer: planning for the surge in cases. *Lancet* 2024; **403:** 1683-1722.

16 van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol* 2021; **31:** 3797–804.

17 Angus DC. Randomized clinical trials of artificial intelligence. *JAMA* 2020; **323:** 1043–45.

18 Suarez-Ibarrola R, Sigle A, Eklund M, et al. Artificial intelligence in magnetic resonance imaging-based prostate cancer diagnosis: where do we stand in 2021? *Eur Urol Focus* 2022; **8:** 409–17.

19 Maier-Hein L, Reinke A, Kozubek M, et al. BIAS: Transparent reporting of biomedical image analysis challenges. *Med Image Anal* 2020; **66:** 101796.

20 Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nat Med* 2020; **26:** 807–08.

21 von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* 2007; **370:** 1453–57.

22 Krüger-Stokke B, Bertilsson H, Langørgen S, Sjøbakk TAE, Bathen TF, Selnæs KM. Multiparametric prostate MRI in biopsy-naïve men: a prospective evaluation of performance and biopsy strategies. *Front Oncol* 2021; **11:** 745657.

23 Wagensveld IM, Osses DF, Groenendijk PM, et al. A prospective multicenter comparison study of risk-adapted ultrasound-directed and magnetic resonance imaging-directe diagnostic pathways for suspected prostate cancer in biopsy-naïve men. *Eur Urol* 2022; **82:** 318–26.

24 Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system. *Am J Surg Pathol* 2016; **40:** 244–52.

25 de Rooij M, Israël B, Tummers M, et al. ESUR/ESUI consensus statements on multi-parametric MRI for the detection of clinically significant prostate cancer: quality requirements for image acquisition, interpretation and radiologists' training. *Eur Radiol* 2020; **30:** 5404–16.

26 Rouvière O, Souchon R, Lartizien C, et al. Detection of ISUP ≥2 prostate cancers using multiparametric MRI: prospective multicentre assessment of the non-inferiority of an artificial intelligence system as compared to the PI-RADS V.2.1 score (CHANGE study). *BMJ Open* 2022; **12:** e051274.

27 Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med* 2022; **28:** 1773–84.

28 Ahmed HU, El-Shater Bosaily A, Brown LC, et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *Lancet* 2017; **389:** 815–22.

29 Sathianathen NJ, Omer A, Harriss E, et al. Negative predictive value of multiparametric magnetic resonance imaging in the detection of clinically significant prostate cancer in the Prostate Imaging Reporting and Data System era: a systematic review and meta-analysis. *Eur Urol* 2020; **78:** 402–14.

30 Woo S, Suh CH, Kim SY, Cho JY, Kim SH. Diagnostic performance of Prostate Imaging Reporting and Data System version 2 for detection of prostate cancer: a systematic review and diagnostic meta-analysis. *Eur Urol* 2017; **72:** 177–88.