Assessing the clinical performance of artificial intelligence software for prostate cancer detection on MRI

Tobias Penzkofer¹ • Anwar R Padhani² • Baris Turkbey³ • Hashim U Ahmed⁴

Received: 5 January 2022 / Revised: 5 January 2022 / Accepted: 20 January 2022 / Published online: 23 February 2022
© The Author(s), under exclusive licence to European Society of Radiology 2022

Multiple researchers are developing computer-aided detection (CAD) algorithms to enable the detection of clinically significant prostate cancer (csPCa) on MRI [1–3]. Commercial vendors are offering early versions of CAD software that claim to improve radiologists' performance and productivity by decreasing the observer variability for detecting suspicious lesions while decreasing time-intensive reporting and data processing tasks. These developments are to be welcomed by radiologists, urologists and healthcare administrators, as we face the increased demand for MRI-influenced prostate biopsies. As software developments gather momentum, developers need to provide clinically relevant metrics to enable comparisons between algorithms, commercial products and the current standard of multi-disciplinary care [2].

It is important to remember that CAD software detects lesions that are likely to represent csPCa thus representing targets for biopsy. The benefits of the software on biopsy procedures arising from improved true radiologic detections or when no lesions are found, need to be weighed against the harms of false alarms and when radiologically important lesions are missed leading to a false sense of security. CAD-influenced biopsy benefits and harms depend on how the results are used for planning biopsy

procedures, which vary according to the urological and patient tolerance to false results.

CAD software when used for radiologic triage of prostate MRI scans requires high sensitivity (low-false negative rate) for detecting men with important cancers (usually defined as grade group (GG) ≥ 2) [4]. A $\geq 90\%$ patient-level sensitivity for GG ≥ 2 is a suitable benchmark against which the patient-level false-positive rates should be used to discriminate between software (Table 1). Lower false-positive rates will ensure that fewer men are biopsied overall. Expert human readers can achieve a detection sensitivity of $> 90\%$ resulting in 1 in 3 biopsy-naïve men avoiding biopsy after a negative MRI for a disease prevalence of 30–50% [4, 5].

For men with suspicious MRI scans, accurate localization of all positive targets is necessary to ensure appropriate tissue sampling. A suitable lesion-level metric is the false detection rate in men deemed to be positive by experienced radiologists [2]. Generally, to mitigate over-diagnoses, there need to be low false-detections to decrease the number of lesions sampled or biopsy cores taken [6]. The number of acceptable false detections depends on whether the urologist or patient is biopsy versus cancer averse [7] which determines their tolerance to false results (Table 1).

With these general considerations in mind, Hosseinzadeh et al reported on the performance of their deep-learning (DL) CAD model for the automated detection and classification of lesions that are likely to harbour csPCa lesions [8, 9]. Their multistage architecture reduced the false-positive detection rates while maintaining high sensitivity for the presence of high-suspicion lesions using bi-parametric MRI. When used on an external biopsy-confirmed testing dataset of 296 men who all underwent 12-core systematic biopsy and those with radiologist determined positive scans underwent additional in-bore MRI targeted biopsies, the DL-CAD system achieved a lesion-level detection sensitivity for PI-RADS 4-5 lesions of 87% (95% CI: 82–91) at an average of 1 false-positive detection per patient. The level of false-positive detections per case was higher than a panel of experienced radiologists. The DL-

This comment refers to the article available at <https://doi.org/10.1007/s00330-021-08320-y>

✉ Anwar R Padhani
anwar.padhani@stricklandscanner.org.uk

¹ Department of Radiology, Charité Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany

² Paul Strickland Scanner Centre, Mount Vernon Cancer Centre, Rickmansworth Road, Northwood, Middlesex HA6 2RN, UK

³ Molecular Imaging Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

⁴ Division of Surgery, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London SW7 2AZ, UK

Table 1 Performance requirements for computer-aided detection (CAD) software for prostate MRI readings at patient and lesion level

Performance requirements		Standard
Lesion-level	Document high suspicion lesions against false-positive detections (false-alarms) on a case-by-case basis	• FROC against expert determined positive lesions (PI-RADS 3-5 and 4-5 lesions)
	Detected suspicious lesions should be considered by expert radiologists likely to be GG ≥ 2 cancers, with few false-positive detections per case for disease prevalence of 30–50%	• < 1 false detections/case in men considered negative by experts • 1–1.5 false detections/case @ detection rate of $\geq 90\%$ in men considered positive by experts • Performance levels according to clinical and patient preferences and tolerance to false detections
	Improved the consistency of lesion detections between readers	• Less variability for true detections compared to average radiologists in multi-reader studies
Patient-level	Document accuracy for high suspicion patient detection against false-positive detections	• ROC against expert determined PI-RADS 3-5 and 4-5 patients
	Document accuracy for csPCa detection against false-positive rates	• ROC curves against the histologic standard of ISUP GG ≥ 2
	High-negative predictive value at high-detection sensitivity	• Detection sensitivity > 90% for GG ≥ 2 • False-positive rate < 50% • Patient-level NPV > 90% for GG ≥ 2
	Improved the consistency of GG ≥ 2 detections between readers	• Less variability compared to average radiologists in multi-reader studies • Inter- and intra-reader variability of NPV/sensitivity
Special use case	Aid in the discrimination of indeterminate MRI cases for selecting men who need biopsies	• < 10% missed GG ≥ 2 in no biopsy group • Additive value to PSAD in multivariable analyses

* Increased consistency of PPV for GG ≥ 2 cancers is not considered a standard because biopsy targeting accuracy is not the exclusive domain of a CAD system

ISUP, International Society of Urological Pathology; GG, grade group; ROC, receiver operating characteristic; FROC, free-response operating characteristic; NPV, negative predictive value; PSAD, prostate-specific antigen density

CAD had 1.67 false-positive detections per patient at a sensitivity of 90% versus the experts who had a detection sensitivity of 91% at 0.3 false-positive detections. Here we should note that experienced radiologists do consider clinical factors that CAD software does not, so a performance degradation is expected. This means that the standalone performance of the system does not in its current form qualify as a radiological expert system, but is a good candidate to assist radiologists.

Nevertheless, the DL-CAD performance did generalize into a patient-level detection accuracy for GG ≥ 2 cancers after systemic and targeted biopsies of 86% (95% CI: 77–83). At a patient-level sensitivity of 90%, the false-positive detection rate was 50%. This histologic level of performance matched the performance of general radiologists who participated in the readings of the prospective MRI-First study [10] but is below the performance of the study's expert readers (91% sensitivity at a false-positive detection rate of 23%), and of the central readers of the 4M study [11]. At this operating point, there was a moderate agreement between their DL-CAD with expert radiologists ($\kappa = 0.53$) and histopathology readings ($\kappa = 0.50$) [8]. Here we should note that when radiologists and CAD agree on the likely presence of suspicious lesions, the positive predictive value for GG ≥ 2 cancers increases without affecting the negative predictive value [12].

However, we must acknowledge the limitations arising from the retrospective nature of the current study which impacts both the training and the subsequent performance of the CAD software. Specifically, we should note that only the lesions seen by the radiologists were biopsied and thus used for training the CAD software. Extended systematic and targeted biopsies would be better for algorithm learning [12]. In a similar vein, prostatectomy histology learning also introduces bias because many men do not undergo prostatectomy after MRI-influenced biopsies [2]. Furthermore, radiologists' determinations on the likely nature of the CAD detected lesions as used in the current study are not accurate because of the dependence of predictive value for csPCa on the PI-RADS category. Despite this, a radiologist's determined false-detection rate per case metric can still be useful because fewer suspicious lesions can help lower biopsy core numbers and over-diagnosis rates [8, 9].

While the documentation of diagnostic performance is commendable, Hosseinzadeh et al do not provide design-related information that would enable the usefulness of their CAD to be judged as a radiological tool. Additional items include software impacts on radiologic interpretations and reading efficiency. Some of the items have been provided by Winkel et al [13] who achieved an accuracy for detecting PI-RADS 4-5 lesions of 86%, which is comparable to

Hosseinzadeh et al of 87%. For the detection of PI-RADS 4–5 lesions in 100 cases, the accuracy of 7 radiologists was 0.84 (95% CI, 0.79–0.89) without CAD, improving by 4.4% (95% CI, 1.1–7.7%; $p = 0.01$). Inter-reader concordance also increased (kappa 0.22 without CAD vs. 0.36 when using the software).

The impact of software on time-intensive biopsy-related tasks such as gland and target outlining and report generation also requires documentation. Winkel et al [13] showed reductions in reading times by 21% (from 103 s to 81 s/case), but this improvement needs to be judged in the context of the time to generate a complete radiological report. Other useful CAD functionality features include levels of suspicion of user-defined ROIs.

To conclude, the heterogeneity of data analyses in the published research impedes the wide clinical adoption of CAD for prostate cancer diagnosis [3]. Consistent reporting will enable comparisons to be made with the current standard of PI-RADS-based multidisciplinary care. Performance metrics focusing on the sensitivity of radiological lesion detections and csPCa in patients should be described. Studies should also evaluate whether CAD systems can remove the need for radiologists to review non-suspicious MRI scans when incorporated with patient-related meta-data. Data on the ability of software to support radiologists by reducing interobserver variability and the times for interpretation, reporting and biopsy outlining tasks are also needed. Ideally, this should be done in prospective, multicentre, validation and cost-effectiveness studies where human and CAD detected lesions are adequately sampled.

Funding The authors state that this work has not received any funding.

Declarations

Guarantor The scientific guarantor of this publication is Prof. Anwar R Padhani.

Conflict of interest The authors declare no competing interests.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was not required for this study because this is an invited editorial and not the subject of a research article.

Ethical approval Institutional Review Board approval was not required because this is an editorial piece and not subject to a research article.

Methodology

- Editorial opinion

References

- Twilt JJ, van Leeuwen KG, Huisman HJ et al (2021) Artificial intelligence based algorithms for prostate cancer classification and detection on magnetic resonance imaging: a narrative review. *Diagnostics* (Basel) 11:959. <https://doi.org/10.3390/diagnostics11060959>
- Turkbey B, Haider MA (2021) Artificial intelligence (AI) for automated cancer detection on prostate MRI: opportunities and ongoing challenges, from the AJR Special Series on AI Applications. *AJR Am J Roentgenol*. <https://doi.org/10.2214/AJR.21.26917>
- Syer T, Mehta P, Antonelli M et al (2021) Artificial intelligence compared to radiologists for the initial diagnosis of prostate cancer on magnetic resonance imaging: a systematic review and recommendations for future studies. *Cancers* (Basel) 13:3318. <https://doi.org/10.3390/cancers13133318>
- Mottet N, van den Bergh RCN, Briers E et al (2021) EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer-2020 update. Part 1: Screening, diagnosis, and local treatment with curative intent. *Eur Urol* 79:243–262. <https://doi.org/10.1016/j.eururo.2020.09.042>
- Drost F-JH, Osses D, Nieboer D et al (2020) Prostate magnetic resonance imaging, with or without magnetic resonance imaging-targeted biopsy, and systematic biopsy for detecting prostate cancer: a Cochrane systematic review and meta-analysis. *Eur Urol* 77: 78–94. <https://doi.org/10.1016/j.eururo.2019.06.023>
- Penzkofer T, Padhani AR, Turkbey B et al (2021) ESUR/ESUI position paper: developing artificial intelligence for precision diagnosis of prostate cancer using magnetic resonance imaging. *Eur Radiol* 31: 9567–9578. <https://doi.org/10.1007/s00330-021-08021-6>
- Van Calster B, Wynants L, Verbeek JFMM et al (2018) Reporting and interpreting decision curve analysis: a guide for investigators. *Eur Urol* 74:796–804. <https://doi.org/10.1016/j.eururo.2018.08.038>
- Saha A, Hosseinzadeh M, Huisman H (2021) End-to-end prostate cancer detection in bpMRI via 3D CNNs: effects of attention mechanisms, clinical priori and decoupled false positive reduction. *Med Image Anal* 73:102–155. <https://doi.org/10.1016/j.media.2021.102155>
- Hosseinzadeh M, Saha A, Brand P, Slootweg I, de Rooij M, Huisman H (2021) Deep learning-assisted prostate cancer detection on bi-parametric MRI: minimum training data size requirements and effect of prior knowledge. *Eur Radiol* 20:153–159. <https://doi.org/10.1007/s00330-021-08320-y>
- Rouvière O, Puech P, Renard-Penna R et al (2019) Use of prostate systematic and targeted biopsy on the basis of multiparametric MRI in biopsy-naïve patients (MRI-FIRST): a prospective, multicentre, paired diagnostic study. *Lancet Oncol* 20:100–109. [https://doi.org/10.1016/S1470-2045\(18\)30569-2](https://doi.org/10.1016/S1470-2045(18)30569-2)
- van der Leest M, Cornél E, Israël B et al (2019) 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* 75:570–578. <https://doi.org/10.1016/j.eururo.2018.11.023>
- Schell P, Kohl S, Radtke JP et al (2019) Classification of cancer at prostate MRI: deep learning versus clinical PI-RADS assessment. *Radiology* 293:607–617. <https://doi.org/10.1148/radiol.2019190938>
- Winkel DJ, Tong A, Lou B et al (2021) A novel deep learning based computer-aided diagnosis system improves the accuracy and efficiency of radiologists in reading biparametric magnetic resonance images of the prostate: results of a multireader, multicase study. *Invest Radiol* 56:605–613. <https://doi.org/10.1097/RLI.0000000000000780>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.