The ReIMAGINE consortium – establishing an infrastructure for the external validation of prostate MRI lesion classification models

Natasha Thorley[1,2], Tom Syer[1,3], Swetha Srikanthan[4], Jacob Antunes[4], Thomas Parry[1], Teresa Marsden[5], Rosemary Clow[1], Aida Santaolalla[6], Mrishta Brizmohun Appayya[1], Giorgio Brembilla[1], Chris Brew-Graves[1], Zhe Min[7], Yipeng Hu[7], David Atkinson[1], Sue Mallett[1], Steve Rodney[4], Paul Jacobs[4], Jonathan Piper[4], Hashim U Ahmed[8,9], Mark Emberton[5], and Shonit Punwani[1,2]

[1]Centre for Medical Imaging, University College London, London, United Kingdom, [2]Imaging Department, University College London Hospital NHS Foundation Trust, London, United Kingdom, [3]Department of Radiology, University of Cambridge, Cambridge, United Kingdom, [4]MIM Software Inc, Cleveland, OH, United States, [5]Division of Surgical and Interventional Science, University College London, London, United Kingdom, [6]School of Cancer and Pharmaceutical Sciences, King's College London, London, United Kingdom, [7]Dept of Med Phys & Biomedical Eng, University College London, London, United Kingdom, [8]Division of Surgery, Imperial College London, London, United Kingdom, [9]Imperial Urology, Imperial College Healthcare NHS Trust, London, United Kingdom

Synopsis

Keywords: Prostate, Prostate

Motivation: Multiparametric MRI is highly sensitive for identifying clinically significant prostate cancer (csPCa), but has a poorer specificity, meaning many men undergo unnecessary prostate biopsies.

Goal(s): To evaluate whether artificial intelligence (AI) could improve the diagnostic accuracy of MRI compared to current clinical methods, including Likert score and PSA density (PSAd).

Approach: We carried out independent evaluation of a prostate MRI lesion classifier model using a large multisite and multivendor prostate MRI dataset (1,039 patients).

Results: The AI model matched the sensitivity and specificity of Likert score plus PSAd cut-offs on data similar to the training set, but did not generalise to other data.

Impact: An infrastructure has been successfully established to allow robust and independent evaluation of prostate MRI lesion classification models to accelerate the development of such tools and to ensure adequate testing pre-deployment.

Introduction

International guidelines recommend multiparametric magnetic resonance imaging (MRI) as the first-line investigation for people with suspected prostate cancer, followed by MRI-directed biopsies in those with suspicious lesions on MRI[1–3]. This pathway has a high sensitivity and negative predictive value for clinically significant prostate cancer (csPCa), but a much poorer specificity and positive predictive value[4]. Consequently, many patients undergo unnecessary biopsy or are diagnosed with clinically insignificant disease, leading to biopsy-related complications or over-treatment.

Artificial intelligence (AI) offers one potential solution; currently published AI algorithms are promising in their ability to classify lesions on MRI accurately, but many are limited by study design flaws, using small, single-centre patient cohorts and lacking robust external validation in representative populations[5]. In addition, the sparsity of open-source algorithms often prevents independent evaluation, particularly for commercial AI products. The 2023 AI Safety Summit specifically highlighted the importance of external testing and included a plan for governments and AI companies to collaboratively test the safety of models pre-deployment, stating "we shouldn't rely upon [AI companies] to mark their own homework"[6].

Through the ReIMAGINE consortium[7], an infrastructure has been established to allow robust independent evaluation of commercial prostate MRI lesion classification models. The results of the first consortium member submission [MIM-PCa-Radiomics-Prototype] are presented here.

Methods

ReIMAGINE Risk[7] is a prospective observational cohort study of patients with suspicious lesions on MRI undergoing standard-of-care MRI-directed biopsy between September 2019 and April 2022. The dataset included baseline data, MRI-derived data, DICOMs (axial T2-weighted images, high *b*-value diffusion-weighted images, apparent diffusion coefficient (ADC) map and dynamic contrast-enhanced images; with segmentation of the whole prostate, central gland and targeted lesions on T2 images) and histopathological data. 1,039 patients were recruited from three sites, with MRIs from all major vendors and both 1.5T and 3.0T field strengths.

For model testing, the dataset was split and the first 201 patients from site 1 were shared with partners for model development and internal validation. The remaining patients formed a withheld test set, including a temporally separated set from site 1 and two external sites.

The prostate lesion classification model aimed to predict the presence of csPCa (Gleason grade ≥3+4) within lesions using the annotated DICOMs and baseline clinical data, with a minimum per-patient sensitivity of 90%[8]. The submitted model [MIM-PCa-Radiomics-Prototype] utilised a random forest decision tree classifier with 20 radiomic features from T2, high *b*-value diffusion and ADC images (Figure 1).

Results

The final test set comprised 761 patients and 1,126 lesions (Figure 2). Descriptors of the final test set are not presented as these continue to be used for validation of algorithms from other ReIMAGINE partners.

Evaluation of the prototype on the withheld site 1 test set achieved per-patient sensitivity of 93% [95%CI:88-96], specificity 38% [95%CI:29-49] and AUC 0.73 [95%CI:0.67-0.80]. This was comparable to the performance of Likert+PSAd (Table 1). Performance decreased when applied to the two external sites with variation between individual scanners (Table 2).

External evaluation on the full withheld test set demonstrated per-patient sensitivity of 91% [95%CI:88-33], specificity 26% [95%CI:21-31] and AUC 0.67 [95%CI:0.64-0.71]. Compared to clinical comparators, the prototype had lower specificity than Likert+PSAd, but higher specificity than Likert≥3 grading alone (Table 1, Figure 3).

Discussion

The submitted prototype achieved the minimum accepted sensitivity to reduce biopsy rates safely, and its performance was comparable to Likert+PSAd when tested on data similar to the training set. However, when applied to a broader test set, it showed performance variability between sites and individual scanners. The drop in performance is most likely attributed to differences in acquisition parameters across sites, which directly affects the computation of radiomic features. This reinforces the importance of training AI models with data that encompasses the full scope of scanning parameters expected in future clinical settings.

Creating a large, multi-scanner, clinically representative dataset and establishing a consortium and infrastructure to allow independent evaluation of commercial AI algorithms ensures models are adequately tested pre-deployment and offers an opportunity to accelerate model development. Other consortium members are in the process of submitting further algorithms which will be presented and allow direct comparison.

Conclusion

AI algorithms can potentially improve the diagnostic accuracy of the MRI-guided prostate cancer diagnostic pathway. However, to establish if performance can surpass current clinical methods and generalise to clinical practice, it is necessary to train these models on heterogenous data that is representative of the target population and engage in robust independent and external evaluation. The ReIMAGINE dataset, consortium and infrastructure have been set up and successfully utilised to aid in developing AI tools for prostate MRI lesion classification.

Acknowledgements

ReIMAGINE Prostate Cancer Risk Study (NCT04060589, IRAS 251166)

The ReIMAGINE Consortium was launched with funding of £4.1 m from the Medical Research Council Grant no: MR/R014043/1 and £1 m from Cancer Research UK, as part of the MRC's Stratified Medicine Initiative.

The ReIMAGINE Study Group:

Eric Aboagye, Hashim Ahmed, Fatima Akbar, Gerhardt Attard, Teresita Beeston, Charlotte Bevan, Chris Brew-Graves, Mrishta Brizmohun, Paul Boutros, Giorgio Brembilla, Louise Brown, Joey Clemente, Rosie Clow, Ton Coolen, Ged Corbett, Caroline Dive, Eytan Domany, Mark Emberton, Andrew Feber, Elena Frangou, Alex Freeman, Francesco Giganti, Miriam Gonclaves, Fiona Gong, Saran Green, Joanna Hadley, Ashling Henderson, Elizabeth Isaac, Richard Kaplan, Douglas Kopcke, Stefano Lise,Annabel Kunzemann Martinez, Teresa Marsden, Malcolm Mason, Neil McCartan, Caroline Moore, Charlotte Moss, Kinnari Naik, Anwar Padhani, Peter Parker, Chris Parker, Shonit Punwani, Nahian Rahman, Francesca Rawlins, Manuel Rodriguez-Justo, Mark Rowley, Aida Santa-Olalla, Harbir Sidhu, Pirruntha Sivaharan, Katerina Soteriou, Andrew Stubbs, Tom Syer, Suparna Thakali, Steve Tuck, Mieke Van Hemelrijck, Anna Wingate, Daniel Wetterskog, Hayley Whitaker, Savahnna Wolfe

References

1. A Collaborative Initiative by the American Urological Association and the Society of Abdominal Radiology Prostate Disease Focus Panel. Standard Operating Procedure for Multiparametric Magnetic Resonance Imaging in the Diagnosis, Staging and Management of Prostate Cancer. 2018.
2. European Association of Urology. Guidelines on Prostate Cancer. 2023.
3. National Institute for Health and Care Excellence (NICE). Prostate cancer: diagnosis and management (NG131). 2019.
4. Mazzone E, Stabile A, Pellegrino F, Basile G, Cignoli D, Cirulli GO, et al. Positive Predictive Value of Prostate Imaging Reporting and Data System Version 2 for the Detection of Clinically Significant Prostate Cancer: A Systematic Review and Meta-analysis. Eur Urol Oncol. 2021;4(5).
5. Syer T, Mehta P, Antonelli M, Mallett S, Atkinson D, Ourselin S, et al. Artificial intelligence compared to radiologists for the initial diagnosis of prostate cancer on magnetic resonance imaging: A systematic review and recommendations for future studies. Vol. 13, Cancers. 2021.
6. Prime Minister's Office 10 Downing Street, Department for Science I and T, The Rt Hon Michelle Donelan MP, The Rt Hon Rishi Sunak MP. Press release: World leaders, top AI companies set out plan for safety testing of frontier as first global AI Safety Summit concludes [Internet]. 2023 [cited 2023 Nov 3]. Available from: https://www.gov.uk/government/news/world-leaders-top-ai-companies-set-out-plan-for-safety-testing-of-frontier-as-first-global-ai-safety-summit-concludes#:~:text=Today%20we've%20reached%20a,with%20countries%20around%20the%20world.
7. Marsden T, McCartan N, Brown L, Rodriguez-Justo M, Syer T, Brembilla G, et al. The ReIMAGINE prostate cancer risk study protocol: A prospective cohort study in men with a suspicion of prostate cancer who are referred onto an MRI-based diagnostic pathway with donation of tissue, blood and urine for biomarker analyses. PLoS One. 2022;17(2 February).
8. Penzkofer T, Padhani AR, Turkbey B, Ahmed HU. Assessing the clinical performance of artificial intelligence software for prostate cancer detection on MRI. Vol. 32, European Radiology. 2022.

Figures

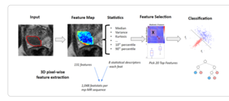

Figure 1 – [MIM-PCa-Radiomics-Prototype] Pipeline Overview: machine learning classification via radiomic features.

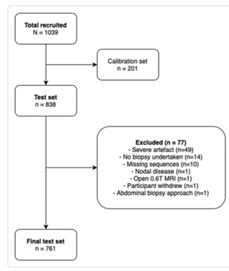

Figure 2 - Flowchart of patient recruitment with reasons for exclusion.

Predictor	Test Set	Sensitivity (95%CI)	Specificity (95%CI)	AUC (95%CI)
MIM model	Site 1 test set (n=237)	93 (88, 96)	38 (29, 49)	0.73 (0.67, 0.80)
MIM model	Full set (n=761)	91 (88, 93)	26 (21, 31)	0.67 (0.64, 0.71)
Clinical comparator				
Likert ≥3	Full set	100 (99, 100)	3 (2, 4)	
Likert ≥4 or PSAd ≥0.10	Full set	96 (94, 98)	28 (23, 33)	
Likert ≥4 or PSAd ≥0.12	Full set	93 (91, 95)	34 (29, 40)	
Likert ≥4 or PSAd ≥0.15	Full set	91 (88, 93)	43 (38, 49)	
Likert ≥4	Full set	83 (79, 86)	52 (47, 57)	

Table 1 – Per-patient diagnostic accuracy of the submitted model and clinical comparators for the final test set.

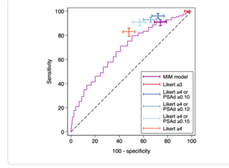

Figure 3 – Receiver operating characteristic curve of the submitted model in the final test set, model operating point and clinical comparator are marked with crosshairs with the centre at the associated sensitivity/specificity and shoulders of 95% confidence intervals.

Subgroup	N	Sensitivity (95% CI)	Specificity (95% CI)	AUC (95% CI)
Site				
1 Internal	237	93 (88, 96)	38 (29, 49)	0.73 (0.67, 0.80)
2 External	371	90 (85, 93)	22 (16, 29)	0.65 (0.60, 0.71)
3 External	153	88 (79, 93)	18 (11, 29)	0.63 (0.55, 0.72)
Field strength (T)				
1.5	518	89 (85, 92)	31 (25, 37)	0.69 (0.64, 0.73)
3	243	95 (90, 97)	15 (10, 23)	0.67 (0.60, 0.74)
MR Scanner Model				
Vendor 1, Model 1 1.5T	322	89 (83, 93)	36 (29, 45)	0.69 (0.64, 0.75)
Vendor 1, Model 2 3T	124	91 (81, 96)	17 (9, 28)	0.64 (0.54, 0.74)
Vendor 1, Model 3 1.5T	61	79 (64, 89)	18 (7, 39)	0.63 (0.48, 0.78)
Vendor 1, Model 4 3T	47	96 (80, 99)	9 (3, 28)	0.70 (0.55, 0.85)
Vendor 2, Model 1 3T	40	100 (87, 100)	29 (12, 55)	0.73 (0.57, 0.89)
Vendor 1, Model 5 1.5T	41	95 (78, 99)	16 (6, 38)	0.61 (0.44, 0.79)
Vendor 2, Model 1 1.5T	24	90 (60, 98)	21 (8, 48)	0.80 (0.57, 1.00)
Other	102	96 (88, 98)	17 (8, 33)	0.66 (0.55, 0.77)

Table 2 - Per-patient subgroup analysis of diagnostic accuracy by site, field strength, scanner.