


## RESEARCH ARTICLE

## Open Access



# Expected impact of MRI-related interreader variability on ProScreen prostate cancer screening trial: a pre-trial validation study

Ronja Hietikko<sup>1,2\*</sup> , Tuomas P. Kilpeläinen<sup>1,2</sup>, Anu Kenttämies<sup>2,3</sup>, Johanna Ronkainen<sup>4</sup>, Kirsty Ijäs<sup>3</sup>, Kati Lind<sup>3</sup>, Suvi Marjasuo<sup>3</sup>, Juha Oksala<sup>4</sup>, Outi Oksanen<sup>3</sup>, Tuomas Saarinen<sup>4</sup>, Ritja Savolainen<sup>3</sup>, Kimmo Taari<sup>1</sup>, Teuvo L. J. Tammela<sup>5</sup>, Tuomas Mirtti<sup>2,6</sup>, Kari Natunen<sup>7</sup>, Anssi Auvinen<sup>7</sup> and Antti Rannikko<sup>1,2</sup>

## Abstract

**Background:** The aim of this study is to investigate the potential impact of prostate magnetic resonance imaging (MRI) -related interreader variability on a population-based randomized prostate cancer screening trial (ProScreen).

**Methods:** From January 2014 to January 2018, 100 men aged 50–63 years with clinical suspicion of prostate cancer (PCa) in Helsinki University Hospital underwent MRI. Nine radiologists individually reviewed the pseudonymized MRI scans of all 100 men in two ProScreen trial centers. All 100 men were biopsied according to a histological composite variable comprising radical prostatectomy histology ( $N = 38$ ) or biopsy result within 1 year from the imaging ( $N = 62$ ). Fleiss' kappa ( $\kappa$ ) was used to estimate the combined agreement between all individual radiologists. Sample data were subsequently extrapolated to 1000-men subgroups of the ProScreen cohort.

**Results:** Altogether 89% men of the 100-men sample were diagnosed with PCa within a median of 2.4 years of follow-up. Clinically significant PCa (csPCa) was identified in 76% men. For all PCa, mean sensitivity was 79% (SD  $\pm$  10%, range 62–96%), and mean specificity 60% (SD  $\pm$  22%, range 27–82%). For csPCa (Gleason Grade 2–5) MRI was equally sensitive (mean 82%, SD  $\pm$  9%, range 67–97%) but less specific (mean 47%, SD  $\pm$  20%, range 21–75%). Interreader agreement for any lesion was fair ( $\kappa$  0.40) and for PI-RADS 4–5 lesions it was moderate ( $\kappa$  0.60). Upon extrapolating these data, the average sensitivity and specificity to a screening positive subgroup of 1000 men from ProScreen with a 30% prevalence of csPCa, 639 would be biopsied. Of these, 244 men would be true positive, and 395 false positive. Moreover, 361 men would not be referred to biopsy and among these, 56 csPCas would be missed. The variation among the radiologists was broad as the least sensitive radiologist would have twice as many men biopsied and almost three times more men would undergo unnecessary biopsies. Although the most sensitive radiologist would miss only 2.6% of csPCa (false negatives), the least sensitive radiologist would miss every third.

**Conclusions:** Interreader agreement was fair to moderate. The role of MRI in the ongoing ProScreen trial is crucial and has a substantial impact on the screening process.

**Keywords:** Prostate cancer, Agreement, Magnetic resonance imaging, PI-RADS version 2, Screening

\* Correspondence: [Ronja.hietikko@hus.fi](mailto:Ronja.hietikko@hus.fi)

<sup>1</sup>Department of Urology, University of Helsinki and Helsinki University Hospital, PL900, 00029 HUS, Helsinki, Finland

<sup>2</sup>Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Early detection of aggressive prostate cancer (PCa) remains challenging. Prostate-specific antigen (PSA) -based screening reduces cancer-specific mortality by approximately 20% by detecting aggressive cancers at an early stage when they can be successfully treated. However, such screening also leads to overdiagnosis of clinically insignificant cancers that are likely to be subsequently overtreated [1]. Therefore, organized PCa screening has not been implemented in Europe.

Traditionally, the standard procedure for men with a clinical suspicion of PCa, with elevated PSA or abnormal digital rectal examination, has been the systematic 10- or 12-core transrectal ultrasound (TRUS) -guided biopsy [2, 3]. The limitations of this approach are that some cases of clinically significant PCa (csPCa) are not detected. In contrast, many cases of clinically insignificant PCa (cisPCa) are overdiagnosed using this approach, and all men with a clinical suspicion of PCa are required to undergo invasive and harmful biopsy procedure [4, 5].

Multiparametric magnetic resonance imaging (mpMRI) of the prostate and targeted biopsies of only lesions identified is a promising diagnostic pathway. A recent study by Drost and colleagues has shown that MRI improves the detection ratio (DR) of csPCa by 12% and decreases the risk of cisPCa diagnosis by 30–40% compared to systematic biopsies in men with a suspected PCa [5]. Of men with a clinical suspicion of PCa, roughly one-third have negative MRI and can therefore avoid prostate biopsy [5]. Thus, the MRI pathway is an appealing tool for PCa screening, as it may be possible to maintain the substantial reduction in PCa mortality and yet avoid the unnecessary biopsies and overdiagnosis of cisPCa.

However, the usefulness of MRI in a population-based screening is highly dependent on the quality of the MRI process (imaging and reporting) per se. The ability of MRI to detect csPCa has improved over the past decade as the Prostate Imaging Reporting and Data System (PI-RADS) was introduced in 2012 [6] and has been updated twice since then [7]. Nevertheless, several uncertainties remain, as the interreader agreement on PI-RADS categories is moderate at best and the experience of an individual radiologist may have an effect on the specificity of reporting [8].

The aim of our study is to evaluate the potential impact of MRI -related interreader variability in an ongoing ProScreen PCa screening trial. We initiated a population-based prospective randomized screening trial (ProScreen) in 2018, which is still ongoing. In ProScreen men with a suspicion of csPCa in a biochemical screening test (PSA and the four kallikrein test (4K) (free PSA, intact and total PSA and kallikrein-like peptidase 2 [hK2]) are referred to mpMRI with targeted biopsies of the visual lesion(s) only [9].

The precision of the ProScreen trial at detecting csPCa while avoiding overdiagnosis, ultimately depends on the

subjective evaluation of the MRI by the radiologist. In this current study, we specifically investigated the potential impact of prostate MRI-related interreader variability on the ProScreen trial.

## Methods

The ProScreen trial is a population-based screening trial with a total of 67,000 men aged 50–63 years who reside in Helsinki or Tampere in Finland that commenced in 2018 [9]. These men are randomized to either a screening arm or a control arm in a 1:3 ratio. The men in the screening arm are invited to consent for the trial and upon giving their informed consent, a serum PSA test is taken, whereas the men in the control arm will not be contacted. A PSA  $\geq 3.0$   $\mu\text{g/l}$  value is considered abnormal and will trigger the next stage of screening, i.e. 4K score test [10]. Men with a 4K score  $\geq 7.5\%$  are referred for prostate MRI in one of the participating urology departments. Men with lesions that have a PI-RADS of 3–5 upon MRI are then invited for transrectal ultrasound guided fusion biopsies (FBx) of the target lesions only. Men with negative MRI are invited for TRUS guided systematic 12-core biopsies only when PSA density is  $\geq 0.15$   $\mu\text{g/l}$ .

Here, we chose a retrospective sample of 100 non-consecutive men who had been referred to the Helsinki University Hospital (HUS) for suspicion of PCa before the initiation of the ProScreen trial to come up with different GGG classes of roughly equal size. Previous MRI or negative biopsies were allowed. Men had varying baseline risks for PCa. Most men ( $n = 91$ ) had undergone MRI before diagnostic biopsies, whereas for nine men the MRI was used post-biopsy in cancer staging before definitive treatment. The 91 men were biopsied within 6 months of the MRI. The mean age at imaging was 67 years (SD  $\pm 9$ ) and the median PSA level was 9.4  $\mu\text{g/l}$  (interquartile range [IQR] 6.7–14.5  $\mu\text{g/l}$ ).

The imaging was performed with 3T scanners Philips Achieva (from 2014) and with Siemens Skyra (from 2017). The protocol included T2 weighted imaging (T2WI), diffusion (DWI) with ADC-mapping and dynamic contrast enhancement (DCE). Surface coil was used and the slice thickness was 3 mm for T2WI and DWI, and 4 mm for DCE. The image resolutions for T2WI were  $0,6 \times 0,6$  mm (Skyra) or  $0,6 \times 0,7$  (Achieva). ADC-maps were calculated from diffusion b-values 0 (Achieva) or 50 (Skyra), 100 and 800. The high b-value images, b2000, for tumour detection were scanned separately (Achieva) or extrapolated up to b1600 by using lower b-value data (Skyra). DCE imaging comprised intravenous administration of gadolinium-based contrast agent (Dotarem<sup>®</sup>, 0,2 ml/kg, 2 ml/s) with the temporal resolution of 7 s (Skyra) or 8 s (Achieva) up to 2 min 30 s, and flip angle 12° (Skyra) or 10° (Achieva). The possible early enhancement was assessed visually, and the data were further processed by using scanner's software (until

2015) or DynaCad to create signal intensity curves of suspected lesions (Fig. 1). Until October 2017 only summary reports of the DCE analyses were stored as jpg-images. These could not be pseudonymized adequately. Hence, for the re-evaluation in the present study, the original DCE data were not available. MRI images were first pseudonymized. Subsequently, the images of all 100 men were each made available to all nine radiologists that were concurrently reporting prostate MRI in the two ProScreen trial centers. Previous experience of the nine radiologists regarding prostate MRI reads varied from 40 to 100 reads (one radiologist) to 100–300 reads (one radiologist) to > 500 reads (seven radiologists) (see Table 1). The PSA concentration, the age of the patient and the 4K score data were the only additional data the radiologists had during assessment. The radiologists were blind to all other relevant data regarding the patients of the sample. The MRI scans were evaluated using version 2 PI-RADS [11] but the DCE image sets were not available, thus the re-evaluation is based on only biparametric MRI (bpMRI).

Structured pathological assessment was given using the five-tier Gleason Grade Groups (GGG): 3 + 3; 3 + 4; 4 + 3; 4 + 4 and > 8 [12]. We focused on the evaluation of the index lesion, which was defined as the largest and highest-grade lesion in the prostate [13]. The gold standard in sensitivity and specificity analyses was a histological composite variable: radical prostatectomy histology (for those who underwent radical prostatectomy,  $n = 38$ ) or biopsy result within 1 year from the imaging (for those who did not undergo radical prostatectomy,  $n = 62$ ). As the positive predictive value (PPV) and negative predictive value (NPV) are highly dependent on the underlying prevalence

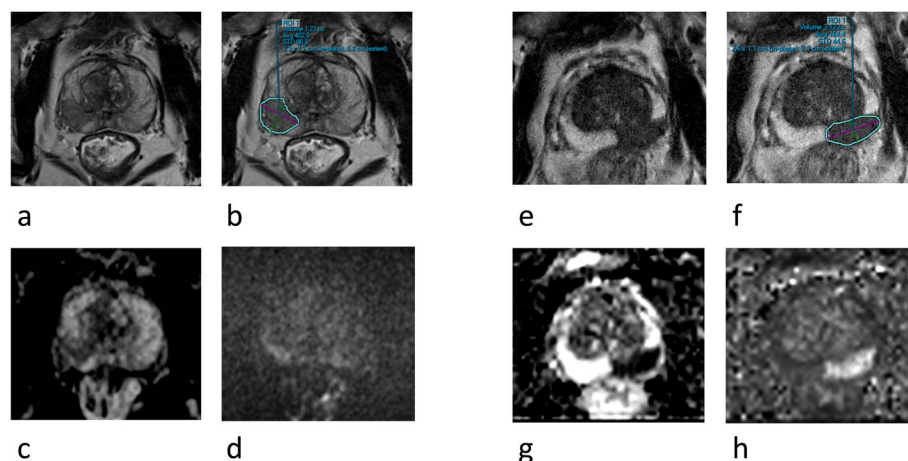
of the condition, we also report positive and negative likelihood (LH) ratios.

Fleiss' kappa was used to estimate the combined agreement between all individual radiologists. Finally, the distribution of PI-RADS score was presented graphically for each patient stratified by composite pathological result. The study protocol was evaluated by the research ethics committee of the HUS Helsinki University Hospital (HUS/333/2019).

## Results

The median follow-up time for men with negative MRI or negative biopsies was 2.4 years (range 1.5–5.4 years and mean 2.6 years, SD  $\pm 0.82$  years). Twenty-two men (22%) had both targeted and standard 12-core biopsies, 68 men (68%) had only targeted biopsies, nine (9%) men had only standard 12-core biopsies and one man had saturation biopsies. Most men ( $n = 87$ , 87%) had PCa upon biopsy. Of the 13 men (13%) with benign biopsy, two (15.4%) were later diagnosed with PCa. Of these two, one had GGG1 cancer diagnosed by transurethral resection of the prostate and the other one had GGG2 cancer diagnosed by subsequent saturation biopsies.

Of the 89 men diagnosed with PCa, 13 (14.6%) had GGG1 cancer, 31 (34.8%) had GGG2 cancer, 28 (31.5%) had GGG3 cancer, 7 (7.9%) had GGG4 cancer and 10 (11.2%) had GGG5 cancer in the biopsies. Thus, 76% of the patients in the cohort sample had clinically significant cancer (GGG2 or worse) and roughly half had GGG3 cancer or worse. Patient level assessments grouped by pathological result are illustrated in Fig. 2.



**Fig. 1** Prostate MRI images from two patients demonstrating poor agreement (images a-d) and good agreement (images e-h). T2 (a and e), T2 with delineated lesion (b and f), ADC (c and g) and high b (d and h) images are shown. In a 64-year old man with PSA of 22 ng/ml (a-d) five radiologist scored a lesion and four did not (fusion biopsies were benign and no PCa has been diagnosed during a 2.5-year follow-up). In a 72-year old man with PSA of 10.5 ng/ml all nine radiologist correctly scored a lesion (GG5 in fusion biopsies)

**Table 1** Comparison of radiological findings and agreement between individual radiologists and clinical reference

Comparison of radiological findings and agreement between individual radiologists and clinical reference.								
	Mean prostate size, cm <sup>3</sup>	SD	Mean Index lesion size, cm <sup>3</sup>	Percentage of PIRADS 3-5 index lesion	Percentage of PIRADS 4-5 index lesion	Percentage of T3-4 disease	Radiologist experience, as N of previous assessments	Radiologist experience, as N of years of previous experience
Radiologist 1	49	( 30 )	1.8 (2.5)	67 %	54 %	16 %	> 500	>5 years
Radiologist 2	47	( 31 )	2.3 (4.1)	69 %	56 %	18 %	> 500	2-5 years
Radiologist 3	47	( 31 )	1.7 (2.0)	66 %	63 %	22 %	> 500	2-5 years
Radiologist 4	51	( 29 )	2.1 (3.0)	79 %	61 %	23 %	> 500	2-5 years
Radiologist 5	52	( 30 )	1.8 (2.4)	82 %	64 %	33 %	> 500	>5 years
Radiologist 6	50	( 30 )	2.4 (3.1)	76 %	61 %	28 %	> 500	>5 years
Radiologist 7	61	( 37 )	1.9 (2.5)	93 %	70 %	30 %	> 500	>5 years
Radiologist 8	49	( 29 )	3.3 (5.3)	85 %	75 %	38 %	40-100	1-2 years
Radiologist 9	52	( 36 )	2.5 (3.5)	57 %	48 %	30 %	100-300	1-2 years
Mean among radiologists 1-9	51		2.2	75 %	61 %	26 %		
Agreement* among radiologists 1-9				0.40	0.60	0.49		

\* Fleiss' Kappa

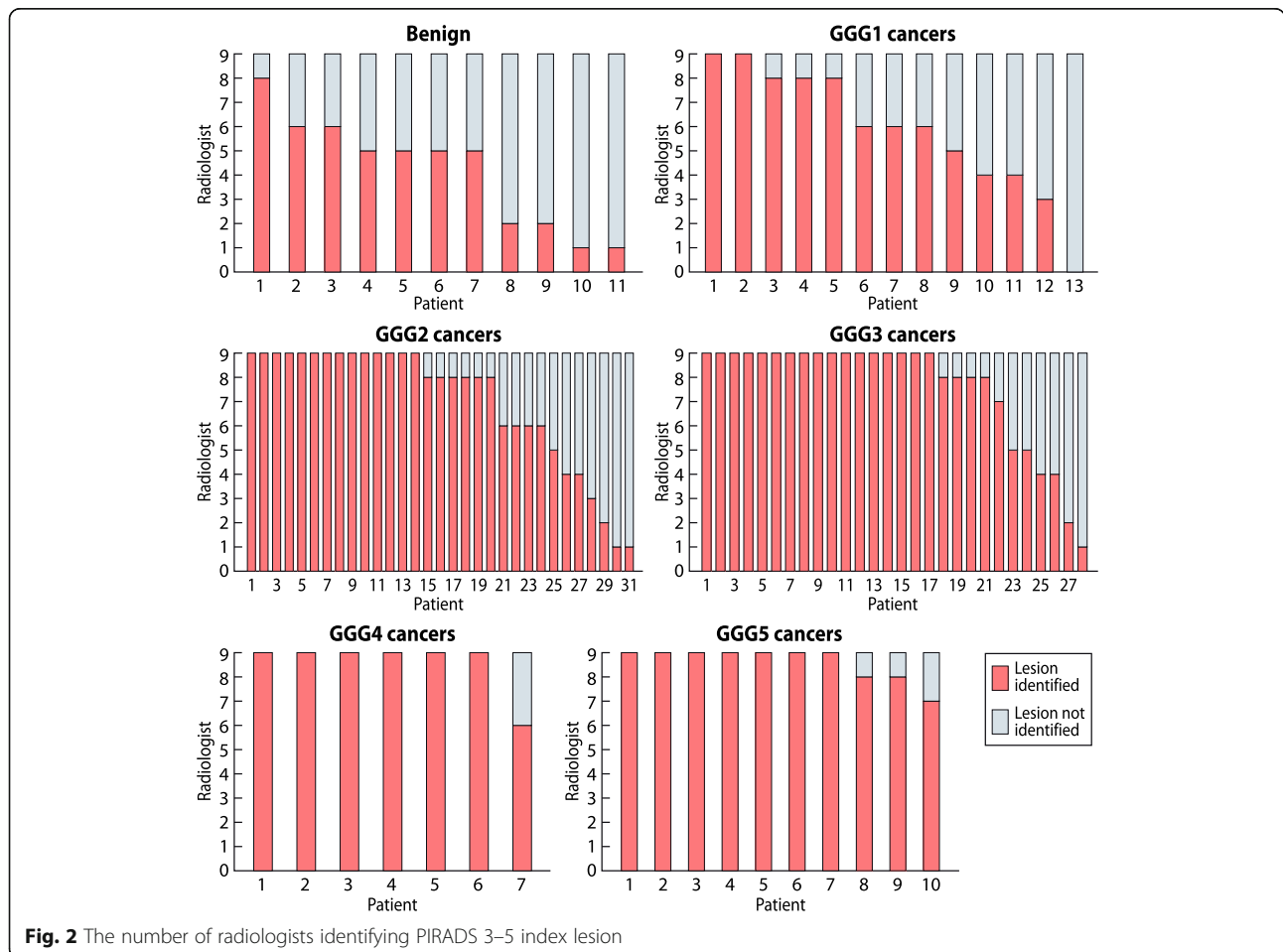
**Agreement between radiologists**

Individual radiologists reported a PI-RADS 3–5 index lesion in 75% of cases (range 57–93%) (Table 1). Agreement among radiologists was fair (Fleiss' kappa 0.40). Agreement was moderate (Fleiss' kappa 0.60) for PI-RADS 4–5 lesions, which were found in 61% of cases (range 48–75%) (Table 1). T3–4 disease was reported in

26% of cases (range 16–38%) and agreement among radiologists was moderate (Fleiss' kappa 0.49) (Table 1).

**Radiological assessment compared to the pathological reference**

The radiological assessment compared to the pathological reference is presented in Table 2. Mean positive



**Fig. 2** The number of radiologists identifying PIRADS 3–5 index lesion

**Table 2** Sensitivity, specificity, positive predictive value and negative predictive value for PIRADS 3–5 or PIRADS 4–5 index lesions

<b>A</b> Any prostate cancer, PIRADS 3-5 index lesion					<b>B</b> Any prostate cancer, PIRADS 4-5 index lesion				
	Sensitivity	Specificity	PPV	NPV		Sensitivity	Specificity	PPV	NPV
Radiologist 1	73 %	82 %	97 %	27 %	Radiologist 1	60 %	91 %	98 %	22 %
Radiologist 2	74 %	73 %	96 %	26 %	Radiologist 2	63 %	100 %	100 %	25 %
Radiologist 3	72 %	82 %	97 %	26 %	Radiologist 3	69 %	82 %	97 %	24 %
Radiologist 4	80 %	27 %	90 %	14 %	Radiologist 4	65 %	73 %	95 %	21 %
Radiologist 5	87 %	55 %	94 %	33 %	Radiologist 5	71 %	91 %	98 %	28 %
Radiologist 6	81 %	64 %	95 %	29 %	Radiologist 6	67 %	91 %	98 %	26 %
Radiologist 7	96 %	27 %	91 %	43 %	Radiologist 7	74 %	64 %	94 %	23 %
Radiologist 8	89 %	45 %	93 %	33 %	Radiologist 8	80 %	64 %	95 %	28 %
Radiologist 9	62 %	82 %	96 %	21 %	Radiologist 9	54 %	100 %	100 %	21 %
Mean	79 %	60 %	94 %	28 %	Mean	67 %	84 %	97 %	24 %
Clinical reference	99 %	18 %	91 %	67 %	Clinical reference	75 %	45 %	92 %	19 %

<b>C</b> Clinically significant prostate cancer (GG2-5), PIRADS 3-5 index lesion					<b>D</b> Clinically significant prostate cancer (GG2-5), PIRADS 4-5 index lesion				
	Sensitivity	Specificity	PPV	NPV		Sensitivity	Specificity	PPV	NPV
Radiologist 1	78 %	67 %	88 %	48 %	Radiologist 1	66 %	83 %	93 %	43 %
Radiologist 2	79 %	63 %	87 %	48 %	Radiologist 2	71 %	92 %	96 %	50 %
Radiologist 3	75 %	63 %	86 %	44 %	Radiologist 3	72 %	67 %	87 %	43 %
Radiologist 4	83 %	33 %	80 %	38 %	Radiologist 4	71 %	71 %	89 %	44 %
Radiologist 5	88 %	38 %	82 %	50 %	Radiologist 5	75 %	71 %	89 %	47 %
Radiologist 6	82 %	42 %	82 %	42 %	Radiologist 6	74 %	79 %	92 %	49 %
Radiologist 7	97 %	21 %	80 %	71 %	Radiologist 7	82 %	67 %	89 %	53 %
Radiologist 8	88 %	25 %	79 %	40 %	Radiologist 8	82 %	46 %	83 %	44 %
Radiologist 9	67 %	75 %	89 %	42 %	Radiologist 9	59 %	88 %	94 %	40 %
Mean	82 %	47 %	84 %	47 %	Mean	72 %	74 %	90 %	46 %
Clinical reference	100 %	12 %	76 %	100 %	Clinical reference	80 %	46 %	81 %	44 %

and negative LR was 2.0 (SD ±1.1, range 1.1–4.0) and 0.4 (SD ±0.2, range 0.2–0.7), respectively.

For more suspicious lesions (PI-RADS 4–5 index lesion, any PCa) (Table 2b), corresponding positive and negative LR median was 4.2 (SD ±3.0, range 0–7.9) and 0.4 (SD ±0.1, range 0.3–0.5), respectively.

For clinically significant PCa (GGG2–5) and PI-RADS 3–5 index lesion (Table 2c), corresponding median and mean positive and negative LR was 1.5 (SD ±0.6, range 1.2–2.7) and 0.4 (SD ±0.1, range 0.1–0.5), respectively.

Positive and negative LR was 2.8 (SD ±2.1, range 1.5–8.9) and 0.4 (SD ±0.1, range 0.3–0.5), respectively, for PI-RADS 4–5 lesions in finding csPCa (Table 2d).

Seven of the nine radiologists were very experienced with prostate MRI assessment and therefore no comparative statistical analysis on experience level was done.

**Extrapolation to the screening cohort**

In the ProScreen trial approximately 16,700 men are randomized to the screening arm. A power calculation determined that roughly 11,690 men would be expected to participate in screening (70%) and of these, 1520 men would have PSA ≥ 3.0 µg/l and subsequently 1000 men would have a 4KScore of ≥7.5% and would therefore have the indication for prostate MRI [9].

Assuming a 30% prevalence of csPCa in the screen-positive subcohort with 1000 men, the mean sensitivity (82%) and specificity (47%) of radiologists would entail 639 men being referred to biopsy of which 244 men would be true positive and 395 false positive. Of the 361 men who would not be referred to biopsy, 305 would be true

negatives and 56 would harbor a cancer that would be missed, i.e., false negative.

If the sensitivity and specificity of the evaluations done by the most sensitive radiologist is assumed in a similar subcohort of 1000 men, 846 men would be biopsied (292 true positive; 554 false positive) and 154 men would not be biopsied (146 true negative; 8 false negative).

Conversely, if the sensitivity and specificity of the evaluations carried out by the least sensitive radiologist is extrapolated into the same subcohort of 1000 men, 376 men would be biopsied (of which 201 would be true positive; 175 false positive) and 624 men would not be biopsied (525 true negative; 99 false negative).

**Discussion**

We estimated the impact of interreader variability on our ongoing ProScreen PCa screening trial which, in addition to the two objective biomarker measurements (PSA and 4 K), is ultimately dependent on subjective evaluation of the MRI by the radiologist. Even though our results correspond reasonably well to the published data [4, 5, 8], the range between the radiologists is broad. We observed a significant difference between radiologists in sensitivity and specificity which can have a substantial impact on the precision of the screening. Assuming a 30% prevalence for csPCa in screen positive men, twice as many men would be biopsied based on MRI interpretation by the most sensitive radiologist compared to the least sensitive radiologist and almost three times more men would undergo unnecessary biopsies (i.e. due to false positive screening results). Conversely, the most sensitive radiologist would miss only 2.6% of csPCa (false negatives), whereas the least

sensitive radiologist would miss every third case. On average 64% of men would be biopsied and 62% of them would undergo unnecessary biopsy but every fifth man with csPCa would be missed.

We could not evaluate the explanatory factors for the observed variation between radiologists, although the radiologists' experience of prostate MRI readings were collected. However, the majority (seven out of nine radiologists) were very experienced, which prevented us from comparing the impact of experience (Table 1). Interestingly, a recent study that evaluated MRI-related interreader variability reported that the sensitivity for detection of index lesion was not dependent on radiologist experience, whereas the specificity was highly dependent on reader experience [8]. Therefore, other causal factors of interreader variability might also exist. It can be assumed that the extremes (clearly malignant and clearly benign) would be reported more consistently although the area in-between these extremes would be more prone to interreader variability. This is at least in part supported by our data as most of the radiologists correctly identified GGG 4–5 cancers whereas there was substantial variability for men with cisPCa as none of the benign prostates were correctly identified by all the radiologists (Fig. 2). Such high variability leads to unnecessary biopsies and, thus, causes unnecessary morbidity and elevated costs.

We found fair interobserver agreement for the detection of index lesion (0.40) and moderate agreement (0.60) for the detection of PI-RADS 4–5 index lesion. While similar relatively modest agreement has been observed previously by Baldisserotto et al. (interobserver agreement of 0.53) [14] we were expecting better agreement [15, 16]. Greer et al. evaluated the interobserver agreement for five radiologist and found a high interobserver agreement [15]. Girometti et al. found substantial agreement in assessing PCa with category three or greater [16]. Greer et al. [8] reported excellent agreement (0.87) for detecting index lesion and substantial agreement (0.74) for true-positive findings. In that same study, nine radiologists evaluated on average 58 MRIs from a group of 163 patients of whom 110 (67%) had a subsequent radical prostatectomy as a reference standard. This might explain their better observed agreement as men selected for RP are more likely to harbor large csPCa. Furthermore, Greer and colleagues noted that not all radiologists interpreted all the images. Similar to the study by Greer et al., biopsy information was not available by the radiologists in our present study. In addition, the interpretation in our study was based on bpMRI as DCE was not available for the radiologists. It has been suggested that readers have a high level of agreement on DCE-MRI assessment in general [17] although agreement regarding the peripheral zone lesions on DCE images may be compromised [18]. It is anticipated that PI-RADS v2.1 will decrease the interreader variation in DCE MRI analyses and reduce

overinterpretations compared to PI-RADS v2 [19]. It has also been reported that DCE may assist in the detection of csPCa in both the peripheral zone (PZ) and the transitional zone (TZ) [17, 20]. Therefore, we expect that the agreement in the ProScreen trial would be better as radiologists that interpret the MRIs of screen positive men in practice have the opportunity to consult colleagues on difficult cases and may have some benefit to evaluating DCE in men with equivocal peripheral zone lesions according to PI-RADS. On the other hand, concern regarding gadolinium deposition in the brain and the extra scanner time (i.e. costs) needed for DCE may promote the use of bpMRI [21]. Also, unknown factors could influence mpMRI interpretation, other than reader experience, and that this needs to be investigated in further studies. Eventually, further standardization in the parameters used for imaging may improve consistency of reporting.

Some evident discrepancies with radiological and pathological assessment were seen in our study (see Fig. 1). A man with a large GG3 pT3a cancer in prostatectomy specimen was correctly identified in MRI by only one of nine radiologists. Conversely, a man with only benign inflammatory histology in biopsies was erroneously scored as cancer by all but one radiologist. These cases aptly demonstrate the inability of prostate MRI to detect some 7% of the csPCas correctly, whereas it is well known that inflammatory changes may confound a reading by appearing suspicious in prostate MRI and is a common cause for false positives [4, 22, 23].

Currently, substantial uncertainty remains about the appropriate actions to be taken on men with clinical suspicion for PCa but negative MRI (nMRI). The true false negative rate, i.e., the rate at which csPCas are missed by MRI, is difficult to assess. Clinically the question is whether systematic biopsies should accompany targeted biopsies. For proper analysis of false negative rate, prostates of all men with PCa suspicion would be removed for pathological evaluation irrespective of MRI/biopsy results, which of course is unethical. The PROMIS trial tackled this dilemma by taking intense 5 mm template mapping biopsies on all men. They showed a false negative rate of 12% for >GGG1 cancers and 7% for >GGG2 cancers [4]. Another way to look at this would be to rely on csPCa incidence during follow-up after negative MRI. Panebianco V et al. showed that csPCa diagnosis free survival (DFS) was 95% after 2 years follow-up [24]. Furthermore, a recent analysis from another cohort largely corroborate this finding by reporting a csPCa DFS of 99.6% after 3 years [25]. In terms of the ProScreen trial this is reassuring as screen positive men (PSA > 3 and 4 K > 7.5%) with a nMRI or negative targeted biopsy are rescreened after 2 years. Furthermore, men with nMRI with PSAD (PSA density) > 0.15 will undergo systematic biopsies as supported by the recent review and meta-analysis [26].

The use of pre-biopsy MRI as a triage test criterion for restricting biopsy to only men with suspicious lesions, could result in one in four men avoiding biopsy. This is in line with the data from PROMIS and PRECISION trials where nMRI rates of 27 and 28% were reported [4, 27]. A recent Cochrane review reported up to one-third of men with nMRI [5] whereas up to one in two has been reported in some expert centers [28].

We found a moderate sensitivity for the detection of any PCa (79%) and approximately the same sensitivity for csPCa (ISUP GG 2–5, 81%). Even though the MRI was quite accurate in detecting csPCa, the sensitivity for more suspicious lesions (PI-RADS 4–5) did not improve. Other studies have obtained better sensitivities (Cochrane review 91%, PROMIS 93%) [4, 5]. This is possibly due to differences in reference standards. The Cochrane review was based on template-guided biopsy and the PROMIS trial was based on 5 mm template mapping biopsy as opposed to the systematic biopsy for most men used in our study. The specificity of the MRI for csPCa in our study was in concordance with the Cochrane review [5].

A low NPV for cisPCa and a high PPV were found in contrast to other studies [4, 5, 29]. The NPV and PPV are highly dependent on the underlying prevalence of the disease and the observed discrepancy, which probably reflects the high prevalence of csPCa (76%) in our study cohort sample.

The threshold used to define positive MRI is equivocal [6]. The intermediate PI-RADS 3 lesions are particularly difficult to define [5]. If the threshold in our study was set at PI-RADS 4 and 5 lesions instead of PI-RADS 3, the proportion of men with nMRI would have increased from 25 to 39%, and the MRI would not have correctly identified 33% of csPCa. This is in concordance with the literature [4, 8]. In respect to the ProScreen trial, it might be an acceptable compromise to increase further the ratio between the benefit and the harm due to built-in “safety tailgate”, whereby men with nMRI would undergo systematic biopsy if PSAD > 0.15, and otherwise (PSAD < 0.15) would be invited for the next screening round in 2 years [9].

+Some inherent limitations to our study must be considered. The PCa prevalence in our study cohort (87%) is higher than in the general population (30%) [30, 31]. Furthermore, the prevalence of csPCa was also high (76%). The prevalence of csPCa in the Cochrane meta-analysis, the MRI-FIRST trial and the 4M trial were 28, 38 and 30%, respectively [5, 28, 29] hence these discrepancies largely restrict any direct comparison of the results. However, our study was not designed to assess the diagnostic performance of MRI, thus the related limitations such as high prevalence of the disease and verification bias are not essential. Instead, the aim of our study was to evaluate the interreader agreement between radiologists and extrapolate these to the ProScreen cohort. The agreement

between radiologists was better for the very high-risk GGG4 and GGG5 cancers as opposed to men with benign histology, which is reassuring in regard to mortality reduction in ProScreen. Nevertheless, the consequence is that more benign prostates would be scored suspicious and thus, the cost-efficiency of screening will be reduced by the taking of unnecessary biopsies.

Though the aim and design of the study were to evaluate interreader agreement, we should also pay attention to the urologist’s role in the diagnostic work up. Accuracy of the fusion biopsy to detect the lesion correctly identified by the radiologist could not be evaluated here.

Re-reading the MRI images did not entirely mimic the routine clinical scenario for several reasons. Although they are still of controversial importance, the DCE sequences were not available for the radiologists [11, 32]. Moreover, contrary to clinical routine, radiologists were not allowed to consult a colleague with challenging cases. These likely underestimate the interreader agreement observed. Finally, nearly all radiologists were relatively experienced, and therefore we had no opportunity to study the effect of experience on the agreement dimension. All these factors may limit the extent to which the results can be generalized, although they should not have a significant effect on the ProScreen trial per se.

## Conclusions

The interreader variability among radiologists whom interpret prostate MRI is significant. In respect to the ongoing ProScreen PCa screening trial, the effect on mortality reduction is expected to be modest. However, poor interobserver agreement especially for men with true benign histology may cause undue sampling of the prostate and thus drive inefficacy of screening.

## Abbreviations

MRI: Magnetic resonance imaging; PCa: Prostate cancer; PSA: Prostate-specific antigen; csPCa: Clinically significant PCa; cisPCa: Clinically insignificant PCa; TRUS: Transrectal ultrasound; mpMRI: Multiparametric magnetic resonance imaging; DR: Detection ratio; PI-RADS: The Prostate Imaging Reporting and Data System; 4K: The four kallikrein; FBx: Fusion biopsies; T2WI: T2-weighted imaging; DWI: Diffusion-weighted imaging; ADC: Apparent diffusion coefficient mapping; DCE: Dynamic contrast enhancement; bpMRI: Biparametric MRI; GGG: Gleason Grade Groups; PPV: Positive predictive value; NPV: Negative predictive value; LH: Likelihood; PZ: Peripheral zone; TZ: Transitional zone; DFS: Diagnosis free survival; nMRI: Negative MRI; PSAD: PSA density

## Acknowledgments

Not applicable.

## Authors’ contributions

AR, TPK and AA made substantial contributions to the conception and design of the study. AK, JR, KI, KL, SM, JO, OO, TS, RS reviewed the MRI scans. TPK did the statistical analysis. TPK and RH have extracted data and drafted the manuscript. AR, KT, TLJT, TM, KN contributed to the conception of the work and substantively revised it. All authors agreed to participate and approved the submitted version.

**Funding**

This study was supported by a grant from the Cancer Society of Finland and the Academy of Finland.

**Availability of data and materials**

Yes.

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Yes.

**Competing interests**

The authors declare that they have no competing interests.

**Author details**

<sup>1</sup>Department of Urology, University of Helsinki and Helsinki University Hospital, PL900, 00029 HUS, Helsinki, Finland. <sup>2</sup>Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland. <sup>3</sup>HUS Diagnostic Center, HUS Medical Imaging Center / Radiology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland. <sup>4</sup>Department of Radiology, Tampere University Hospital, Tampere, Finland. <sup>5</sup>Department of Urology, Tampere University Hospital, Tampere, Finland. <sup>6</sup>HUSLAB Laboratory Services, Department of Pathology, HUS Helsinki University Hospital, Helsinki, Finland. <sup>7</sup>Faculty of Social Sciences, Tampere University, Tampere, Finland.

Received: 7 April 2020 Accepted: 24 September 2020

Published online: 09 October 2020

**References**

- Hugosson J, Roobol M, Månsson M, et al. A 16-yr follow-up of the European randomized study of screening for prostate cancer. *Eur Urol*. 2019;76(1):43–51.
- Heidenreich A, Bastian P, Bellmunt J, et al. EAU guidelines on prostate cancer. Part 1: screening, diagnosis, and local treatment with curative intent-update 2013. *Eur Urol*. 2014;65:124.
- Carter H. American urological association (AUA) guideline on prostate cancer detection: process and rationale. *BJU Int*. 2013;112:543.
- Ahmed H, El-Shater Bosaily A, Brown L, et al. Diagnostic accuracy of multiparametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *Lancet*. 2017;389(10071):815–22.
- Drost FH, Osses DF, Nieboer D, Steyerberg EW, Bangma CH, Roobol MJ, et al. 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*. 2020;77:78–94.
- Barentsz J, Richenberg J, Clements R, et al. ESUR prostate MR guidelines 2012. *Eur Radiol*. 2012;22(4):746–57.
- American College of Radiology. MR Prostate Imaging Reporting and Data System version 2.0. 2015.
- Greer M, Shih J, Lay N, et al. Interreader variability of prostate imaging reporting and data system version 2 in detecting and assessing prostate cancer lesions at prostate MRI. *Am J Roentgenol*. 2019;212:1197–205.
- Auvinen A, Rannikko A, Taari K, et al. A randomized trial of early detection of clinically significant prostate cancer (ProScreen): study design and rationale. *Eur J Epidemiol*. 2017;32:521–7.
- Bryant R, Sjoberg D, Vickers A. Predicting high-grade cancer at ten-core prostate biopsy using four kallikrein. *J Natl Cancer Inst*. 2015;107:7.
- Weinreb J, Barentsz J, Choyke P, et al. PI-RADS prostate imaging - reporting and data system: 2015, version 2. *Eur Urol*. 2016;69:16–40.
- Pierorazio PM. Prognostic Gleason grade grouping: data based on the modified Gleason scoring system. *BJU Int*. 2013;111:753–60.
- Donaldson IA, Emberton M, Freeman A, Ahmed HU. The concept of the index lesion. In: Barret E, Durand M. (eds) *Technical aspects of focal therapy in localized prostate cancer*. Springer. 2015. [https://doi.org/10.1007/978-2-8178-0484-2\\_2](https://doi.org/10.1007/978-2-8178-0484-2_2).
- Baldissertotto M, Neto E, Carvalho G, et al. Validation of PI-RADS v.2 for prostate cancer diagnosis with MRI at 3T using an external phased-array coil. *J Magn Reson Imaging*. 2016;44:1354–9.
- Greer M, Brown A, Shih J, et al. Accuracy and agreement of PIRADSv2 for prostate cancer mpMRI: a multireader study. *J Magn Reson Imaging*. 2017;45:579–85.
- Girometti R, Giannarini G, Greco F, et al. Interreader agreement of PI-RADS v. 2 in assessing prostate cancer with multiparametric MRI: A study using whole-mount histology as the standard of reference. *J Magn Reson Imaging*. 2018;49(2):546–55.
- Greer M, Shih J, Lay N, et al. Validation of the dominant sequence paradigm and role of dynamic contrast-enhanced imaging in PI-RADS version 2. *Radiology*. 2017;285:859–69.
- Rosenkrantz A, Ginocchio L, Cornfeld D, et al. Interobserver reproducibility of the PI-RADS version 2 lexicon: A multicenter study of six experienced prostate radiologists. *Radiology*. 2016;280:793–804.
- Turkbey B, Rosenkrantz A, Haider M, 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(3):340–51.
- Rosenkrantz A, Babb J, Taneja S, Ream J. Proposed adjustments to PI-RADS version 2 decision rules: impact on prostate cancer detection. *Radiology*. 2017;283:119–29.
- Kanda T, Nakai Y, Oba H, Toyoda K, et al. Gadolinium deposition in the brain. *Magn Reson Imaging*. 2016;34(10):1346–50.
- Jyoti R, Jina N, Haxhimolla H. In-gantry MRI guided prostate biopsy diagnosis of prostatitis and its relationship with PIRADS V.2 based score. *J Med Imaging Radiat Oncol*. 2017;61:212–5.
- Rourke E, Sunnapwar A, Mais D, et al. Inflammation appears as high prostate imaging-reporting and data system scores on prostate magnetic resonance imaging (MRI) leading to false positive MRI fusion biopsy. *Investig Clin Urol*. 2019;5:388–95.
- Panebianco V, Barchetti G, Giuseppe S, et al. Negative multiparametric magnetic resonance imaging for prostate Cancer: What's next? *Eur Urol*. 2018;41(1):48–54.
- Venderink W, van Luitelaar A, van der Leest M, et al. Multiparametric magnetic resonance imaging and follow-up to avoid prostate biopsy in 4259 men. *BJU Int*. 2019;5(124):775–84.
- Pagniez MA, Kasivisvanathan V, Puech P, Drumez E, Villers A, Olivier J. Predictive factors of missed clinically significant prostate cancers in men with negative magnetic resonance imaging: a systematic review and meta-analysis. *J Urol*. 2020;204:1:24–32.
- Kasivisvanathan V, Rannikko A, Borghi M, et al. MRI-targeted or standard biopsy for prostate-cancer diagnosis. *N Engl J Med*. 2018;378:1767–77.
- 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. *Eur Urol*. 2018;18:30880–7.
- Rouvière O, Puech P, Renard-Penna R, et al. MRI-FIRST investigators. 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*. 2019;20(1):100–9.
- Bell K, del Mar C, Wright G, et al. Prevalence of incidental prostate cancer: a systematic review of autopsy studies. *Int J Cancer*. 2015;137:1749–57.
- Jahn J, Giovannucci E, Stampfer M, et al. The high prevalence of undiagnosed prostate cancer at autopsy. *Int J Cancer*. 2015;137:2795–802.
- Scialpi M, Rondoni V, Aisa M, et al. Is contrast enhancement needed for diagnostic prostate MRI? *Transl Androl Urol*. 2017;6:499–509.

**Publisher's Note**

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