Factors Influencing Variability in the Performance of mp-MRI in Detecting Clinically Significant Prostate Cancer: A Systematic Literature Review

Armando Stabile¹, Francesco Giganti²,³, Veeru Kasivisvanathan²,⁴, Gianluca Giannarini⁵, Caroline M Moore²,⁴, Anwar R Padhani⁶, Valeria Panebianco⁷, Andrew B Rosenkrantz⁸, Georg Salomon⁹, Baris Turkbey¹⁰, Geert Villeirs¹¹, Jelle O Barentsz¹²

¹ Department of Urology and Division of Experimental Oncology, URI, Urological Research Institute, Vita-Salute San Raffaele University, IRCCS San Raffaele Scientific Institute, Milan, Italy
² Division of Surgery and Interventional Science, University College London, London, UK
³ Department of Radiology, University College London Hospitals NHS Foundation Trust, London, UK
⁴ Department of Urology, University College London Hospitals NHS Foundation Trust, London, UK.
⁵ Urology Unit, Academic Medical Centre, Santa Maria della Misericordia Hospital, Udine, Italy.
⁶ Paul Strickland Scanner Centre, Mount Vernon Cancer Centre, Northwood, UK.
⁷ Department of Radiology, Sapienza Rome University, Policlinico Umberto I, Rome, Italy.
⁸ Department of Radiology, NYU Langone Health, New York, NY, USA.
⁹ Prostate Cancer Center, Martini-Klinik Hamburg, University Hospital Hamburg-Eppendorf, Hamburg, Germany.
¹⁰ Molecular Imaging Program, National Cancer Institute, NIH, Bethesda, MD, USA
¹¹ Department of Radiology, Ghent University Hospital, Ghent, Belgium.
¹² Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, The Netherlands.
**Corresponding author:**
Armando Stabile, MD
Department of Urology and Division of Experimental Oncology
Urological Research Institute, IRCCS San Raffaele Scientific Institute,
Via Olgettina 60, 20132 Milan, Italy
Tel. +39 02 26435663; Fax. +39 02 26437298
E-mail: armando.stabile88@gmail.com

**Word count (text):** 6758  
**Text pages:** xx  
**Tables:** 4  
**Word count (abstract):** 348  
**References:** 110  
**Figures:** 2

**Keywords:** prostate cancer; diagnosis, multiparametric magnetic resonance imaging; MRI; targeted biopsy

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The authors declare no conflicts of interest in preparing this article. Francesco Giganti is funded by the UCL Graduate Research Scholarship and the Brahm PhD scholarship in memory of Chris Adams. Veeru Kasivisvanathan is an Academic Clinical Lecturer funded by the United Kingdom National Institute for Health Research (NIHR). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.
Abstract

Context: There is a lack of comprehensive data regarding the factors that influence the diagnostic accuracy of multiparametric magnetic resonance imaging (mpMRI) to detect and localize clinically significant prostate cancer (csPCa).

Objective: To systematically review the current literature assessing the factors influencing the variability of mpMRI performance in csPCa diagnosis.

Evidence acquisition: A computerized bibliographic search of Medline/PubMed database was performed for all studies assessing magnetic field strength, the use of an endorectal coil (ERC), the assessment system used by radiologists and inter-reader variability, radiologist and urologist experience, the use of a contrast agent and the use of computer-aided diagnosis (CAD) tools in relation to mpMRI diagnostic accuracy.

Evidence synthesis: A total of 77 articles were included. Both radiologist’s reading experience and urologist’s/radiologist’s biopsy experience were the main factors that influenced diagnostic accuracy. Therefore, it is mandatory to indicate the experience of the interpreting radiologists and biopsy-performing urologists to support the reliability of the findings. The most recent PI-RADS guidelines are recommended to use, as the main assessment system for csPCa given the simplified and standardized approach as well as its particular added value for less experienced radiologists. Biparametric MRI had similar accuracy to mpMRI, however, bpMRI performed better with experienced readers. The limited data available suggests that the combination of CAD and radiologist readings may positively influence diagnostic accuracy.

Conclusions: Multiple factors affect the accuracy of mpMRI and MRI-targeted biopsy to detect and localize csPCa. The high heterogeneity across the studies underlines the need
to define the experience of radiologist and urologist, to implement quality control, and to adhere to the most recent PI-RADS assessment guidelines. Further research needs to clarify how and which factors impact the accuracy of the MRI-pathway.

**Patient summary:** We systematically reported the factors influencing mpMRI accuracy in detecting csPCa. These factors are significantly related to each other with the experience of the radiologists being the dominating factor. In order to deliver the benefits of mpMRI to diagnose csPCa, it is necessary to develop expertise for both radiologists and urologists, to implement quality control, and to adhere to the most recent PI-RADS assessment guidelines.
1. **Introduction**

Over the last 10 years the diagnostic pathway of prostate cancer (PCa) is significantly changed by the advent of multiparametric magnetic resonance imaging (mpMRI) [1,2]. As shown by recently published randomized controlled trials, head-to-head comparisons and a Cochrane meta-analyses [3–8] mpMRI is the best technique to detect and localize suspicious areas for clinically significant prostate cancer (csPCa), and it allows to perform MRI targeted biopsy (MRI-TBx) [9]. The value of mpMRI and MRI-TBx over systematic transrectal ultrasound (TRUS)-biopsy is in reducing diagnoses of insignificant (is)PCa and potentially avoiding unnecessary prostate biopsies in men with negative mpMRI scans. The detection and localization of csPCa are slightly but not significantly better compared to TRUS-biopsy.

The acquisition and interpretation of prostate mpMRI are evolving with ongoing improvements, that influence its accuracy. These include magnetic field- and gradient-strength, the use of an endorectal coil (ERC)[10], different versions of assessment-systems, reader experience and inter-reader variability, the potential to avoid a contrast agent (i.e. biparametric MRI - bpMRI), and the use of computer-aided diagnosis (CAD) and deep-learning tools. Another factor that influences the accuracy of the MRI-pathway is the experience of operators performing MRI-TBx. These show a significant variation across reported series [4], affecting the risk of bias of the available data and preventing robust systematic analyses.

Despite the large number of studies reporting the accuracy of mpMRI, there is a lack of comprehensive data, that specifically address the difference of mpMRI execution and performance. Given the promising role of MRI in csPCa diagnosis, there is a need to
systematically review the current literature on the factors that influence the variability of mpMRI in the diagnosis of csPCa.

2. Evidence acquisition

2.1. Objective

We aimed to systematically review the current literature assessing the factors influencing the variability of mpMRI performance in detecting csPCa. The magnetic field strength (1.5T vs 3.0T, including importance of the gradient-strength), the use of an endorectal coil, the assessment system used by the radiologist, the inter-reader variability, the radiologist and urologist experience, the use of bpMRI and the use of CAD or Deep or Machine Learning (DL, ML) for mpMRI assessment were considered as potential influencing factors.

2.2. Search strategy

Data collection was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) statement [11]. A computerized bibliographic search of Medline/PubMed database was searched from inception to 23rd June 2019. The search terms used were (prostate cancer OR prostate adenocarcinoma) AND (MRI OR magnetic resonance) AND (coil OR endorectal coil OR surface coil OR magnetic field OR reporting system OR PI-RADS OR Likert OR inter-reader variability OR inter-reader agreement OR biparametric OR radiologist experience OR urologist experience OR learning curve OR CAD OR machine learning OR computer-aided OR artificial intelligence OR neural network).
2.3. Inclusion criteria

As recommended by the PRISMA guidelines, we used the Population, Intervention, Comparator and Outcome (PICO) approach to define study eligibility [11]. Reports were considered relevant if they provided comparative data on the relationship between the aforementioned factors. MpMRI performance was defined as PCa detection at either prostate biopsy or after radical prostatectomy. Thus, studies assessing one of the factors without providing any comparison (e.g. providing data regarding the 1.5T mpMRI or bpMRI alone accuracy) were not included in this review. Non-comparative studies, case reports, editorials, letters, review articles, and meeting abstracts were not included.

2.4. Systematic review process

Two authors (AS and FG) independently reviewed a total of 2013 abstracts and selected 77 studies that were finally included in the systematic review for full-text evaluation. Figure 1 shows the PRISMA flowchart describing the selection process.

2.5. Data extraction

Data were independently extracted from all included studies by the same two authors. A standardized data extraction form was created a priori and used to collect data such as the study design, number of participants, mpMRI protocol, radiologist experience, and outcome.

2.6. Data analysis
A comprehensive and narrative synthesis of included studies was performed, since quantitative meta-analytic synthesis was not possible due to the heterogeneity of the studies.

2.7. Risk of bias assessment

The risk of bias and applicability concern in individual studies was assessed independently by the same two authors using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) criteria [12]. The presence of baseline confounding factors or selection bias, as well as the presence of any bias within mpMRI protocols, mpMRI interpretation, biopsy protocol, and histopathological reference standard, were assessed (Figure 2; Supplementary Figure 1).

3. Evidence synthesis

Overall, 2013 publications were found. If it was not clear from the abstract whether the paper might contain relevant data, the full paper was assessed. Seventy-seven articles were included in the final analysis (Fig. 1). Single studies are described in detail in Tables 1-4 and Supplementary tables 1-3.

3.1. Risk of bias within studies

The overall risk of bias and applicability concern is given in Figure 2. The overall methodological quality of the studies was moderate with 17 studies having a low risk of bias and applicability concern across all domains assessed. Supplementary Fig. 1 shows the risk of bias and applicability concerns for each study.
3.2. Magnetic field

MpMRI at 3.0T has an increased signal-to-noise ratio compared to 1.5T scanners, resulting in a higher spatial resolution of T2-weighted (T2-WI) and diffusion-weighted imaging (DWI).

In total, seven studies assessed the value of the magnetic field relating to mpMRI staging accuracy (Supplementary Table 1) [13–19]. Two studies from 2004 [14,17] were among the first reporting a comparison of 1.5T and 3.0T scanners even though with slightly different results. Bloch et al.[14] reported the higher image quality at 3.0T compared to 1.5T (both with ERC) while Sosna et al.[17] reported a comparable quality for non-endorectal 3.0T and endorectal 1.5T mpMRI in forty men receiving mpMRI and subsequent biopsy. However, these studies are not comparable given that ERC was not used at 3.0T [17]. Most importantly only two studies addressed diffusion-weighted imaging (DWI) [18,19] and thus only these studies could evaluate the detection difference between insignificant (is)PCa and csPCa, as for this DWI is the most important sequence in the peripheral zone (PZ). These studies showed a similar PI-RADS assessment for 1.5 and 3T.

Overall, the majority of the studies did not investigate the detection of csPCa but instead focused on the recognition of any PCa. In this respect both magnetic field strengths performed equally [13,16,18] but the small sample size and high heterogeneity of these studies make an objective comparison difficult to conduct. In summary, no reliable information could be obtained regarding the detection of csPCa according to field strength regardless of reception coils usage.
3.3. Reception Coil

A total of eleven studies comparing the use of mpMRI with an external pelvic phased-array coil with or without ERC were included (Supplementary Table 2)[20–30]. Four studies did not use DWI.

The use of an ERC significantly improves the signal-to-noise ratio irrespective of the magnetic field strength, providing T2-WI with higher spatial resolution and potentially a more accurate delineation of the structures in the transition zone (TZ), which is the key factor in assessment of csPCa in this zone. However, the addition of an ERC is associated with increased costs, increased artifacts [31], organ deformation and discomfort for patients. Mirak et al.[27] investigated the performance of 3.0T mpMRI with and without ERC to detect PCa using PI-RADS v.2 guidelines, with whole-mount histopathology as reference standard. Two sub-cohorts, with (n=260) and without (n=169) ERC, were analyzed. They concluded that detection rates for any PCa, for the index lesion and for csPCa were similar in both cohorts, but there was a significantly lower detection of anterior and TZ csPCa in the ERC sub-cohort due to signal drop off in the anterior gland when ERC was used without an accompanying abdominal phased array coil.

Another study with a direct comparison of ERC vs non-ERC at 3.0T [21] showed no differences in detecting PCa using PI-RADS v2 guidelines.

Costa et al.[25] reported that using an ERC at 3.0T provides superior sensitivity (78%) for PCa detection when compared with standard and augmented protocols (i.e. those with twice as many signal averages; 43% and 60%, respectively) without ERC.
Turkbey et al. [22] compared the utility of T2-WI and DWI at 3.0T with and without ERC in detecting PCa in twenty men receiving mpMRI before radical prostatectomy. The authors demonstrated a higher sensitivity when using an ERC. The sensitivity and positive predictive value (PPV) of ERC use vs non-ERC MRI were 76% vs 45% and 80% vs 64%, respectively.

Torricelli et al. [29] reported that the image quality with ERC at 1.5T is superior to 3.0T without ERC in evaluating tumor conspicuity, capsular infiltration, and seminal vesicle involvement. No significant differences between the two techniques for apex and neurovascular bundles involvement and comparable performance for the diagnosis of capsular involvement were observed. The authors concluded that during preoperative PCa staging, 3.0T mpMRI can provide comparable diagnostic information to 1.5T mpMRI with ERC.

In summary, the use of ERC improves signal reception, that slightly improved sensitivity to visualize any PCa. Only one study addressed the value of the ERC with respect to the PI-RADS assessment of csPCa [27]. This study showed a minor disadvantage of the ERC to detect anterior TZ PCa. Due to the lack of significant evidence that the ERC improves csPCa assessment, associated increased costs, duration of examination and patient discomfort, the PI-RADS v. 2.1 guidelines [32] recommend the ERC to be only used only for older 1.5T scanners with lower gradient-strength [29].

Still, the lack of standardized protocols in the available studies makes robust comparisons hard to assess.

3.4. Assessment system
The development of a simple, structured, standardized assessment system was one of the most considerable challenges in prostate mpMRI. Assessment systems have been refined during the years, in order to increase the inter-reader agreement, to decrease the gap between differently skilled radiologists and to improve communication between radiologists and urologists. Ten studies comparing the use of different assessment systems were included (Table 1)[33–43]. The majority compared PI-RADS v1[44] and v2[45]. Differently from PI-RADS v1, PI-RADS v2 defined dominant sequences (T2-WI for the TZ and DWI for the PZ) and decreased the role of dynamic contrast-enhanced (DCE) MRI-imaging[46].

Most studies reported a similar diagnostic accuracy for both PI-RADS v1 and v2. However, three studies[35,36,38] showed a higher sensitivity of PI-RADS v2 for TZ lesions and another study by Krishna et al.[39] showed that PI-RADS v1 detected approximately 10% more tumors than PI-RADS v2.

De Visschere et al.[34] compared the performance of PI-RADS v2 and v1 in 245 biopsy naïve men with an elevated prostate-specific antigen (PSA). They found that PI-RADS v1 and PI-RADS v2 yielded similar accuracy to detect csPCa. However, PI-RADS v2 had lower sensitivity than PI-RADS v1 when a score of 3 was used as a threshold for a positive mpMRI. The authors suggested that the majority of discrepancies were caused by a suspicious lesion in the PZ on T2-WI but with normal DWI, scored positive with PI-RADS v1 but negative with PI-RADS v2. Indeed, sensitivity of PI-RADS v2 might be less when suspicious lesions in T2-WI but negative on DWI are present in the PZ. Nonetheless the authors concluded that PI-RADS v2 is preferable because of the more structured and standardized and simpler approach. Similar results were found by Krishna.
et al.[39] who compared the two assessment systems for the detection of csPCa lesions in 47 men before radical prostatectomy. They found no difference in the overall detection of csPCa but observed a higher sensitivity of PI-RADS v1 on T2-WI and DCE.

Feng et al. [36] compared PI-RADS v.1 and v.2 in 401 consecutive biopsy naïve men with clinical suspicion of PCa at 3.0T mpMRI. Both assessment systems had a good diagnostic performance for the detection of csPCa, but the diagnostic accuracy increased from 0.82 to 0.88 with the use of PI-RADS v2 compared to v1 when also non-biopsy naïve men were included. Interestingly, PI-RADS v2 had a better performance in the TZ (0.92 vs 0.81).

Hoffmann et al.[37] evaluated in 58 men whether PI-RADS v2 was more accurate in assessing anterior prostate csPCa compared to PI-RADS v1. PI-RADS v2 did not improve the accuracy for diagnosing anterior csPCa when compared to PI-RADS v1, however, PI-RADS v.2 was more reproducible between radiologists.

Schaudinn et al.[42] reported moderate interobserver reliability (k=0.48) for PCa detection of two radiologists in 40 men before radical prostatectomy, and similar results have been reported by Tewes et al.[43], with a moderate interobserver agreement (k=0.56) for PI-RADS v2 and fair agreement (k=0.39) for PI-RADS v1. The authors concluded that PI-RADS v2 had equivalent diagnostic accuracy to PI-RADS v1 for PCa detection, but with a shorter interpretation time for PI-RADS v2.

Two studies compared PI-RADS v1 and two different Likert assessment systems[40,41]. Both showed good inter-reader agreement, even if one study reported a higher accuracy of the Likert scale for TZ lesions compared to PI-RADS v1[41].
In summary, it is recommended to use the most recent PI-RADS guidelines as the main assessment system given the simplified, objective, and standardized approach as well as its particular added value for less experienced radiologists.

3.5. **Radiologist and urologist experience**

The use of mpMRI to detect csPCa represents a multidisciplinary approach, that includes both the skills in acquisition and interpretation of mpMRI and in performing MRI-TBx.

Five studies assessed the impact of reader experience on mpMRI diagnostic performance (Table 2). Akin et al. and Garcia-Reyes et al. were the first to test the learning curve in prostate mpMRI reporting using whole-mount pathology as a reference standard. They showed that a dedicated training curriculum is useful to improve mpMRI interpretation [47,48]. Rosenkrantz et al. [49] assessed the variation of mpMRI diagnostic accuracy in detecting and localizing csPCa of six 2nd-year radiology residents reporting 124 prostate mpMRI scans (both negatives and positives). Three out of six readers (50%) received feedback after each examination showing the preceding case’s solution. For both readers with and without feedback, there was an initial rapid improvement, that slowed down after 40 examinations. In the group receiving feedback, accuracy, sensitivity and specificity improved from 58%, 59% and 56% to 72-77%, 72-77% and 74-82%, respectively. Interestingly, the presence of feedback did not significantly affect the accuracy as compared to the group without feedback, showing the effects of self-directed learning, even though readers with feedback felt more confident. Moreover, the feedback was more useful for TZ lesions, suggesting a higher challenge in detecting these tumors,
in line with previous studies [47,48]. Pickersgill et al. [50] conducted a retrospective review of 459 men receiving mpMRI according to PI-RADS (v1 implemented with v2 during the study period) and a subsequent MRI-TBx if necessary. They showed, that the radiologist’s experience did not improve the accuracy in csPCa detection. The authors speculated that the use of PI-RADS might have attenuated the impact of the reader’s experience. However, this study had serious limitations, such as the implementation of PI-RADS version 1 to version 2 during the study period and an arbitrary definition of radiologist experience (i.e. more than 500 mpMRI examinations). Following the widespread use of mpMRI and the need for dedicated training for radiologists, an online interactive case-based website for prostate mpMRI interpretation using PI-RADS version 2 has been proposed [51]. This training course increased the sensitivity (58% to 73%, p=0.003) and the negative predictive value (NPV) (69% to 78%, p=0.049) of three 2nd-year radiology residents who evaluated separate sets of sixty MR scans before and after the course. Interestingly, there were no significant improvements in the accuracy of the PI-RADS assessment scores (from 46% to 53%, p=0.149) [51].

The quality of MRI-TBx performance plays an equally important role in defining the final diagnostic accuracy of this technique. Similar to systematic ultrasound-guided biopsy [52,53], the experience of the biopsy-operator significantly influences the outcome of an MRI-TBx, that can be visual/cognitive (cogn), MR-US-fusion (fus), or direct-MRI-TBx (in-bore) [54], using either a transrectal or transperineal approach.

A total of eight studies assessing the learning curve of MRI-TBx were included (Table 2). Gaziev et al. [55] demonstrated a significant gradual increase in csPCa detection (27% to 63%) in 70 men receiving fusion MRI-TBx. Similarly, Calio et al. [56]
reported data from three consecutive cohorts of biopsy-naïve men receiving MRI-TBx over a study period of nine-years. There was a 13% increase in csPCa detection by MRI-TBx from the early to the most recent cohort. Meng et al.[57] reported a 26% increase of csPCa detection in 1500 men receiving repeat biopsy. Mager et al.[58] attempted in demonstrating the presence of the learning curve effect proposing the MRI-TBx quotient, defined as the ratio between the number of positive targeted cores and the total number of targeted cores. The authors showed a significant learning process, in both detection-quotient and biopsy times; for a novice performer a sufficient learning process occurred after 42 procedures reaching a flattening after 63 biopsies. Kasabwala et al.[59] calculated the distance between the planned and the actual core route in the prostatic tissue during a fusion MRI-TBx, and demonstrated a significant improvement in MRI-TBx accuracy after 98 cases. Halstuch et al.[60] attempted to identify a minimum number of procedures to reach the best PCa detection using a mathematical algorithm. The authors demonstrated that at least 104 transrectal fusion MRI-TBx and 119 transperineal fusion MRI-TBx are necessary for men with visible PI-RADS 3 lesions, before reaching the plateau phase of csPCa detection. In this context, Stabile et al.[61] demonstrated the presence of a learning curve affecting csPCa detection rate even when accounting for several confounders (such as PSA, prostate volume and PI-RADS score) for both visual and fusion MRI-TBx. The authors showed a steep increase in csPCa detection during the first 60 procedures and a flattening after 80 procedures. Interestingly, it was suggested that the transperineal approach might be less affected by the learning curve effect, hence it might be easier compared to the transrectal approach when considering MRI-TBx [61]. Finally, Westhoff et al.[62] proposed at least eight procedures as experience threshold.
However, this result should be cautiously interpreted considering the limited number of MRI-TBx performed by each of the twenty-two urologists included in this study.

In summary, when assessing the performance of mpMRI in detecting csPCa it is nowadays mandatory to indicate the experience of the interpreting radiologists and biopsy performing urologists to support the reliability of the findings. Less experienced readers and biopsy-operators should always be supervised by experienced readers and operators. Moreover, mpMRI performance should be internally validated before widespread adoption. According to Rosenkrantz et al. radiologists should have reported at least 100 expert-supervised prostate mpMRI examinations after a dedicated training course [49] and urologists should have performed between 60 and 100 MRI-TBx before they potentially reach an acceptable csPCa detection. Most importantly additional quality assurance tests are needed. Further development of quality-criteria, assessment, and the development of training platforms/courses are needed.

3.6. **Inter-reader variability**

Although reader experience plays a substantial role in determining mpMRI accuracy, the reporting process is affected by an almost inevitable variability among different radiologists which varies across different studies and centers. Fifteen studies were included for this topic (Table 3).

Quentin et al.[63] assessed the inter-reader agreement of mpMRI using a 5-point (Likert) scale [64]. The authors showed high inter-reader reliability (PPV: 88-96%; k=0.90) between three blinded radiologists scoring 108 lesions. After the introduction of PI-RADS v1 guidelines, Schimmöller et al.[65] reported the agreement of three
experienced radiologists scoring 164 pre-marked lesions. The overall agreement was good to moderate and higher for malignant than for benign lesions. Nonetheless, the clinical application of PI-RADS was still premature, and a diagnostic cut-off had not yet been proposed. Rosenkrantz et al.[66] published a series of 55 patients undergoing prostate mpMRI in a single institution that were retrospectively reviewed by three radiologists (two moderately experienced and one inexperienced) using both PI-RADS v1 and Likert scores. For both assessment methods the agreement was strong between the two experienced readers but poor when compared to the inexperienced reader. Interestingly, the Likert assessment scale had better inter-reader reproducibility than PI-RADS score in the TZ. This was probably due to the greater experience of the readers with their ‘own’ Likert assessment. Since the widespread use of PI-RADS assessment score and the introduction of PI-RADS v2, few studies have assessed its reproducibility, with conflicting results. Mussi et al.[67] reported moderate to good agreement between eight radiologists with different levels of experience using PI-RADS v2. However, this study is hardly applicable to clinical practice since each reader evaluated only one MR slice containing a single specified lesion. Similarly, Glazer et al.[68] conducted a retrospective study with three radiologists (with different levels of experience, ranging from 1 to 11 years) who scored pre-selected lesions, with a moderate agreement for PZ (k=0.46) and fair agreement for TZ (k=0.36). Moreover, the authors disclosed as PI-RADS v2 had been recently introduced in their clinical practice, potentially influencing the level of reproducibility. Girometti et al.[69] supported the higher level of agreement among experienced radiologists in a monocentric study including three radiologists analyzing forty-eight MRI scans, with a moderate agreement both for PI-RADS cut-offs
Even though the readers were aware of the pre-operative reason of mpMRI, hence of the presence of csPCa. Müller et al.[70] reported a poor level of agreement in a cohort of 126 men receiving two consecutive MRI scans at two different institutions. Nonetheless, the design of this study was not devoid of many limitations and significant bias. In fact, among 292 patients referred to the first institution, 126 men had mpMRI lesions considered challenging to be reliably accessed by systematic or cognitive biopsy. Since equipment for MRI targeted biopsy with fusion technique was not available in the first institution, these men were referred to a second institution where they received a second mpMRI before undergoing an MRI targeted fusion biopsy. Also, all readers and authors had limited experience and training in reading prostate MRI according to PI-RADS. Moreover, no information regarding the used PI-RADS version was provided. For these reasons, these results should be very cautiously considered. Rosenkrantz et al.[71] carried out a multi-center study with six experienced radiologists reporting at two different time points (40 and 80 MRI scans per session) and receiving a training session in between. The authors reported moderate reproducibility of PI-RADS v2 suggesting no benefit from the training session[71]. However, this study was limited by suboptimal image quality in a number of the included centres. Smith et al.[72] provided results regarding intra- and inter-reader agreement with a multi-centre study on four differently experienced readers. Overall, intra-reader agreement was moderate to substantial (60%-77% of agreement across different radiologists). Inter-reader agreement was poor to moderate and higher for more experienced radiologists. Hansen et al.[73] reported the value of a second opinion by a sub-specialized tertiary care center in reviewing mpMRI from seven different regional
hospitals. Overall disagreement was 54% (86/158 MRIs). Specifically, the second reading had significantly improved NPV (0.89 vs 0.72) and PPV (0.61 vs 0.28). Greer et al. [74] reported an excellent agreement on index lesion identification (k=0.87) and moderate on individual PI-RADS v2 category assignment (k=0.419). Other two well-designed studies [75,76] reported similar results with an area under the curve (AUC) for PCa ranging between 0.88 and 0.95 among six blinded readers [75]. Conversely, two recent studies showed high variability in PI-RADS v2 reporting [77,78]. In particular, Sonn et al. [78], in a retrospective study of real-life mpMRI reporting by taking into account nine radiologists and 409 patients, while reporting a low variation in the number of lesions identified, the authors demonstrated high variability in PI-RADS distribution and csPCa detection. The AUC for csPCa ranged between 0.61 and 0.81 [78]. Finally, van der Leest et al. [79] in their prospective multicenter head-to-head comparison study showed high inter-reader agreement of the participating expert radiologists. This was obtained after similar training compared to Rosenkranz et al. [49].

In summary, most of the well-designed dedicated studies reported moderate agreement when PI-RADS v2 is taken into consideration. Furthermore, the radiologist’s experience is crucial to increase mpMRI reproducibility, with the major concerns related with the variability in csPCa yield and high false-positive rates. The heterogeneity across the studies is still high. Most of the studies on this topic did not provide results about MRI-acquisition, or reader-experience, or training. There is thus still a need for standardized mpMRI-assessment training protocols that should be widely available in order to improve the general performance of mpMRI and to provide more reliable data in this context. Only Rosekrantz et al. [49] and van der Leest et al. [79], describe that
radiologists should have reported at least 100 expert-supervised prostate mpMRI examinations after a dedicated training course. Further standardization of assessment systems, education, and certification will likely help in reducing the subjectivity and improving the reproducibility also among less experienced readers.

3.7. Biparametric-MRI vs multiparametric-MRI

Since the introduction of a standardized reporting system for mpMRI[44], the role of DCE-MRI has been controversial. Indeed, version 2 of PI-RADS downgraded the role of DCE-MRI to an additional sequence only for upgrading a PI-RADS 3 to 4 PZ lesion on DWI[32]. In the light of the increasing use of mpMRI in the assessment of csPCa and the need for more optimized and efficient protocols, the use of a bpMRI based only on T2-WI and DWI has been proposed by multiple authors [80,81]. The benefits of omitting DCE-MRI are related to reduced examination times, reduced costs and avoiding the risk of adverse events related to the use of contrast agent. Results coming from prospective trials assessing the diagnostic accuracy of bpMRI are promising[80,81] regarding biopsy avoidance and for reducing the detection of insignificant cancers. Comparative studies of mpMRI and bpMRI are mostly retrospective with significant differences in the methods and not negligible risk of bias (Table 4). To the best of our knowledge, Stanzione et al.[82] showed the diagnostic accuracy of bpMRI compared to mpMRI. The authors reported a series of 82 men undergoing mpMRI for suspicion of csPCa then receiving systematic biopsy plus eventual MRI-TBx, with 35% of patients receiving radical prostatectomy. Two experienced radiologists blindly reported bpMRI first, followed by mpMRI (i.e. with DCE), after an interval of 20-30 days to avoid any recall bias. The
The overall AUC of bpMRI and mpMRI for csPCa detection was 0.91 and 0.93 (p>0.05). Thstrup et al.[83] reported similar accuracies of bpMRI and mpMRI in detecting csPCa, though without providing any formal statistical comparisons. Lee et al.[84] compared two cohorts undergoing mpMRI and bpMRI for a suspicion of PCa and then receiving visual MRI-TBx in addition to standard systematic biopsy. The authors reported a similar detection of csPCa among men who had suspicious lesions in the bpMRI and mpMRI group (63% and 62%, respectively). Nonetheless, these results should be carefully interpreted since the two cohorts were not randomly matched.

Further studies reported promising results though being affected by significant bias mainly concerning the MRI interpretation process (mpMRI and bpMRI read by the same radiologist during the same session)[85] and the reference standard (no biopsy result in men with negative MRI)[86]. The similar diagnostic performance of these two techniques was confirmed in other series[87–89] that attempted to better identify the eventual differences. Specifically, omitting DCE MRI was related to an increasing rate of PI-RADS 3 lesions and a slightly better specificity and worse sensitivity (though never significant)[87,89]. Furthermore, DCE-MRI was not needed for determination of the overall assessment category in 81% of patients[89].

Choi et al.[90] compared the ability of bpMRI vs mpMRI in detecting the index lesion using radical prostatectomy as reference standard. Two independent radiologists (7 and 13 years of experience) retrospectively reviewed prebiopsy MRI of 113 men using PI-RADS v2. No significant differences in csPCa diagnostic accuracy for bpMRI vs mpMRI for both readers using PI-RADS ≥3 as cut-off were found. Interestingly, both readers reported a significantly higher sensitivity for bpMRI compared to mpMRI[90].
Furthermore, inter-reader agreement on PI-RADS assessment score was moderate for both bpMRI (k=0.540) and mpMRI (k=0.478). However, both readers of this study were aware that all men underwent radical prostatectomy for csPCa and this might have affected the overall results. In a similar study, Scialpi et al.[91] evaluated bpMRI and mpMRI ability in detecting single lesions in a cohort of 41 men receiving radical prostatectomy. For both bpMRI and mpMRI the sensitivity was similar and was 100% in PZ and 98 and 95% in the entire prostate and TZ, respectively. BpMRI detected 181/131 lesions detected at final pathology resulting in 28% false positives and 3% false negatives rates[91]. Nonetheless, no information regarding readers’ background was provided and specific information regarding the experience is often scarce.

In this context, Gatti et al.[92] compared bpMRI and mpMRI according to the reader’s experience. The authors conducted a retrospective study on 6 blinded radiologists, divided into three groups according to the level of experience, reviewing bpMRI and mpMRI protocols of 68 men receiving a prostate biopsy and eventual radical prostatectomy. The authors used a modified version of PI-RADS v2 [89] for bpMRI reading and a cut-off≥4 for both protocols. Interestingly, the specificity was quite stable regardless of the protocol and the readers’ experience. Significant differences were found for sensitivity and AUC in detecting PCa index lesions, mainly related to the rate of true positives. The effect of experience was more evident when considering bpMRI than mpMRI. Consequently, in the highly experienced group, the performance of bpMRI vs mpMRI was similar (AUC:0.86 vs 0.93, p=0.10; true positive: 82% vs 86%, p=0.13). The accuracy of bpMRI became progressively less if compared to mpMRI with the decrease of experience (0.68 vs 0.77 in the less experienced group). Furtherly, the authors
attempted to provide a rough indication on the number of necessary cases to reach an AUC and sensitivity ≥0.80: 150-200 for mpMRI and 700-800 for bpMRI[92].

Differently, Di Campli et al.[93] showed no diagnostic differences between bpMRI and mpMRI and no significant influence by readers’ experience.

In summary, available evidence from comparative studies suggests that bpMRI might be a potentially valid alternative to mpMRI, particularly for experienced readers, on the condition that DWI is of excellent quality. These findings have also been confirmed in a recent meta-analysis, demonstrating the non-inferiority of bpMRI and showing on overall non-significant higher sensitivity and lower specificity of mpMRI [94]. Moreover, a recent prospective, multi-reader, blinded direct comparison between bpMRI and mpMRI showed similar diagnostic performance in ruling out the presence of high-grade PCa [79].

That being said, the high methodological heterogeneity might have represented a great confounder, and it remains unclear how the performance of bpMRI will translate to less experienced centers and less quality images. Indeed, the assessment system used (i.e. PI-RADS, dedicated bpMRI-PI-RADS), the choice of the cut-off, the magnetic field, the choice of the outcome (i.e. PCa, csPCa) and the reference standard are the factors varying the most across the studies. Ultimately, randomized prospective studies using non-inferiority designs, in men with variable prevalence with clinically meaningful endpoints (biopsy avoidance, detection of csPCa and cisPCa) will be needed to decide on which patient groups can avoid contrast enhancement.

3.8.  Computer-aided diagnosis and deep learning
The first study on PCa detection [95] was carried out by Chan et al. in 2003 (Supplementary Table 3)[96]. The CAD mainly consisted of quantitative analysis of medical imaging data (i.e. mpMRI) with the aim to provide results potentially related to clinical conditions (i.e. PCa). Different CAD systems have different strategies for imaging feature analysis with different diagnostic accuracies[97]. The imaging analysis process includes multiple steps that can change across different CAD systems [97,98].

The CAD has recently been assessed in aiding radiologists during mpMRI interpretation and reporting. Interestingly, the evaluation of CAD as an assisting tool began at a similar time to the proposal of the first standardized assessment system (i.e. PI-RADS v1)[44,99]. The inclusion of CAD systems in mpMRI interpretation process would potentially overcome some of the issues affecting diagnostic accuracy such as reader experience, reproducibility, as well as enhancing the accuracy of mpMRI itself.

Hambrock et al.[99] published the first study evaluating the effect of CAD for both less and more experienced radiologists on the differentiation of benign from malignant lesions at mpMRI. The authors demonstrated as the addition of CAD significantly improved the performance of less experienced radiologists detecting PCa, and when less experienced radiologists used CAD, they achieved similar performance to that of experienced radiologists. Furthermore, stand-alone CAD had similar diagnostic accuracy compared to experienced readers. However, this study did not replicate a real-life mpMRI diagnostic pathway, since the CAD system used was able to differentiate only pre-identified regions between benign and malignant but not PCa detection in a whole gland. In a similar study, Niaf et al.[100], demonstrated that CAD increased both experienced and less-experienced readers’ performance (AUC increase of 2% and 4%,
respectively, \( p=0.08 \). Litjens et al.\[101\], in a standard clinical scenario, showed that the combination of CAD and PI-RADS assessment categorization achieved higher diagnostic accuracy than PI-RADS categorization alone to discriminate between both benign vs PCa (0.88 vs 0.81, \( p=0.013 \)) and PCa vs csPCa (0.88 vs 0.78, \( p<0.01 \))\[101\]. In a similar study based on prostatectomy specimens, Wang et al.\[102\] demonstrated that the combination of CAD information with PI-RADS v2 increased the clinical net benefit for PCa identification as compared to PI-RADS categorization alone. Further studies demonstrated the clinical utility of quantitative analyses of ADC in improving the diagnostic performance when compared to the scoring system alone\[103–105\]. Giannini et al.\[106\] was the first to replicate a real-life diagnostic pathway including the use of CAD. On per-patient analysis, the use of CAD achieved a higher sensitivity for csPCa (91% vs 81%, \( p=0.046 \)), while specificity was not affected. The per-lesion analysis showed a higher number of single lesions detected with CAD assistance. Interestingly, the average reading time with CAD was significantly lower (60s vs 220s; \( p<0.001 \))\[106\]. Greer et al.\[107\] tested the effect of CAD in a multi-institutional group of differently experienced and blinded readers interpreting mpMRI in a cohort of men receiving radical prostatectomy for PCa. Sensitivity increased for all experience levels; however, specificity was dependent on reader experience. Improved sensitivity came from lesions scored as PI-RADS\(<3\) as sensitivity for lesions scored as PI-RADS\(\geq3\) was equal. The authors observed that CAD likely helped readers to overcome the “satisfaction of search” limitation which stems from reduced detection for subsequent lesions after one lesion. However, the improved sensitivity of CAD was balanced by decreased specificity. It is noteworthy that the use of CAD improved agreement between all readers\[107\]. In an
effort to assess the clinical value of a CAD system in a “real-world” scenario, Gaur et al.[108] tested a multi-institution population of 216 men receiving mpMRI. The cases used in this study were diverse in terms of institution-specific acquisition, MR manufacturer, and patient population and nine readers of different experiences were included. Overall, sensitivity for index lesions of mpMRI without and with CAD assistance was comparable. The highest benefit of CAD, as compared to mpMRI alone, was observed for moderately experienced readers detecting TZ csPCa. The authors concluded that CAD might help in reducing the rate of false-positivity and increase sensitivity of moderately experienced readers[108]. Furthermore, recent studies confirmed the clinical utility of CAD use in combination with a Likert score and its generalizability to different field strengths[109,110].

Taken together these findings suggest that CAD represents a promising tool, that generally improves mpMRI efficiency in terms of both diagnostic accuracy and reporting time. The greatest benefit related to the use of CAD potentially lies in improving mpMRI sensitivity for multifocal csPCa, improving diagnostic accuracy in less experienced readers, improving inter-reader agreement and reducing reporting time. Nonetheless, the majority of the available CAD systems are limited to site-specific pre-defined sequences with most of them not considering DCE sequences, raising therefore similar caveats of those raised when comparing bpMRI and mpMRI. To date CAD is limited to research use only, as several aspects still need to be investigated and standardized.

3.9. Discussion

When assessing the role of mpMRI for PCa diagnosis, there are multiple
additional aspects that need to be taken into account before drawing conclusions regarding clinical utility. In this study, the factors that might influence the diagnostic accuracy of MRI have been systematically explored. On the basis of our findings, some points need to be discussed.

First, it is clear that all the factors described in this study affect mpMRI accuracy, and are extremely related to each other in determining the ability of mpMRI in detecting csPCa. This is of crucial importance when comparing different studies. For instance, the use of an ERC helps to increase mpMRI accuracy for anatomic image detail in the posterior part of the prostate, but without DWI this does not help in better detecting csPCa. The benefit of the ERC is related to the magnetic field strength, which varies across the studies. In fact, according to our findings, a 3.0T scan without ERC and a 1.5T mpMRI with ERC reach similar staging accuracies, and thus anatomic detail. However, the ERC causes compression of the PZ, which may even result in missing small central posterior clinically significant cancers [24]. Considering that the addition of ERC is also associated with increased costs, duration of examination, increased artifacts, and is uncomfortable for patients, the use of ERC is suggested only for older 1.5T MR scanners. 3.0T is the recommended magnetic field strength for prostate mpMRI [32]. However, an important issue that is discussed in literature is the gradient-strengths of scanners: steep-gradients are crucial for good DWI. Therefore, gradient-strength steepness may be more important than the field-strength.

Second, the radiologist’s and urologist’s experience is pivotal and affects most of the factors related to mpMRI accuracy, such as the inter-reader agreement and the accuracy of different assessment systems. In this regard, different assessment systems
seem to perform differently according to the experience of the reader. Indeed, standardization of the use of one assessment system (i.e. PI-RADS v2) might actually help radiologists with lower experience in prostate mpMRI to easily reach acceptable accuracy. Also, training and certification are important issues. E.g. following an expert hands-on-course, followed by 100 supervised (double) reads may contribute to better interpretation quality and lower inter-reader variability [49,79]. Similarly, the use of bpMRI appears to be most effective for experienced readers, when good image quality is available, whereas those with low experience might still need DCE-MRI as a backup in order to achieve acceptable accuracy. The non-inferiority of bpMRI would represent a great step towards the widespread use of MRI, allowing to reduce the costs (up to $300 spared/test [84,87]) and to increase the availability of the test (15 minutes saved per patient [79,87]). With this aim in mind, well-designed prospective trials are necessary to provide reliable evidence and to draw solid conclusions on bpMRI for PCa diagnosis [79]. Even when assessing the use of CAD, the highest benefit in terms of diagnostic accuracy is observed in low-experienced radiologists.

To the best of our knowledge, this is the first systematic review exploring the different factors influencing mpMRI accuracy in detecting csPCa. However, this study is limited by the unsatisfactory evidence of most of the studies included, especially related to different MRI protocols, different outcomes, different mpMRI indications, different csPCa prevalence, variable readers’ experience and pathological reference standards with only 22% of studies having both low risk of bias and applicability concerns. Moreover, the lack of reliable pathological reference standards prevented the exploration of the variability of mpMRI performance in the presence of a negative report in the majority of
studies. This heterogeneity prevents the possibility to conduct a proper meta-analysis on
the effect of these factors. Lastly, further factors might be included as potential
influencers of mpMRI accuracy (e.g. apparent coefficient diffusion and b-value, the
gradient-strength of the MRI scanner, type of scanner, communication methods between
radiologist and urologist and the type of study population). Nonetheless, studies
specifically addressing these further factors eligible for inclusion in this review are
scarce. These limitations are the cause that the findings of this review should be
interpreted with caution and within the appropriate context.

4. Conclusions

Even though the role of mpMRI in predicting PCa has been widely demonstrated, several
factors influence its diagnostic accuracy and are affecting each other, with the
radiologist/biopsy-operator experience being the key-confounders. The high
heterogeneity across the studies underlines the need for further studies that clarify how
they impact on the clinical utility of mpMRI for prostate cancer diagnosis. Indeed, the
factors assessed in this study concur with the high variability of mpMRI performance and
the related clinical utility, consequently limiting the widespread use of mpMRI. In order
to deliver the benefits of mpMRI to diagnose csPCa, it is necessary to develop expertise
for both radiologists and urologists, to implement quality control, and to adhere to the
most recent PI-RADS assessment guidelines.
Acknowledgments

Study conception and design: Stabile, Kasivisvanathan, Giannarini, Moore, Padhani, Panebianco, Rosenkrantz, Salomon, Turkbey, Villeirs, Barentsz

Acquisition of data: Stabile, Giganti

Analysis and interpretation of data: Stabile, Giganti

Drafting of the manuscript: Stabile, Giganti

Critical revision of the manuscript for important intellectual content:
Kasivisvanathan, Giannarini, Moore, Padhani, Panebianco, Rosenkrantz, Salomon, Turkbey, Villeirs, Barentsz

Statistical analysis: Stabile

Obtaining funding: None

Administrative, technical, or material support: None

Supervision: Giannarini, Moore, Padhani, Panebianco, Rosenkrantz, Salomon, Turkbey, Villeirs, Barentsz

Other: None

Financial disclosures: Francesco Giganti is funded by the UCL Graduate Research Scholarship and the Brahm PhD scholarship in memory of Chris Adams. Veeru Kasivisvanathan is an Academic Clinical Lecturer funded by the United Kingdom National Institute for Health Research (NIHR). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.
**Figure legend**

**Figure 1:** Preferred Reporting Items for Systematic Reviews and Meta-analysis flow diagram showing the outcome of the initial and additional searches resulting in the full studies included in the review.

**Figure 2:** Overall summary of risk of bias and applicability concerns across studies based on Quality Assessment of Diagnostic Accuracy Studies-2.

**Supplementary Figure 1:** Risk of bias assessment and applicability concern in included studies according to Quality Assessment of Diagnostic Accuracy Studies-2.
References


[34] Visschere P De, Pattyn E, Ost P, Claeys T, Lumen N. Comparison of the Prostate Imaging Reporting and Data System (PI-RADS) Version 1 and 2 in a Cohort of 245 Patients with Histopathological Reference and 2016;100:1–10.


transrectal ultrasound-guided prostate biopsies: Implications for training programs.


Nieuwenhove S, Saussez T, Thiry S, Trefois P, Annet L. Prospective comparison of a fast 1.5-T biparametric with the 3.0-T multiparametric ESUR magnetic resonance imaging protocol as a triage test for men at risk of prostate cancer. BJU Int 2018:0–2.


Alabousi M, Salameh JP, Gusenbauer K, et al. Biparametric vs multiparametric prostate magnetic resonance imaging for the detection of prostate cancer in treatment-naïve patients: a diagnostic test accuracy systematic review and meta-


