Accuracy of Magnetic Resonance Imaging for Local Staging of Prostate Cancer: A Diagnostic Meta-analysis

Maarten de Rooij*, Esther H.J. Hamoen, J. Alfred Witjes, Jelle O. Barentsz, Maroeska M. Rovers

Radboud University Medical Centre, Radboud Institute for Health Sciences, Nijmegen, The Netherlands

Abstract

Context: Correct assessment of tumour stage is crucial for prostate cancer (PCa) management.

Objective: To assess the diagnostic accuracy of magnetic resonance imaging (MRI) for local PCa staging and explore the influence of different imaging protocols.

Evidence acquisition: We searched the PubMed, Embase, and Cochrane databases from 2000 up to August 2014. We included studies that used MRI for detection of extracapsular extension (ECE; T3a), seminal vesicle invasion (SVI; T3b), or overall stage T3 PCa, with prostatectomy as the reference standard. Methodologic quality was assessed using the Quality Assessment of Diagnostic Accuracy Studies tool by two independent reviewers. Data necessary to complete 2 × 2 tables were obtained, and patient, study, and imaging characteristics were extracted. Accuracy was reported for the most experienced or first reader. Results were pooled and plotted in summary receiver operating characteristics plots.

Evidence synthesis: A total of 75 studies (9796 patients) could be analysed. Pooled data for ECE (45 studies, 5681 patients), SVI (34 studies, 5677 patients), and overall stage T3 detection (38 studies, 4001 patients) showed sensitivity and specificity of 0.57 (95% confidence interval [CI] 0.49–0.64) and 0.91 (95% CI 0.88–0.93), 0.58 (95% CI 0.47–0.68) and 0.96 (95% CI 0.95–0.97), and 0.61 (95% CI 0.54–0.67) and 0.88 (95% CI 0.85–0.91), respectively. Functional imaging in addition to T2-weighted imaging and use of higher field strengths (3 T) improved sensitivity for ECE and SVI. ECE sensitivity was not improved by endorectal coil use.

Conclusions: MRI has high specificity but poor and heterogeneous sensitivity for local PCa staging. An endorectal coil showed no additional benefit for ECE detection, but slightly improved sensitivity for SVI detection. Higher field strengths and the use of functional imaging techniques can slightly improve sensitivity.

Patient summary: We pooled the results from all previous studies that evaluated magnetic resonance imaging (MRI) for detection of tumour growth outside the prostate. MRI is not sensitive enough to find all tumours with extraprostatic growth.

* Corresponding author. Radboud University Medical Centre, Radboud Institute for Health Sciences, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands. Tel. +31 24 3616770; Fax: +31 24 3540866. E-mail address: maarten.derooij@radboudumc.nl (M. de Rooij).
1. **Introduction**

Prostate cancer (PCa) has emerged as the most common malignancy among Western males, and is the second leading cause of cancer-related mortality [1]. Traditionally, PCa detection and local staging depend on a combination of diagnostic tests. Serum prostate-specific antigen (PSA) and digital rectal examination (DRE) are used to identify men who need subsequent transrectal ultrasound (TRUS)-guided biopsies [2]. Although these techniques are able to detect PCa and estimate disease aggressiveness, they often underestimate tumour stage and are not accurate for detection of locally advanced disease [3,4]. Correct assessment of the tumour stage is crucial for disease management. Curative treatment is most likely when the TNM stage is T2c, that is, when extracapsular extension (ECE; stage T3a), seminal vesicle invasion (SVI; stage T3b), and metastatic disease (N+ and/or M+) are not present. Magnetic resonance imaging (MRI) is increasingly used to aid prostate biopsy targeting and more accurate detection of PCa [5,6]. MRI can also improve the determination of tumour extent [4]. Many studies have investigated the accuracy of MRI in local staging. Studies differ in their use of magnetic field strengths, endorectal coils (ERCs), and combinations of anatomic and functional MRI techniques. Heterogeneous results are driving the ongoing debate regarding the usefulness of MRI and the best imaging protocol for PCa staging.

So far, two meta-analyses on local staging accuracy have been published. Engelbrecht et al [7] included literature up to 2000, and a more recent survey by Silva et al [8] was restricted to studies using 1.5-T devices with an ERC. Since 2000, local PCa staging has been studied extensively using many different imaging protocols. An updated search without limitations on field strength or coil use is warranted to provide a comprehensive overview of evidence available for the most common imaging protocols. The aim of this meta-analysis was therefore to assess the diagnostic accuracy of MRI for local staging and to analyse the influence of different imaging protocols in men with biopsy-proven PCa, with radical prostatectomy specimens as the reference standard.

2. **Evidence acquisition**

We performed this systematic review and meta-analysis according to the Preferred Reporting Items for Systematic Reviews and Meta-analysis statement [9].

2.1. **Literature search**

We performed a systematic search in PubMed, the Cochrane Central Register of Controlled Trials, and Embase for studies evaluating the diagnostic accuracy of MRI for local PCa staging. Our search strings combined "prostate cancer" synonyms with synonyms for "MRI" and "staging". Reference lists for the articles and reviews included were checked, and related articles were traced to complement the electronic query. Searches were performed from 2000 up to August 12, 2014, and were restricted to publications in English. The EndNote X5 (Thomas Reuters, New York, NY, USA) bibliographic database was used to filter duplicate articles.

2.2. **Study selection**

We included studies if (1) accuracy was assessed for local staging (ECE/T3a, SVI/T3b, or overall stage T3 disease when there was no stratification between T3a and T3b) using MRI as the index test in patients with biopsy-proven PCa; (2) radical prostatectomy was used as the reference standard; and (3) we could reconstruct 2 x 2 tables of ECE, SVI, and/or overall stage T3. We excluded studies that focused on restaging or lymph node or bone staging, or used other imaging techniques for local staging (eg, positron emission tomography [PET], computed tomography [CT], TRUS-guided biopsy, and CT). Two reviewers (M.d.R. and E.H.) independently assessed the eligibility of the papers identified. Any disagreements were resolved by discussion with a third reviewer (M.M.R.).

2.3. **Data extraction**

We extracted data on patient, study, and imaging characteristics for all studies included. Patient characteristics comprised age, PSA level, Gleason score for biopsy tissue, clinical risk group, and prevalence of ECE, SVI, or overall stage T3 disease. Study characteristics included design, sample size, inclusion and exclusion criteria, ECE/SVI criteria, number of readers, reader experience, consensus reading, radical prostatectomy technique, reference standard technique, consecutive patient selection, blinding for reference and index tests, interval between biopsy and MRI, and interval between MRI and radical prostatectomy. Imaging characteristics included the device manufacturer, device model, magnetic field strength, type of coil, and imaging sequence details, that is, T2-weighted imaging (T2WI), dynamic contrast-enhanced MRI (DCE), diffusion-weighted imaging (DWI), and/or magnetic resonance spectroscopic imaging (MRSI). One reviewer extracted data for all studies included using a standardised data extraction form. A second reviewer was contacted to resolve unclear issues by consensus.

2.4. **Methodologic quality assessment**

We assessed the risk of bias and the applicability at study level using the validated Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) scoring system [10]. Four domains are scored: (1) patient selection, which describes the method for patient selection and the patients included; (2) index test, which describes the test being studied and how it was conducted and interpreted; (3) reference standard, which describes the reference standard test used and how it was conducted and interpreted; and (4) flow and timing, which describe the flow of patient inclusion and exclusion and the interval between the index test and the reference standard. The quality assessment was performed by two independent reviewers (M.d.R. and E.H.). Any disagreements were resolved by discussion with a third reviewer (M.M.R.).
2.5. Data synthesis and analysis

Data from each study were summarized in $2 \times 2$ tables of true positive, false positive, true negative, and false negative values to calculate sensitivity and specificity values for ECE, SVI, and/or overall stage T3 detection. If analyses of different imaging protocols were performed within a single study (eg, T2WI and T2WI + DWI + DCE), we chose the most up-to-date technique (ie, T2WI + DWI + DCE). In the case of different cutoff thresholds, we used the most clinically appropriate one, and when accuracy was reported for different readers we chose the most experienced reader, or the first reader when experience was not reported. If staging accuracy had been assessed on a patient basis (eg, ECE/SVI/T3 present) and a region basis (eg, hemi-prostates, sextants) within a study, we included the patient-based analysis to approximate the clinical practice. Authors of studies that did not report sufficient data were asked to provide additional information. To graphically display the sensitivity and specificity at study level, we used RevMan5 software (Cochrane Collaboration, London, UK). We drew forest plots to show variation and to explore heterogeneity for sensitivity and specificity, and plotted the results on a receiver operating characteristic (ROC) curve. Primary outcomes are pooled estimates of sensitivity and specificity with 95% confidence intervals (CIs). We used the MetaDAS tool within SAS statistical software (SAS Institute, Cary, NC, USA) to carry out the meta-analyses [11]. The analyses were imported into RevMan5 and used for fitting hierarchic summary ROC plots. Publication bias was studied for ECE, SVI, and overall stage T3 separately using Deeks funnel plots. This analysis was performed using the R statistical package system (R Foundation for Statistical Computing, Vienna, Austria). To study heterogeneity, we performed sensitivity analyses of several clinically relevant covariates: analysis method (patient or region level), use of an ERC, magnet field strength (1.0, 1.5, or 3.0 T), use of functional imaging techniques in addition to T2WI (DCE, DWI, and/or MRSI), number of participants (<50 or ≥50), absolute prevalence of ECE/SVI/overall stage T3 (<10% or ≥10%), QUADAS applicability risk (high risk absent or present), and risk category for the study population (low, mixed, high, or unclear risk).

3. Evidence synthesis

3.1. Literature search

Figure 1 provides an overview of the literature search and study selection. Our search yielded 4682 unique records, of which 315 remained after screening titles and abstracts. The full-text of these studies was reviewed for eligibility. Studies were excluded if no staging accuracy was reported, when we were unable to reconstruct a $2 \times 2$ contingency table, or when the study was not in English. Manual checking of references cited in the studies included and relevant review articles resulted in two additional papers. This yielded 75 studies for inclusion, of which 45 reported
on ECE (5681 patients), 34 on SVI (5677 patients), and 38 that did not stratify between ECE and SVI, but only reported on overall stage T3 detection (4001 patients) [12–86].

3.2. Study characteristics

Table 1 summarises the patient, technical, and study characteristics. Table 2 provides the characteristics of the studies separately.

3.3. Quality assessment

Overall, the quality was moderate (Fig. 2 and Supplementary Fig. 1). In particular, the risk of bias was unclear for many studies because of a lack of reporting on patient enrolment, on blinding to the index test during interpretation of the reference test, and on the interval between the index test and surgery. In the patient selection domain, seven studies had a high risk of bias because of a case-control design or inappropriate exclusion criteria. Three studies were assigned high concern regarding applicability to our review question because of unclear patient selection and study design, or because only transition-zone tumours were analysed. For the index test domain, three studies had a high risk of bias. Two of these studies used a likelihood scale for assessment of tumour extension, but did not provide a cutoff level. In one study, the readers were not blinded to the reference test when interpreting the index test. Thirteen of the 75 studies had high concern regarding applicability of the index test because there was no information provided on index test characteristics, readers, blinding, and/or interpretation of the MR images. Other studies used a functional technique in only a subset of patients without reporting the results separately. One study performed both 1.5-T and 3.0-T imaging in the same patients and used both techniques together during interpretation. For the reference test domain, only one study had a high risk of bias because the reference standard was interpreted without blinding to the index test. All the studies used prostatectomy specimens as the reference standard, so there were no concerns regarding applicability

![Fig. 2 – Risk of bias and applicability concerns: review of author judgments about each domain presented as percentages for the studies included.](image)
Table 2 – Patient, study, and imaging characteristics of all the studies included

<table>
<thead>
<tr>
<th>First author (year)</th>
<th>Patient characteristics</th>
<th>Study characteristics</th>
<th>Imaging characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age (yr)</td>
<td>PSA (ng/ml)</td>
<td>Gleason score</td>
</tr>
<tr>
<td>Akin (2006) [13]</td>
<td>60</td>
<td>41–75</td>
<td>5.9 2.1-28 Cat 6-8</td>
</tr>
<tr>
<td>Armitage (2013) [15]</td>
<td>64.1</td>
<td>46–74</td>
<td>NR NR NR NR</td>
</tr>
<tr>
<td>Augustin (2009) [16]</td>
<td>62.8</td>
<td>47–71</td>
<td>8.9 1.7–31.7 Cat 6-8</td>
</tr>
<tr>
<td>Bernstein (2000) [17]</td>
<td>NR</td>
<td>60.2</td>
<td>4.3 1.3–36.6 Cat 6-8</td>
</tr>
<tr>
<td>Beyerdsorf (2005) [18]</td>
<td>62</td>
<td>50–72</td>
<td>7.5 2–14 NR NR NR</td>
</tr>
<tr>
<td>Block (2007) [19]</td>
<td>65.4</td>
<td>42–78</td>
<td>9.3 1–48.3 Cat 6-8</td>
</tr>
<tr>
<td>Borre (2005) [21]</td>
<td>NR NR NR NR NR NR NR</td>
<td>148</td>
<td></td>
</tr>
<tr>
<td>Braftjord (2011) [22]</td>
<td>59.3</td>
<td>66 NR NR NR 61-10</td>
<td>179 Cohort R E, S, T3 Patient 2</td>
</tr>
<tr>
<td>Brassell (2004) [23]</td>
<td>62 NR NR NR NR NR 124</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brown (2009) [24]</td>
<td>58 42–73 6.1 0.5-30.5 Cat 6-8</td>
<td>62 Cohort P T3 Patient 2</td>
<td>No description</td>
</tr>
<tr>
<td>Cerantola (2013) [25]</td>
<td>67 NR NR NR NR NR 38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cornud (2000) [28]</td>
<td>65 NR NR NR NR NR 94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cornud (2002) [29]</td>
<td>NR NR NR 14.5 3–66 6 NR 332</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Cornud (2012) [30]  | 63 47–76 7 2.8–23 Cat 6-8 | 178 Cohort T3 | | | | | \n
237
Table 2 (Continued)

<table>
<thead>
<tr>
<th>First author (year)</th>
<th>Patient characteristics</th>
<th>Study characteristics</th>
<th>Imaging characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age (yr)</td>
<td>PSA (ng/ml)</td>
<td>Gleason score</td>
</tr>
<tr>
<td>Park (2007) [60]</td>
<td>63</td>
<td>43–74</td>
<td>8.6</td>
</tr>
<tr>
<td>Park (2010) [61]</td>
<td>63.5</td>
<td>50–72</td>
<td>7.4</td>
</tr>
<tr>
<td>Pugh (2012) [63]</td>
<td>67</td>
<td>42–76</td>
<td>49</td>
</tr>
<tr>
<td>Ren (2009) [64]</td>
<td>68</td>
<td>56–84</td>
<td>12.9</td>
</tr>
<tr>
<td>Renard-Penna (2011)</td>
<td>60</td>
<td>39–71</td>
<td>8</td>
</tr>
<tr>
<td>Roethke (2013) [67]</td>
<td>62.7</td>
<td>42–77</td>
<td>8.9</td>
</tr>
<tr>
<td>Roethke (2014) [66]</td>
<td>62.8</td>
<td>43–76</td>
<td>8.8</td>
</tr>
<tr>
<td>Rosenkrantz (2013) [68]</td>
<td>61.1</td>
<td>47–81</td>
<td>6.6</td>
</tr>
<tr>
<td>Rud (2014) [69]</td>
<td>64</td>
<td>42–74</td>
<td>7.8</td>
</tr>
<tr>
<td>Ruprecht (2012) [70]</td>
<td>63</td>
<td>49–78</td>
<td>NR</td>
</tr>
<tr>
<td>Sala (2000) [71]</td>
<td>59</td>
<td>44–73</td>
<td>5.8</td>
</tr>
<tr>
<td>Somford (2013) [72]</td>
<td>62.4</td>
<td>NR</td>
<td>10</td>
</tr>
<tr>
<td>Soulié (2001) [73]</td>
<td>64.2</td>
<td>49–74</td>
<td>10.9</td>
</tr>
<tr>
<td>Soylu (2013) [74]</td>
<td>68</td>
<td>43–75</td>
<td>12.1</td>
</tr>
<tr>
<td>Tan (2008) [75]</td>
<td>61</td>
<td>33–71</td>
<td>NR</td>
</tr>
<tr>
<td>Tanaka (2013) [76]</td>
<td>67</td>
<td>51–74</td>
<td>7</td>
</tr>
<tr>
<td>Torricelli (2007) [78]</td>
<td>66</td>
<td>57–75</td>
<td>7.8</td>
</tr>
<tr>
<td>Tiao (2013) [79]</td>
<td>68.9</td>
<td>50–85</td>
<td>16.9</td>
</tr>
<tr>
<td>Wang (2004) [81]</td>
<td>57.5</td>
<td>32–74</td>
<td>7.6</td>
</tr>
<tr>
<td>Wang (2007) [82]</td>
<td>59</td>
<td>40–86</td>
<td>NR</td>
</tr>
<tr>
<td>Wang (2014) [80]</td>
<td>64</td>
<td>57–68</td>
<td>7</td>
</tr>
<tr>
<td>Wetter (2006) [83]</td>
<td>66</td>
<td>48–78</td>
<td>8.5</td>
</tr>
<tr>
<td>Xylinas (2011) [84]</td>
<td>63.8</td>
<td>45–75</td>
<td>11.3</td>
</tr>
<tr>
<td>Zhang (2007) [86]</td>
<td>62.2</td>
<td>NR</td>
<td>8.1</td>
</tr>
<tr>
<td>Zhang (2009) [85]</td>
<td>58</td>
<td>40–76</td>
<td>5.3</td>
</tr>
</tbody>
</table>

Cat = categorical; CC = case-control; DCE = dynamic contrast-enhanced MRI; DWI = diffusion-weighted imaging; E = extracapsular extension; ERC = endorectal coil; mpMRI = multiparametric magnetic resonance imaging; MRSI = magnetic resonance spectroscopic imaging; N = no, NR = not reported; P = prospective; PSA = prostate-specific antigen; R = retrospective, S = seminal vesicle invasion; Y = yes.

* Median.
1 interquartile range.
2 consensus reading.
3 Case-control study: 31 patients with and 46 patients without extracapsular extension.
4 Patients underwent staging MRI with and without an endorectal coil. Results reported separately for both groups.
5 Case-control study: 30 patients with and 136 patients without seminal vesicle invasion.
for this domain. For the flow and timing domain, three studies had a high risk of bias. One study did not avoid inappropriate exclusions and the staging accuracy was dependent on the accuracy of tumour detection in the first step of the protocol. The two remaining studies did not use the same reference standard for all patients. No studies were excluded from the analysis on the basis of the quality assessment.

### 3.4. Diagnostic accuracy

Figure 3 and Supplementary Figure 2 show hierarchic summary ROC plots with summary point and 95% CI areas for ECE, SVI, and overall stage T3 detection. The pooled sensitivity and specificity for ECE, SVI, and overall stage T3 were 0.57 (95% CI 0.49–0.65) and 0.91 (95% CI 0.88–0.93), 0.58 (95% CI 0.47–0.68) and 0.97 (95% CI 0.95–0.98), and 0.61 (95% CI 0.54–0.67) and 0.88 (95% CI 0.85–0.91), respectively.

### 3.5. Sensitivity analyses

Figure 4 and Supplementary Figure 3 show the results of sensitivity analyses performed for subgroups of studies to explore the influence of patient characteristics, methodologic differences, and technical details on pooled sensitivity and specificity estimates. Overall, specificity estimates showed comparable results, but sensitivity differences were observed in several sensitivity analyses.

#### 3.5.1. Sensitivity analyses for ECE (stage T3a) detection

Figure 4A shows the results for ECE (T3a) detection. Studies with patient-level analysis (n = 42) had lower sensitivity (0.55, 95% CI 0.47–0.63) than studies reporting at a region level (n = 9; 0.70, 95% CI 0.53–0.83). Studies using T2WI as the only imaging modality (n = 30) had lower sensitivity (0.53, 95% CI 0.44–0.63) than studies using an additional functional technique (n = 18; 0.63, 95% CI 0.51–0.74). Furthermore, the 15 studies with a 3.0-T device had higher sensitivity (0.61, 95% CI 0.48–0.72) but lower specificity (0.88, 95% CI 0.82–0.92) in comparison to studies with 1.0 or 1.5 T (n = 33; sensitivity 0.55, 95% CI 0.45–0.65; specificity 0.92, 95% CI 0.88–0.94). Studies that used a 3.0-T device and no ERC (n = 5) had the highest sensitivity (0.71, 95% CI 0.51–0.86), with specificity of 0.90 (95% CI 0.72–0.97). We found comparable sensitivity and specificity between studies with and without an ERC. In addition, pooled sensitivity and specificity estimates were comparable between studies with ≤50 or ≥50 participants, with absolute ECE prevalence of <10% or ≥10%, and with or without high-risk QUADAS-2 assessment.

#### 3.5.2. Sensitivity analyses for SVI (stage T3b) detection

Figure 4B shows the results for SVI (T3b) detection. Studies reporting on SVI that used an ERC showed a sensitivity of 0.59 (95% CI 0.50–0.67) and specificity of 0.97 (95% CI 0.95–0.98), compared to 0.51 (95% CI 0.23–0.78) and 0.94 (95% CI 0.90–0.97), respectively, in studies without an ERC. Sensitivity was higher for studies that used one or two functional techniques in addition to T2WI (n = 15; 0.64, 95% CI 0.48–0.76) than for studies using T2WI alone (n = 22; 0.53, 95% CI 0.39–0.67). Studies using a 1.0- or 1.5-T device (n = 27) had comparable sensitivity (0.58, 95% CI 0.46–0.70) and specificity (0.97, 95% CI 0.95–0.98) to studies with a 3.0-T scanner (n = 10; 0.57, 95% CI 0.37–0.75; 0.95, 95% CI 0.91–0.97). In studies with a 3.0-T scanner, sensitivity was higher without an ERC (n = 5; 0.65, 95% CI 0.30–0.89) than with an ERC (0.45, 95% CI 0.30–0.60). In studies with a 1.5-T scanner, sensitivity was higher with an ERC (0.62, 95% CI 0.51–0.71) than without an ERC (0.37, 95% CI 0.08–0.80). The combination of 3 T and multiparametric (mp)MRI had the highest sensitivity of 0.73 (n = 5; 95% CI 0.45–0.90) and specificity of 0.95 (95% CI 0.89–0.98). Because of the limited
Fig. 4 – Forest plots of pooled sensitivity and specificity estimates and corresponding 95% confidence intervals (CIs) for all studies overall and for the different sensitivity analyses for (A) extracapsular extension (ECE) and (B) seminal vesicle invasion (SVI). ERC = endorectal coil; mpMRI = multiparametric MRI, T2WI = T2-weighted imaging, QUADAS = Quality Assessment of Diagnostic Accuracy Studies. * = unstable pooled estimates.
number of studies per stratum and the low number of true-positive and false-negative cases, several sensitivity analyses resulted in unstable pooled estimates. Therefore, we were unable to reliably compare studies with and without an absolute SVI prevalence of $\geq 10\%$, with and without $\geq 50$ participants, with different risk populations, and with patient versus region-level analyses.

3.5.3. Sensitivity analyses for overall stage T3 detection

Supplementary Figure 3 shows the results for studies that did not stratify between stage T3a and T3b, but reported on overall stage T3 only. Studies that used an ERC ($n = 25$) had lower sensitivity ($0.57$, $95\%$ CI $0.49–0.64$) and higher specificity ($0.90$, $95\%$ CI $0.86–0.93$) than studies without an ERC ($n = 13$; sensitivity $0.66$, $95\%$ CI $0.55–0.76$; specificity $0.85$, $95\%$ CI $0.78–0.90$). Use of a higher magnetic field strength of $3.0$T yielded higher sensitivity ($0.63$, $95\%$ CI $0.52–0.73$) than $1.0$ or $1.5$ T ($0.60$, $95\%$ CI $0.52–0.67$). Addition of functional imaging techniques to T2WI yielded higher sensitivity ($0.62$, $95\%$ CI $0.52–0.71$) and lower specificity ($0.86$, $95\%$ CI $0.81–0.90$) in comparison to T2WI alone (sensitivity $0.60$, $95\%$ CI $0.51–0.67$, specificity $0.90$, $95\%$ CI $0.85–0.94$). The absolute prevalence and the number of participants did not influence the sensitivity and specificity. We were unable to reliably compare studies with different risk populations, T3 prevalence of $\geq 10\%$, and patient versus region analysis.

3.6. Publication bias

The slope coefficients for Deeks funnel plots for detection of ECE ($p = 0.88$), SVI ($p = 0.88$), and overall stage T3 ($p = 0.80$ for T2 vs T3) suggest symmetry in the data and a low likelihood of publication bias [87].

3.7. Discussion

This meta-analysis of the diagnostic accuracy of MRI for local PCa staging revealed high specificity and poor and heterogeneous sensitivity overall. The pooled sensitivity and specificity for detection of ECE/T3a (45 studies, 5681 patients), SVI/T3b (34 studies, 5677 patients), and overall stage T3 disease (38 studies, 4001 patients) were $0.57$ ($95\%$ CI $0.49–0.64$) and $0.91$ ($95\%$ CI $0.88–0.93$), $0.58$ ($95\%$ CI $0.47–0.68$) and $0.96$ ($95\%$ CI $0.95–0.97$), and $0.61$ ($95\%$ CI $0.54–0.67$) and $0.88$ ($95\%$ CI $0.85–0.91$), respectively. Several of the patient, study, and imaging characteristics explored clearly influenced staging accuracy. This affected sensitivity more, whereas specificity remained relatively stable.

In particular, use of functional techniques in addition to T2WI and of devices with a higher field strength appeared to have a large influence on sensitivity. Use of functional techniques is recommended and widely adopted for PCa detection; for example, the European Society of Uroradiology consensus statement recommends the use of at least two functional techniques for PCa detection [88]. For local staging, functional techniques with high-resolution T2WI are also suggested to help radiologists to focus on lesions suspicious for local ECE [89]. This is in line with the current meta-analysis. When one additional functional technique (DWI, DCE, or MRSI) was used, the detection sensitivity improved for ECE, SVI, and overall stage T3 compared to T2WI alone. When two or more functional techniques were used, the detection sensitivity further improved for ECE, but slightly decreased for overall T3 disease. For SVI detection, the pooled estimates for studies with two additional functional techniques were unstable.

Use of a higher field strength ($3.0$ T instead of $1.0$ or $1.5$ T) improved the detection sensitivity for ECE and overall stage T3. ERC use appeared useful for a field strength of $1.5$ T or in the absence of mpMRI. However, when higher field strengths or additional functional techniques were used, studies that used an ERC showed lower sensitivity than studies without an ERC.

Region-based image interpretation showed higher sensitivity for local staging than patient-based analysis. This is possibly because this technique artificially increases the number of true positives, leading to inflated sensitivity. It is important to be aware of this effect when interpreting studies that report at a region instead of a patient level.

The major strength of our meta-analysis is that it provides a complete and unique overview of the literature since the last extensive meta-analysis by Engelbrecht et al [7] and the survey was not restricted to certain imaging parameters as in the meta-analysis by Silva et al [8]. Therefore, many studies could be included and sensitivity analyses of the most important patient, study, and imaging characteristics were possible. The sensitivity results could aid radiologists and urologists in deciding on which imaging protocols to use for local PCa staging. Some methodologic issues also need to be considered. First, we could not completely explain the heterogeneity, because many studies did not include sufficient information for all of the study characteristics. Information was often missing for blinding to clinical information while interpreting the index test, the risk profile of the study population, and image interpretation methods. In addition, some of the strata were too small to result in stable pooled estimates.

Second, because several studies included multiple analyses within the same patient group (eg, for different readers, functional techniques, or coils), we had to choose one set among the accuracy results presented. Several studies reported results for different experience levels [20,32–34, 39,48,65–67,70]. We decided to include the most up-to-date technique, and preferred the most experienced radiologist. We believe this choice resembles clinical practice and demonstrates the potential benefit of the newest techniques. Even though this is the best approximation of the clinical situation, specificity results were comparable between the least experienced and most experienced readers. However, there was greater variation for the sensitivity values, with worse results for the less experienced readers, although comparable or even better results were also reported. Use of the results from the less experienced readers would have led to only minor changes in the results, and would therefore not bias the overall conclusions of our meta-analysis.

Currently, MRI is the best imaging technique available for assessing ECE in clinical practice. PSA, DRE, and TRUS are
not accurate enough for local staging (T stage), but other imaging techniques (eg, PET-CT) can be of value for detection lymph nodes or distant metastases. These techniques are not accurate enough to assess ECE \[2,90\].

The studies included used different methods to standardise reporting: dichotomisation, Likert scales, or a standardised lexicon. Many studies did not sufficiently describe the reporting method, which precluded a sensitivity analysis regarding this characteristic. Wibmer et al\[91\] showed that a standardised reporting system using a 5-point Likert scale with a standardised lexicon could improve staging accuracy over a nonstandardised approach. In working towards a robust method, an international language should preferably be used, similar to the Prostate Imaging Reporting and Data System (PI-RADS) for PCa detection, to improve detection accuracy \[92\].

MRI is limited for detection of focal (microscopic) ECE, a disease category with favourable prognosis compared to more extensive ECE \[93,94\]. Staging accuracy appears to decrease when cases of focal ECE are incorporated \[20,34\]. However, there is no internationally accepted definition of focal compared to established ECE \[95\]. Refinement is needed for both clinicopathologic and imaging criteria to further investigate the clinical value of MRI in detecting focal ECE.

The current meta-analysis shows that MRI has high specificity but low sensitivity. Traditionally, radiologists have focused on high-specificity reading to minimise unnecessary exclusion of men from curative treatment. This is probably why the meta-analysis revealed high specificity and low sensitivity for MRI. Nowadays, urologists become more interested in high-sensitivity reading to reduce positive surgical margins and preserve neurovascular bundles. Our analyses show that a combination of high magnetic field strength (3.0 T) and functional imaging techniques can slightly improve MRE sensitivity. However, on its own the technique is not good enough to accurately stage local PCa. Prediction of the correct T stage can improve when MRI findings are combined with clinical data such as D’Amico risk categories \[2\]. In the future, a risk-tailored approach might be more appropriate. Unfortunately, results for the sensitivity analyses for these risk groups could not be pooled because of unstable estimates, especially for the low-risk group. The trend showed lower sensitivity values for ECE and SVI detection when the ECE risk was low. A risk-tailored approach is therefore warranted whereby radiologists perform high-specificity reading for high-risk patients to reduce the risk of positive surgical margins, and high-sensitivity reading for patients with low to intermediate risk to select candidates for curative treatment or active surveillance \[72\].

4. Conclusions

MRI appears to have high specificity but poor and heterogeneous sensitivity for detection of ECE, SVI, and overall stage T3. ERC use yielded no additional benefit for ECE detection, but slightly improved the sensitivity of SVI detection. Higher field strengths and the use of additional functional imaging techniques seemed to improve the accuracy of local staging.

Author contributions: Maarten de Rooij had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: de Rooij, Rovers, Barentsz.

Acquisition of data: de Rooij, Hamoen.

Analysis and interpretation of data: de Rooij, Hamoen, Rovers.

Drafting of the manuscript: de Rooij, Rovers, Witjes, Barentsz, Hamoen.

Critical revision of the manuscript for important intellectual content: de Rooij, Rovers, Witjes, Barentsz, Hamoen.

Statistical analysis: de Rooij, Hamoen, Rovers.

Obtaining funding: None.

Administrative, technical, or material support: de Rooij, Hamoen.

Supervision: Witjes, Barentsz, Rovers.

Other: None.

Financial disclosures: Maarten de Rooij certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (eg, employment/affiliation, grants or funding, consultantcies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: None.

Funding/Support and role of the sponsor: None.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.eururo.2015.07.029.

References


