Eur Radiol (2002) 12:2294–2302
DOI 10.1007/s00330-002-1389-z

Marc R. Engelbrecht
Gerrit J. Jager
Robert J. Laheij
André L. M. Verbeek
H. J. van Lier
Jelle O. Barentsz

Local staging of prostate cancer using magnetic resonance imaging: a meta-analysis

Received: 13 September 2001
Revised: 4 February 2002
Accepted: 11 February 2002
Published online: 19 April 2002
© Springer-Verlag 2002

M.R. Engelbrecht (✉) · G.J. Jager
J.O. Barentsz
Department of Radiology,
University Hospital Nijmegen,
P.O. Box 9101, 6500 HB, Nijmegen,
The Netherlands
e-mail: M.Engelbrecht@rdiag.azn.nl
Tel.: +31-24-3614545
Fax: +31-24-3540866

R.J. Laheij · A.L.M. Verbeek
H.J. van Lier
Department of Epidemiology
and Biostatistics, University of Nijmegen,
P.O. Box 9101, 6500 HB, Nijmegen,
The Netherlands

Abstract Our objective was to determine the influence of patient-, study design-, and imaging protocol characteristics on staging performance of MR imaging in prostate cancer. In an electronic literature search and review of bibliographies (January 1984 to May 2000) the articles selected included data on sensitivity and specificity for local staging. Subgroup analyses examined the influence of age, prostate specific antigen, tumor grade, hormonal pretreatment, stage distribution, publication year, department of origin, verification bias, time between biopsy and MR imaging; consensus reading, study design, consecutive patients, sample size, histology preparation, imaging planes, fast spin echo, fat suppression, endorectal coil, field strength, resolution, glucagon, contrast agents, MR spectroscopy, and dynamic contrast-enhanced MRI. Seventy-one articles and five abstracts were included, yielding 146 studies. Missing values were highly prevalent for patient characteristics and study design. Publication year, sample size, histologic gold standard, number of imaging planes, turbo spin echo, endorectal coil, and contrast agents influenced staging performance (*p*=0.05). Due to poor reporting it was not possible to fully explain the heterogeneity of performance presented in the literature. Our results suggest that turbo spin echo, endorectal coil, and multiple imaging planes improve staging performance. Studies with small sample sizes may result in higher staging performance.

Keywords Prostate cancer · Meta-analysis · MR imaging · Diagnostic staging

Introduction

Since 1984, MR imaging has been available for use as a local staging modality for prostate cancer; however, a large variation (heterogeneity) in staging performance remains present in the literature [1, 2, 3, 4]. Various causes may account for heterogeneity in staging performance, such as differences in used reference tests, studied patient population, study methodology, random error, and used thresholds [5].

The purpose of this meta-analysis was to determine how and to which degree these mentioned characteristics influence staging performance of MR imaging in prostate cancer.

Materials and methods

Data sources

Relevant publications were identified in Medline and Embase databases (between January 1984 and May 2000) with the following medical subject heading terms: Prostatic neoplasms; Magnetic resonance imaging; Neoplasm staging; Sensitivity and specificity; Prostat*; Cancer*; MRI*; Magnetic-reson*; Neoplas*; Tumour*; Tumor*; Nuclear Magnetic resonance imaging; Stagin* (all subheadings); these were not restricted to any language. To identify additional relevant references, reference lists of retrieved articles were checked manually, and co-authors were consulted. Furthermore, a manual library search of abstract books of the Radiological Society of North America (RSNA), the International Society of Magnetic Resonance in Medicine (ISMRM), and the European Society of Magnetic Resonance in Medicine (ESMRMB), 1988 to May 2000, was conducted.

Study selection

All retrieved articles were checked by three independent reviewers for the following exclusion criteria (in the used order):

1. Reanalysis/review
2. Only data on nodal staging
3. No comparison with the surgically resected prostate
4. No information on specificity or sensitivity (if sensitivity or specificity could be calculated, the study was included)

Disagreements between reviewers were resolved in consensus. If an exclusion criterion was found, the study was excluded and the reason was recorded. Only the first found exclusion criterion was recorded.

Data extraction

For each study sensitivity and specificity of MR imaging for detection of extracapsular extension (ECE), seminal vesicle invasion (SVI), and detection of clinical (c) stage cT3 were recorded or calculated. Additionally, data were abstracted according to patient group characteristics, methodological characteristics, and MR imaging protocol characteristics using a standardized form.

Patient group characteristics included group average age, prostate specific antigen (PSA) level, tumor grade, hormonal pre-treatment, and percentage of patients with pathological (p) stage T3 (pT3).

Methodological characteristics included publication year, department of origin (radiology, urology, other), verification bias (Were all MR imaging results verified by a reference standard?), time between biopsy and MR imaging, consensus reading, prospective or retrospective study design, consecutive patients, sample size, and histology preparation (whole mount opposed to random sectioning and slice thickness).

The MR imaging protocol characteristics incorporated the number of imaging planes, the imaging sequence (spin echo vs fast spin echo and the use of fat suppression), inclusion of the endorectal coil, magnetic field strength (in Tesla), image resolution (voxel size), use of glucagon, and contrast agents. Finally, the effect of MR spectroscopy and dynamic contrast-enhanced MR imaging on staging performance was evaluated.

It was not possible to perform a subgroup analysis on the criteria for ECE, because the names for the various criteria for ECE differ considerably in the literature. Also the role of microscopic capsular penetration was not analyzed, because in general no definition is given of microscopic capsular extension.

Trapezoidal area under the receiver operating characteristics curve analysis

For each sensitivity and specificity pair, the area under the receiver operating curve (AUC) was calculated using the trapezium method [6]. An advantage of using the AUC, instead of sensitivity and specificity, is that inter-study variability due to different cutoff points of primary studies is decreased [7]. Although the trapezium method underestimates the AUC, it facilitated the comparison between studies. A limitation of using the trapezium method is that comparisons between AUCs are only meaningful if there is a good likelihood that the sensitivity–specificity lines are parallel [8].

We stratified studies according to possible determinants of staging performance and we compared AUCs to evaluate if statistically significant differences were present.

We first used a univariate analysis to determine which characteristics were significant sources of variation. Then we attempted to model the variation between studies by means of multivariate analysis, in which all patient characteristics, methodological characteristics, and MR imaging protocol characteristics (which were significant in the univariate analysis) were simultaneously included. Unfortunately, this was not possible, because there were no studies which reported all mentioned characteristics (convergence problems). A best subset analysis was also not possible for the same reasons. To correct for the variation in the precision of the AUCs caused by studies using smaller and larger numbers of patients, we performed a weighted (for sample size) regression analysis (for the characteristic sample size itself, we used an unweighted regression analysis) [9]. To correct for the dependence between AUCs within the same study population, we used a random-effect model (multilevel model). Student’s *t*-test was used to test for differences between subgroups. A *p*-value of 0.05 or less was considered statistically significant. Analyses were performed with Statistical Analysis System software (SAS 6.12, SAS Institute, Cary, N.C.).

Summary receiver operating characteristics analysis

Characteristics, which caused significant variation in staging performance and low missing values (n=24 missing values) in the trapezium subgroup analyses, were additionally analyzed using summary receiver operating characteristics (ROC) curves. Summary ROC analysis [5] was performed for publication year, consensus reading, prospective vs retrospective study design, sample size, imaging planes, turbo spin-echo imaging, the endorectal coil, and contrast agents. Summary ROC curves were constructed only for studies which used the per-prostate histologic gold standard. The per-prostate histologic gold standard is used when a study compares MR imaging predictions of cT2 vs cT3 with pathology regardless of the location of the tumor extension seen at pathology. For example, MR imaging may predict stage cT3, because ECE is seen on the left side of the prostate. If the pathological ECE is actually on the right side, this fact is ignored using the per-prostate histologic gold standard and the prediction is scored as a correct hit for MR imaging. We used only the per-prostate reference standard to decrease heterogeneity due to different reference standards and consequently to determine more accurately other causes of heterogeneity.

Subgroup analyses using summary ROC curves are accomplished by a transformation of sensitivity and specificity into the logit of the true-positive rate and false-positive rate. Subsequently, the sum and difference of the logit terms were calculated. The sum and difference of the logit terms were plotted and simple linear regression provided a slope and intercept. When the slopes of both regression lines of two subgroups are near zero, a comparison of the intercepts indicates the presence or absence of a statistically significant difference between subgroups. Following the guidelines [5] for fitting summary ROC curves, we obtained corresponding single-number summaries. These are the points on the summary ROC curve where sensitivity and specificity are equal. A *p*-value of 0.05 or less was considered statistically significant. Analyses were performed with Statistical Analysis System (SAS 6.12, SAS Institute, Cary, N.C.).

Results

Literature search

We found 134 articles with the Medline and Embase databases. Articles were excluded for the following reasons: review or reanalysis (*n*=16); nodal staging data instead of local staging (*n*=8); no histologic reference standard (*n*=5); previously published article (*n*=1); not available in library (*n*=3); or no data on sensitivity and specificity (*n*=50). Using bibliographies of retrieved articles and knowledge of co-authors, we additionally included 20 articles. Furthermore, we retrieved 35 eligible abstracts of which we excluded 30 due to republication as an article (*n*=18) or due to absent data on sensitivity and specificity (*n*=12). Finally, we included 71 articles and 5 abstracts for further analysis, containing 146 studies. A study was defined as set of sensitivity and specificity, resulting from one diagnostic evaluation. Therefore, one article or abstract can contain more than one study, e.g., when one article evaluates the same group of subjects using spin-echo imaging and turbo spin-echo imaging. A list of all included articles and abstracts with relevant characteristics is available on request from the authors.

Trapezoidal area under the ROC curve analysis

The patient group characteristics, methodological characteristics, as well as MR imaging protocol characteristics, which resulted in significantly different AUCs, are summarized in Tables 1, 2, 3, and 4.

Although significant differences were found in AUCs between different patient populations (Table 1) and study-design characteristics (Tables 2, 3), missing values were highly prevalent. For example, age showed 59 missing values out of 83 studies; PSA showed 70 missing values out of 83 studies. Because of the high number of missing values for these characteristics, these data may be highly biased.

Table 1 Patient group characteristics. *n* no. of studies; *NS* not significant; *AUC* trapezoidal area under the curve; *ECE* extracapsular extension; *SVI* seminal vesicle invasion; *PSA* prostate specific antigen

Characteristic	AUC ECE		AUC SVI		AUC T3	
Age≤64 years	0.50±0.21	*n*=35	0.56±0.19	*n*=20	0.57±0.12	*n*=17
Age>64 years	0.70±0.21	*n*=7	0.68±0.23	*n*=7	0.67±0.14	*n*=7
Missing value	0.59±0.20	*n*=50	0.64±0.23	*n*=53	0.61±0.13	*n*=59
	p<0.001		*p*=0.01		*p*=0.02	
PSA: 5.9–16.1					0.66±0.09	*n*=8
PSA: 16.2–21.3					0.52±0.05	*n*=5
Missing value	NS		NS		0.60±0.14	*n*=70
					p<0.01	
%pT3<50%	0.43±0.22	*n*=20				
%pT3≥50%	0.65±0.17	*n*=24				
Missing value	0.57±0.2	*n*=48				
		p<0.001		NS		NS

Table 2 Methodological characteristics

Characteristic	AUC ECE		AUC SVI		AUC T3	
Publication year						
1985–1993			0.57±0.25	*n*=16		
1993–2001			0.64±0.21	*n*=64		
	NS		*p*=0.04		NS	
Without verification bias	0.49±0.15	*n*=14				
With verification bias	0.64±0.08	*n*=2				
Missing values	0.58±0.22	*n*=76				
		p=0.03		NS		NS
Not in consensus	0.51±0.20	*n*=49	0.59±0.23	*n*=36	0.55±0.10	*n*=49
Consensus	0.68±0.13	*n*=12	0.73±0.20	*n*=9	0.67±0.10	*n*=14
Missing value	0.61±0.22	*n*=31	0.64±0.21	*n*=35	0.70±0.13	*n*=20
	p<0.001		*p*=0.04		*p*<0.001	
Prospective			0.58±0.24	*n*=35		
Retrospective			0.67±0.23	*n*=22		
Missing values			0.65±0.18	*n*=23		
	NS		*p*<0.01		NS	
Consecutive	0.51±0.24	*n*=29	0.57±0.20	*n*=21	0.56±0.12	*n*=16
Non-consecutive	0.43±0.17	*n*=4	0.84	*n*=1	0.64	*n*=1
Missing values	0.60±0.19	*n*=59	0.64±0.23	*n*=58	0.61±0.14	*n*=66
	p=0.06		*p*=0.04		*p*=0.02	

Table 3 Additional methodological characteristics

Characteristic	AUC ECE		AUC SVI	AUC T3	
Sample size <51	0.62±0.18	*n*=55		0.66±0.14	*n*=39
Sample size >50	0.47±0.22	*n*=36		0.56±0.12	*n*=43
Missing value	0.68	*n*=1		0.65	*n*=1
	p<0.001		NS	*p*<0.001	
ECE per site	0.47±0.22	*n*=27		0.52±0.10	*n*=9
ECE per patient	0.60±0.19	*n*=61		0.62±0.13	*n*=74
Missing value	0.67±0.17	*n*=4			
	p<0.001		NS	*p*<0.001	

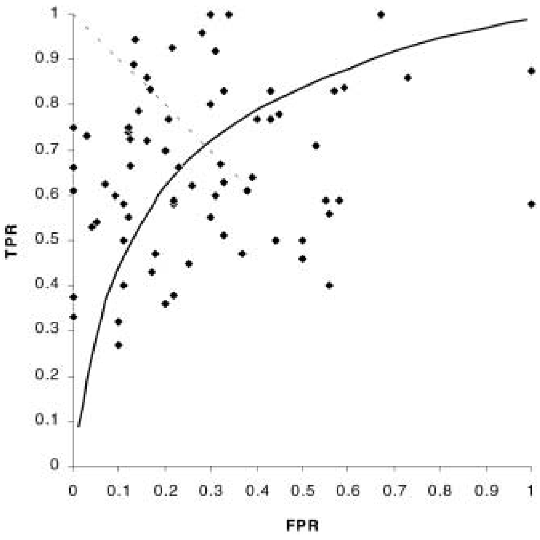

Fig. 1 Summary of receiver operating characteristics (ROC) curve (detection stage T3; all 74 studies) including only studies with the per-prostate scoring method

Publication year, sample size, and reference gold standard yielded limited missing values. Staging performance was lower in studies using more than 50 subjects (*p*<0.001; Tables 2, 3). Staging performance was lower in studies using per-prostate scoring compared with the per-site scoring (*p*<0.001; Tables 2, 3).

Most studies provided enough information about MR imaging protocol characteristics. The number of imaging planes influenced AUCs (*p*=0.012; Table 4). The highest AUC was achieved using two or more imaging planes. Use of turbo spin-echo imaging and the endorectal coil resulted in significant higher AUCs (*p*=0.05). Staging performance was also improved using contrast agents (*p*=0.0024); however, the number of studies was limited (*n*=8). Not enough information was provided on image resolution (missing values: 68 of 83 studies; Table 4).

The following characteristics did not have a significant effect on staging performance: hormonal pre-treatment; department of origin; histology preparation; fat suppression; magnetic field strength; the use of glucagon; MR spectroscopy and dynamic contrast-enhanced MR imaging. This does not necessarily mean that these features are not sources of heterogeneity; however, it may also be possible that too limited studies reported on these characteristics.

Summary ROC analysis

After excluding studies using the per-site histologic gold standard, we included 50 articles and 5 abstracts for summary ROC analysis, yielding 87 studies.

The results of the summary ROC subgroup analyses are summarized in Figs. 1, 2, 3, 4, 5 and 6, and in

Table 4 MR imaging protocol characteristics

Characteristic	AUC ECE		AUC SVI		AUC T3	
1 plane	0.50±0.21	*n*=22	0.43±0.19	*n*=16	0.52±0.10	*n*=33
≥2 planes	0.57±0.23	*n*=49	0.66±0.22	*n*=41	0.64±0.13	*n*=26
Missing value	0.62±0.14	*n*=21	0.70±0.15	*n*=23	0.68±0.12	*n*=24
		p<0.01		*p*=0.012		*p*<0.001
Spin echo			0.49±0.24	*n*=22	0.55±0.12	*n*=39
Turbo spin echo			0.69±0.19	*n*=44	0.65±0.14	*n*=33
Missing value			0.60±0.16	*n*=14	0.66±0.09	*n*=11
		NS		*p*=0.05		*p*<0.01
Without endorectal coil			0.58±0.23	*n*=27	0.54±0.11	*n*=29
With endorectal coil			0.67±0.21	*n*=46	0.65±0.13	*n*=41
Missing value			0.51±0.21	*n*=7	0.60±0.12	*n*=13
		NS		*p*=0.01		*p*=0.01
Voxel>3.0 mm^3			0.59±0.24	*n*=13	0.60±0.16	*n*=9
Voxel≤3.0 mm^3			0.74±0.19	*n*=12	0.76±0.11	*n*=6
Missing value			0.61±0.22	*n*=55	0.59±0.12	*n*=68
		NS		*p*=0.05		*p*=0.02
With contrast agents	0.70±0.15	*n*=8	0.74±0.17	*n*=7	0.76±0.12	*n*=7
Without contrast agents	0.55±0.21	*n*=80	0.61±0.22	*n*=71	0.59±0.13	*n*=74
Missing value	0.70±0.19	*n*=4	0.85±0.14	*n*=2	0.58±0.08	*n*=2
	p<0.001		*p*=0.02		*p*<0.01	

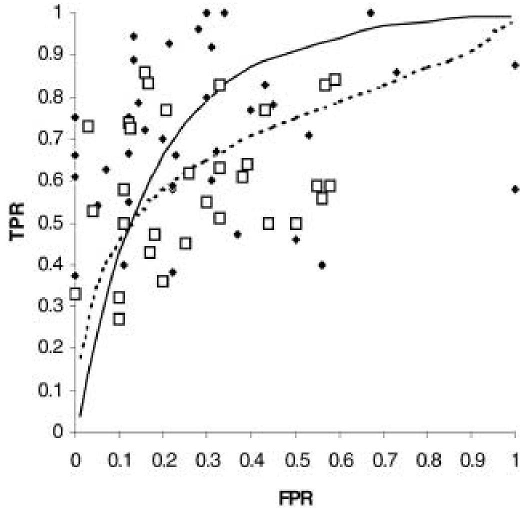

Fig. 2 Summary ROC curve (detection stage T3) for study size. *Diamonds* indicate studies with less than 50 subjects and *squares* indicate studies with 50 or more subjects. *Dotted line* indicates summary ROC curve for studies with 50 or more subjects and *solid line* indicates studies with less than 50 subjects. Differences between both summary ROC curves were significant (p<0.05). *TPR* equal sensitivity; *FPR* equal specificity

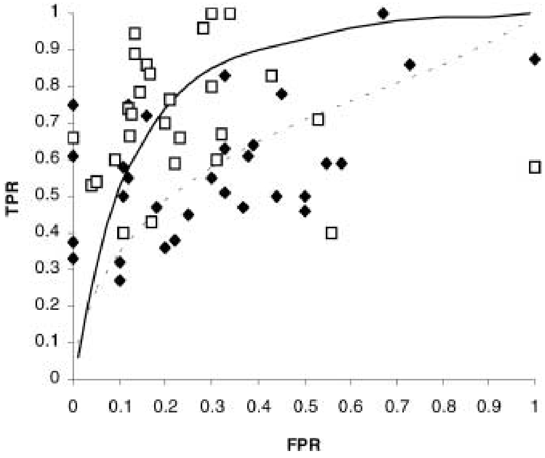

Fig. 4 Summary ROC curves (detection stage T3) for type of spin-echo (SE) imaging used. *Diamonds* indicate studies using SE imaging and *squares* indicate studies using turbo SE (TSE) imaging. The *solid line* indicates summary ROC curve for studies using TSE imaging and the *dotted line* indicates summary ROC curve for studies using SE imaging. Differences between both summary ROC curves were significant (p<0.001)

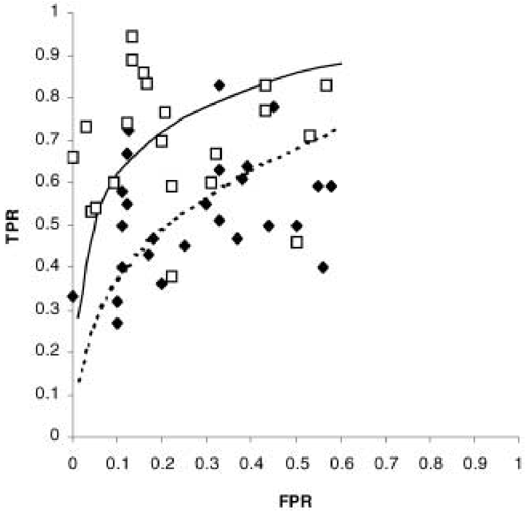

Fig. 3 Summary ROC curves (detection stage T3) for number of imaging planes. *Diamonds* indicate studies using one imaging plane and *squares* indicate studies using two or more imaging planes. The *dotted line* indicates a summary ROC curve for studies using one imaging plane and the *solid line* indicates ROC curves for studies using two or more imaging planes. Differences between both ROC summary curves were significant (p<0.001)

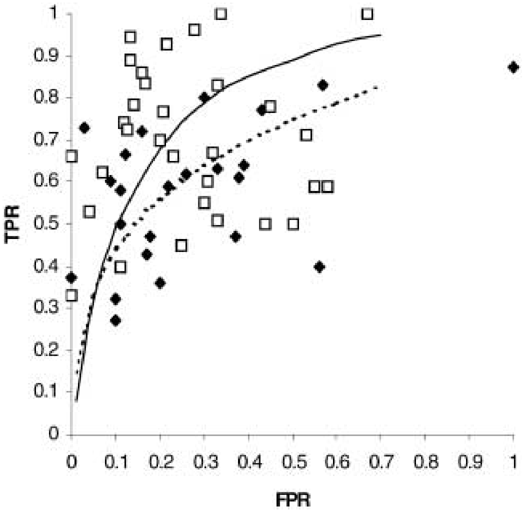

Fig. 5 Summary ROC curves (detection stage T3) for coil type. *Diamonds* indicate studies using no endorectal coil, *squares* indicate studies using the endorectal coil. The *solid line* indicates the summary ROC curve for studies using the endorectal coil and the *dotted line* represents the summary ROC curve for studies using no endorectal coil. Differences between both summary ROC curves were significant (p<0.05)

Table 5. Higher test accuracy is reflected in a summary ROC curve by proximity to the left upper corner of the plot.

Statistically significant differences in staging performance occurred with differences in sample size (p<0.05; Fig. 2), number of imaging planes (p<0.001; Fig. 3), type of spin-echo imaging (p<0.001; Fig. 4), use of the endorectal coil (p<0.05; Fig. 5), and contrast agents (p<0.001; Fig. 6). We did not find significant differences in staging performance for publication year (p=0.49), consensus reading (slopes differed significantly from 0; p=0.03), and prospective vs retrospective study (p=0.52).

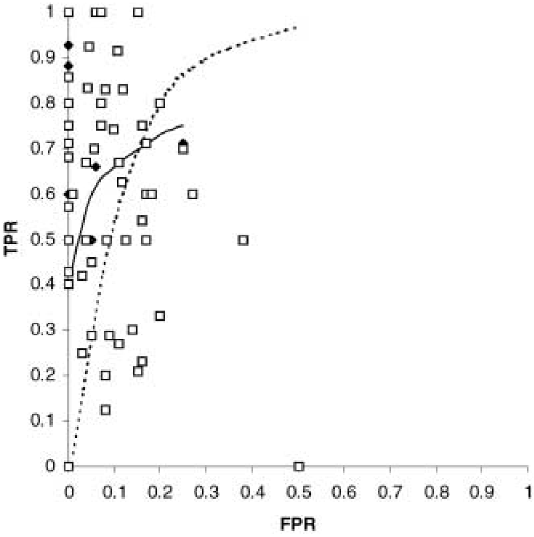

Fig. 6 Summary ROC curves for use of contrast in the detection of seminal vesicle invasion. *Diamonds* indicate studies using the contrast and *squares* indicate studies without contrast. The *solid line* indicates the summary ROC curve for studies using contrast and the *dotted line* indicates the summary ROC for studies without contrast. Differences between both summary ROC curves were significant (p<0.001)

The overall summary ROC curve for studies using per-prostate reference standard appears in Fig. 1. This curve may be considered to be symmetric because the slope of the regression line constructed by regressing D on S for all studies is not statistically different than zero. For a symmetric ROC curve, reporting a single value for test accuracy is both convenient and appropriate, because the odds ratio remains the same at any point along such a curve. This number, representing test accuracy, is the joint maximum sensitivity and specificity, which is the point at which the summary ROC curve intersects the 450 diagonal line (broken line in Fig. 1) designating equal sensitivity (TPR) and specificity (1-FPR). The summary ROC curve for MR imaging in prostate cancer staging (cT2 vs cT3) has a joint maximum sensitivity and specificity of 71%. At a specificity of 80% on this curve, sensitivity was 62%, and at a specificity of 95%, sensitivity was 29%. The summary ROC curve for detection of seminal vesicle invasion has a joint maximum sensitivity and specificity of 82%. At a specificity of 80% on this curve, sensitivity was 85%, and at a specificity of 95%, sensitivity was 27%. The summary ROC curve for detection of extracapsular extension has a joint maximum sensitivity and specificity of 64%. At a specificity of 80% on this curve, sensitivity was 64%, and at a specificity of 95%, sensitivity was 23%.

Discussion

Meta-analysis is a statistical analysis that combines or integrates the results of several independent studies considered to be combinable [10]. Using meta-analysis it is possible to explain variations in study results. Additionally, meta-analysis may be used to highlight important defects in the quality of primary studies and to identify areas of future research [11, 12].

A large heterogeneity in local staging performance of MR imaging in prostate cancer is present in the literature; however, it is not fully understood why staging per-

Table 5 Summary ROC subgroup analysis results. Results of t-tests for symmetry and subgroup comparisons. The characteristics were tested for the detection of ECE, SVI, and T3. Only the significant results are shown

Characteristic	Slope	Intercept	p-value for slope	p-value for difference in intercept	No. of studies
Total no. of patients[a]					
≤50	0.09	2.13	0.61		n=27
>50	–0.27	1.41	0.07	p=0.03	n=31
No. of imaging planes[a]					
1 plane	–0.27	0.95	0.07		n=25
≥2 planes	–0.26	2.24	0.24	p<0.001	n=18
Type of SE imaging used[a]					
SE	–0.19	1.06	0.12		n=26
TSE	0.09	2.48	0.66	p<0.001	n=22
Missing value					n=12
Coil used[a]					
Endorectal coil	–0.004	2.15	0.98		n=27
Other	–0.24	1.35	0.17	p=0.04	n=21
Contrast agents[b]					
Without contrast agents	0.19	2.77	0.29		n=62
With contrast agents	–0.44	2.20	0.30	p<0.001	n=4

[a] For staging T2 vs T3
[b] For detection of SVI

formance varies so much. We could not completely explain the heterogeneity in staging performance. This was partly caused by large numbers of missing values on patient characteristics and study design in the literature, making multivariate analysis not possible. Secondly, we could not evaluate all possible sources of heterogeneity. For example, the role of criteria for capsular penetration, the role of experience, and the role of clinical knowledge could not be investigated.

At the present time the most specific criterion for ECE is asymmetry of the neurovascular bundle (sensitivity 38%, specificity 95%) [13]. The most sensitive criterion is overall impression (sensitivity 68%, specificity 72%) [14]. Other reliable criteria are obliteration of the rectoprostatic angle [13], bulge [14], and extracapsular tumor [14]; however, in this meta-analysis we could not determine the effect of criteria for ECE on staging performance, due to the following reasons: Firstly, each study used different sets of criteria, which made classifying criteria into groups not possible. For example, we found more than 20 different criteria for ECE. Furthermore, the used criteria were often poorly defined or not mentioned.

The role of reading experience could not be analyzed, because most studies did not state a definition of experience; however, from the radiology literature it is known that a learning curve is present for local staging of prostate cancer and that MR staging performance improves with experience [3, 13, 15, 16, 17, 18].

Finally, we tried to determine the effects of clinical information (age, PSA) before reading the images; however, in most studies the available clinical information was not reported. The substantial influence of demographic variables such as age, PSA, and Gleason score on staging performance, has been demonstrated by Getty et al. [19]. Other investigators [20] have shown that the variation in staging performance may be related to the selection of patients, i.e., the number of patients with clinical T1c tumors has increased during the past few years. The prevalence of pathological stage T3 has decreased as well as the size of capsular penetration. Furthermore, the frequency of patients with only microscopic ECE has increased. Due to the high number of missing values (Table 1) we could not reliably evaluate the role of patient selection on the staging performance.

Although large numbers of missing values were present, we did identify characteristics contributing to heterogeneity. Consistent with the meta-analysis by Sonnad et al. [21], we found that studies with small patient numbers and studies using turbo spin echo achieved higher staging performance. Contrary to Sonnad et al. [21], we found that the endorectal coil improved staging performance and we did not find any effect on magnetic field strength on staging performance. Additionally, this meta-analysis demonstrated that more than one imaging plane and contrast media result in higher staging performance; however, limited studies have been performed on the role of contrast agents (*n*=8). We did not find MR spectroscopy to be of significant value in improving staging performance. Although there is some data [22] to support the value of MR spectroscopy in staging prostate cancer, too limited number of studies have been performed in order to perform a meta-analysis.

This meta-analysis covers a time period of 17 years. Due to the fast-moving technology in the field of MR imaging, it is not surprising that current studies differ significantly than previous work. Nevertheless, when evaluating the literature the incremental value of certain technologic improvements on test performance remains unclear. In order to remain as unbiased as possible we included early studies and analyzed each characteristic separately.

The subgroup analysis using the trapezoidal area under the ROC curve proved that staging performance was lower when verification was performed per site of ECE. Apparently, the type of gold standard influences staging performance. Consequently, we used only studies which used one type of reference standard (per prostate) in the summary ROC analysis. We chose the per-prostate reference standard, because the majority of papers used the per-prostate reference standard and because this was considered clinically more relevant; however, the per-prostate reference standard may be less appropriate, because it allows for the following: to call extra-capsular extension on the right, have it be on the left at pathology, and determine that to be a correct hit for MR imaging.

The maximum joint sensitivity and specificity numbers in this study (71%) for detecting stage cT3 are similar to the numbers reported by Sonnad et al. (74%) [21]. A limitation of these estimates is that, due to the heterogeneity in staging performance, it is difficult to discuss the average local staging performance of MR imaging.

Despite its widespread use, meta-analysis continues to be a controversial technique. The pooling of results from a particular set of studies may be inappropriate and meta-analyses of the same issue may reach opposite conclusions; however, by integrating the actual evidence, meta-analysis allows a more objective appraisal, which can help to resolve uncertainties when the original research disagrees. Furthermore, contrary to single studies, it is possible using meta-analysis to reach the necessary number of patients to detect or exclude small effects with confidence [23].

The quality of included studies is of obvious importance for meta-analysis. If the raw material used is flawed, then the conclusions of meta-analytic studies will be equally invalid; however, the type of scale used to assess trial quality can dramatically influence the interpretation of meta-analytic studies [24]. Instead of using quality criteria, we included all studies that met basic entry criteria and analyzed relevant study characteristics individually to determine their influence on staging performance [11, 24, 25].

This study presents the use of two available meta-analytic techniques to analyze local staging performance of MR imaging: the trapezoidal area under the curve and the summary ROC curve. First we analyzed all characteristics using the AUC. Only those characteristics which yielded significant differences and low missing values were additionally analyzed using the summary ROC curves. The advantage of using the trapezoidal area under the curve is that this method facilitates the analysis of large and complicated data sets; however, the accuracy of this method when only one point is present is not proven. The advantage of summary ROC curves is that the curves may be tested for symmetry and therefore it can be determined if subgroup analysis is appropriate. A disadvantage of summary ROC curves is that if a large variability is present between subgroups, the goodness of fit using summary ROC curves will be limited.

In conclusion, it was not possible to fully explain the present heterogeneity in the literature, partly due to poor reporting in primary studies and partly because we could not evaluate the role of clinical information and reader experience; therefore, the quality of reporting in future studies should be improved. Secondly, it is important to consider that those who perform MR imaging in case of prostate cancer should determine their own standard of accuracy by carefully comparing their imaging results with histopathologic findings [16, 26]. Yet, our results suggest that turbo spin echo, the endorectal coil, and multiple imaging planes improve staging performance. Furthermore, we found that studies with small sample sizes may result in higher staging performance, which may be of importance in interpreting the literature.

Acknowledgement Financial support was provided by The Dutch Cancer Society.

References

1. D'Amico AV, Whittington R, Malkowicz SB, Schultz D, Schnall M, Tomaszewski JE et al. (1996) Critical analysis of the ability of the endorectal coil magnetic resonance imaging scan to predict pathologic stage, margin status, and postoperative prostate specific antigen failure in patients with clinically organ confined prostate cancer. J Clin Oncol 14:1770–1777
2. Hricak H, White S, Vigneron D, Kurhanewicz J, Kosco A, Levin D et al. (1994) Carcinoma of the prostate gland: MR imaging with pelvic phased-array coils versus integrated endorectal-pelvic phased-array coils. Radiology 193:703–709
3. Tempany CM, Zhou X, Zerhouni EA, Rifkin MD, Quint LE, Piccoli CW et al. (1994) Staging of prostate cancer: results of radiology diagnostic oncology group project comparison of three MR imaging techniques. Radiology 192:47–54
4. Rifkin MD, Zerhouni EA, Gatsonis CA, Quint LE, Paushter DM, Epstein JI et al. (1990) Comparison of magnetic resonance imaging and ultrasonography in staging early prostate cancer. Results of a multi-institutional cooperative trial. N Engl J Med 323:621–626
5. Littenberg B, Moses LE (1993) Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 13:313–321
6. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36
7. Van der Schouw Y, Straatman H, Verbeek ALM (1994) ROC curves and the areas under them for dichotomized tests: empirical findings for logistically and normally distributed diagnostic test results. Med Decis Making 14: 374–381
8. Habicht J-P (1980) Assessing diagnostic technologies. Science 207:1414
9. Laheij RJF, Straatman H, Jansen JBMJ, Verbeek ALM (1998) Evaluation of commercially available *Helicobacter pylori* serology kits: a review. J Clin Microbiol 36:2803–2809
10. Huque MF (1988) Experiences with meta-analysis in DNA submissions. Proc Biopharm Sect Am Stat Assoc 2:28–33
11. Irwig L, Macaskill P, Glasziou P, Fahey M (1995) Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol 48:119–130
12. Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC et al. (1994) Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 120:667–676
13. Yu KK, Hricak H, Alagappan R, Chernoff DM, Bacchetti P, Zaloudek CJ (1997) Detection of extracapsular extension of prostate carcinoma with endorectal and phased-array coil MR imaging: multivariate feature analysis. Radiology 202:697–702
14. Chelsky MJ, Schnall MD, Seidmon EJ, Pollack HM (1993) Use of endorectal surface coil magnetic resonance imaging for local staging of prostate cancer. J Urol 150:391–395
15. Langlotz C, Schnall M, Pollack H (1995) Staging of prostatic cancer: accuracy of MR imaging. Radiology 194:645–646
16. Harris RD, Schned AR, Heaney JA (1995) Staging of prostate cancer with endorectal MR imaging: lessons from a learning curve. Radiographics 15:813–829
17. Schiebler ML, Yankaskas BC, Tempany CMC, Spritzer CE, Rifkin MD, Pollack HM et al. (1992) MR imaging in adenocarcinoma of the prostate: interobserver variation and efficacy for determining stage C disease. Am J Roentgenol 158:559–562
18. Seltzer SE, Getty DJ, Tempany CMC, Picket RM, Schnall MD, McNeil BJ et al. (1997) Staging prostate cancer with MR imaging: a combined radiologist–computer system. Radiology 202: 219–226
19. Getty DJ, Seltzer SE, Tempany CMC, Picket RM, Swets JA, McNeil BJ (1997) Prostate cancer: relative effect of demographic, clinical, histologic, and MR imaging variables on the accuracy of staging. Radiology 204:471–479
20. Rorvik J, Halvorsen OJ, Albrektsen G, Ersland L, Daehlin L, Haukaas S (1999) MRI with an endorectal coil for staging of clinically localised prostate cancer prior to radical prostatectomy. Eur Radiol 9:29–34
21. Sonnad SS, Langlotz CP, Schwartz JS (2001) Accuracy of MR imaging for staging prostate cancer; a meta-analysis to examine the effect of technologic change. Acad Radiol 8:149–157

22. Yu KK, Scheidler J, Hricak H, Vigneron DB, Zaloudek CJ, Males R et al. (1999) Prostate cancer: prediction of extracapsular extension with endorectal MR imaging and three dimensional proton MR spectroscopic imaging. Radiology 213:481–488
23. Egger M, Davey Smith G (1997) Meta-analysis: potentials and promise. Br Med J 315:1371–1374
24. Juni P, Witschi A, Bloch R, Egger M (1999) The hazards of scoring the quality of clinical trials for meta-analysis. J Am Med Assoc 282:1054–1060
25. Egger M, Davey Smith G (1998) Meta-analysis bias in location and selection of studies. Br Med J 316:61–66
26. Jager GJ, Ruijter ETG, van de Kaa CA, Rosette JJMCH de la, Oosterhof G, Thornbury JR et al. (1996) Local staging of prostate cancer with endorectal MR imaging: correlation with histopathology. Am J Roentgenol 166:845–852