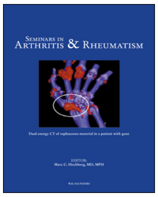
Seminars in Arthritis and Rheumatism 45 (2016) 556–563

Contents lists available at ScienceDirect

Seminars in Arthritis and Rheumatism

journal homepage: www.elsevier.com/locate/semarthrit



Clinical assessment of effusion in knee osteoarthritis—A systematic review

Nasimah Maricar[a,*], Michael J. Callaghan[a], Matthew J. Parkes[a], David T. Felson[a,b,c], Terence W. O'Neill[a,b]

[a] Arthritis Research UK Centre for Epidemiology, Institute of Inflammation and Repair, Faculty of Medical and Human Sciences, Manchester Academic Health Science Centre, University of Manchester, Manchester M13 9PT, UK
[b] NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester NHS Foundation Trust, Manchester Academic Health Sciences Centre, Manchester, UK
[c] Clinical Epidemiology Unit, Boston University School of Medicine, Boston, MA

ARTICLE INFO

Keywords:
Clinical
Tests
Knee
Effusion
Osteoarthritis
Reliability
Sensitivity
Specificity

ABSTRACT

Objective: The aim of this systematic review was to determine the validity and inter- and intra-observer reliability of the assessment of knee joint effusion in osteoarthritis (OA) of the knee.

Methods: MEDLINE, Web of Knowledge, CINAHL, EMBASE, and AMED were searched from their inception to February 2015. Articles were included according to a priori defined criteria: samples containing participants with knee OA; prospective evaluation of clinical tests and assessments of knee effusion that included reliability, sensitivity, and specificity of these tests.

Results: A total of 10 publications were reviewed. Eight of these considered reliability and four on validity of clinical assessments against ultrasound effusion. It was not possible to undertake a meta-analysis of reliability or validity because of differences in study designs and the clinical tests. Intra-observer kappa agreement for visible swelling ranged from 0.37 (suprapatellar) to 1.0 (prepatellar); for bulge sign 0.47 and balloon sign 0.37. Inter-observer kappa agreement for visible swelling ranged from −0.02 (prepatellar) to 0.65 (infrapatellar), the balloon sign −0.11 to 0.82, patellar tap −0.02 to 0.75 and bulge sign kappa −0.04 to 0.14 or reliability coefficient 0.97. Reliability and diagnostic accuracy tended to be better in experienced observers. Very few data looked at performance of individual clinical tests with sensitivity ranging 18.2–85.7% and specificity 35.3–93.3%, both higher with larger effusions.

Conclusion: The majority of unstandardized clinical tests to assess joint effusion in knee OA had relatively low intra- and inter-observer reliability. There is some evidence experience improved reliability and diagnostic accuracy of tests. Currently there is insufficient evidence to recommend any particular test in clinical practice.

Introduction

Knee effusion is common among people with knee osteoarthritis (OA). In those with knee pain and having radiographic OA, nine out of 10 people will have imaging evidence of effusion with 55% having a moderate to large effusion [1]. Clinical assessment for knee effusion is thus an integral part of routine physical examination in knee OA. A range of clinical tests have been used to assess the presence of knee effusion including visible inspection of swelling [2], palpation of the knee [3,4], and a range of dynamic tests including ballottement, patellar tap and the sweep test, though the terminology used in the literature to describe the tests is inconsistent [2,4–7]. To have clinical value and utility, a clinical test should be both valid, in that it should detect an effusion if present and exclude it if not present, and also reliable, meaning that repeated assessments with either the same or different observers result in the same conclusion. Knowledge of the reliability and validity of the currently used clinical tests for knee effusion in knee OA would help inform the optimum approach taken in clinical assessment of these patients and help in informing diagnostic or intervention strategies. To the best of our knowledge there has been no systematic review that pool reliability and validity data from individual studies. We therefore undertook a systematic review with the aim of determining (1) intra- and inter-observer reliability and (2) performance characteristics, of clinical tests for the presence of effusion in knee OA.

[*] Corresponding author.
E-mail address: nasimah.maricar@postgrad.manchester.ac.uk (N. Maricar).

http://dx.doi.org/10.1016/j.semarthrit.2015.10.004

Methods

Search strategies

Publications which reported either intra- and inter-observer reliability of clinical assessments of knee effusion or the validity of clinical assessments using imaging such as ultrasound (US) and magnetic resonance imaging (MRI) were identified from searching six databases up to February 2015: Medline (1948 onwards), Embase (1974 onwards), AMED (1985 onwards), Cumulative Index to Nursing and Allied Health Literature (CINAHL) Plus (1937 onwards), Web of Knowledge (1950 onwards), and the Cochrane Central Registers for Controlled Trials. The databases were searched individually for all possible terms and combination of terms to accommodate differences in their search engines. All medical subject-heading searches (MeSH) were exploded when possible. The key terms used in combination ("OR") were test, examination, assessment, physical, clinical, MRI, US, effusion and swelling which were then used in combination ("AND") with knee. Full details of the MEDLINE search strategy appear in Appendix 1. Hand searches were also performed in addition to additional searches through Google Scholar and Reference Manager Search engines. Reference lists of publications were also searched for other relevant publications. There was no language restriction. Eligibility assessment of trials for inclusion in the review was performed unblinded by 1 reviewer (NM).

Study selection criteria

Publications considered were those that included adults with knee OA based on the American College of Rheumatology (ACR) clinical classification criteria for OA [8] or from detailed clinical and/or radiographic assessment of the knee joint. Due to a limited number of studies on knee OA, for this review, studies with mixed samples of OA and inflammatory arthritis (IA) or knee pain were also included. Studies were selected if they reported evaluation of the validity of clinical assessment of knee effusion against imaging assessment of effusion, or the reliability of clinical assessment of effusion undertaken either by the same observer (intra-observer reliability) or different observers (inter-observer reliability).

Quality assessment

For publications evaluating reliability of the clinical tests for knee joint effusion, study quality was graded using the "reliability" section of COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) with close reference to a separate publication which provided detailed description of the scoring system [9]. To evaluate the quality of diagnostic accuracy studies on knee joint effusion, the diagnostic study appraisal checklist from Centre of Evidence-Based Medicine (CEBM) was used. Two assessors (NM, MJC) independently assessed and scored the publications for quality and reached consensus in cases of disagreement.

Data extraction and analyses

The data from the selected papers were extracted independently by two reviewers (NM, MJP) using a standardised form. Information about the type of study, methods evaluated, subject and observer characteristics, assessment protocols and their results including relevant data for calculation of sensitivity, specificity and reliability of the effusion tests were extracted. In this review, "sensitivity" refers to the ability of a test to identify correctly the presence of effusion in individuals with knee effusion and "specificity" the ability of the test to correctly exclude individuals without knee effusion, assessed by an US or MRI as "gold standard." Intra- and inter-observer reliability is a measure of the degree of agreement when the test is repeated by the same observer or between different observers, respectively. The kappa (κ) statistic was used to denote agreement for dichotomous variables while weighted kappa (κ^{ω}) and reliability coefficient (R_c) was used for ordinal and continuous variables, respectively. We did not undertake a formal meta-analysis for reliability and performance characteristics data because of the different methodologies and differences in reporting among the studies. For kappa and R_c, we considered values of less than 0.2 to indicate poor agreement, between 0.21 and 0.40 fair, 0.41–0.60 moderate, 0.61–0.80 as good and values more than 0.80 to indicate excellent agreement [10].

Results

Search outcome

The results of the search are summarized in Figure 1. In total 10 publications met the inclusion criteria; some of these addressed intra- and/or inter-observer reliability only, others reliability and validity (Tables 1 and 2). Several authors were contacted for further information. Where information was obtained it was included in the review [11,12]. Two articles reported results from the same trials and were considered as one study to avoid potential bias from over-reporting/counting [11,13].

Characteristics of included studies

Eight publications evaluated intra- and/or inter-observer reliability testing for clinical tests of effusion [12,14–20]. The number of assessors varied from two to six [12,14–20]. Two studies reported on reliability of skilled rheumatologists, orthopedic specialists and medical consultants [14,18], four on clinicians of mixed experience such as trainees and consultants [15,16,19,20] and one study reported on reliability of physiotherapists with 7–10 years of post-qualification experience [12].

Four publications reported sensitivity and/or specificity of knee effusion tests where the clinical tests were compared against US as the standard [11,18,20,21]. We found one study [22] that compared clinical assessment of effusion against MRI assessment. However, we were unable to include this study as it did not provide sufficient data to allow the determination of sensitivity and specificity and also the clinical test used to assess the knee effusion was not stated. Of the 10 publications included in the review, seven comprised patients with knee OA [11,14–16,19–21], one comprised patients with knee pain (which included those with OA and inflammatory arthritis) [18], one recent presentation to primary care or to rheumatology clinics or on the waiting list for total knee replacement that included patients with a range of knee pain severity and OA [12], and one comprised a population sample including people with knee OA [17].

Clinical tests used in studies

Clinical tests used to assess knee effusion were categorized as either visual inspection, or dynamic testing involving (1) movement of fluid across the knee with the presence of effusion denoted by reappearance of fluid distension (bulge sign) and (2) pressure over the patella with the presence of effusion determined by palpable ballottement or tapping of the patella against the femoral condyle (patellar tap), or palpable distension of the underlying joint line by fluid fluctuation (balloon sign). We looked

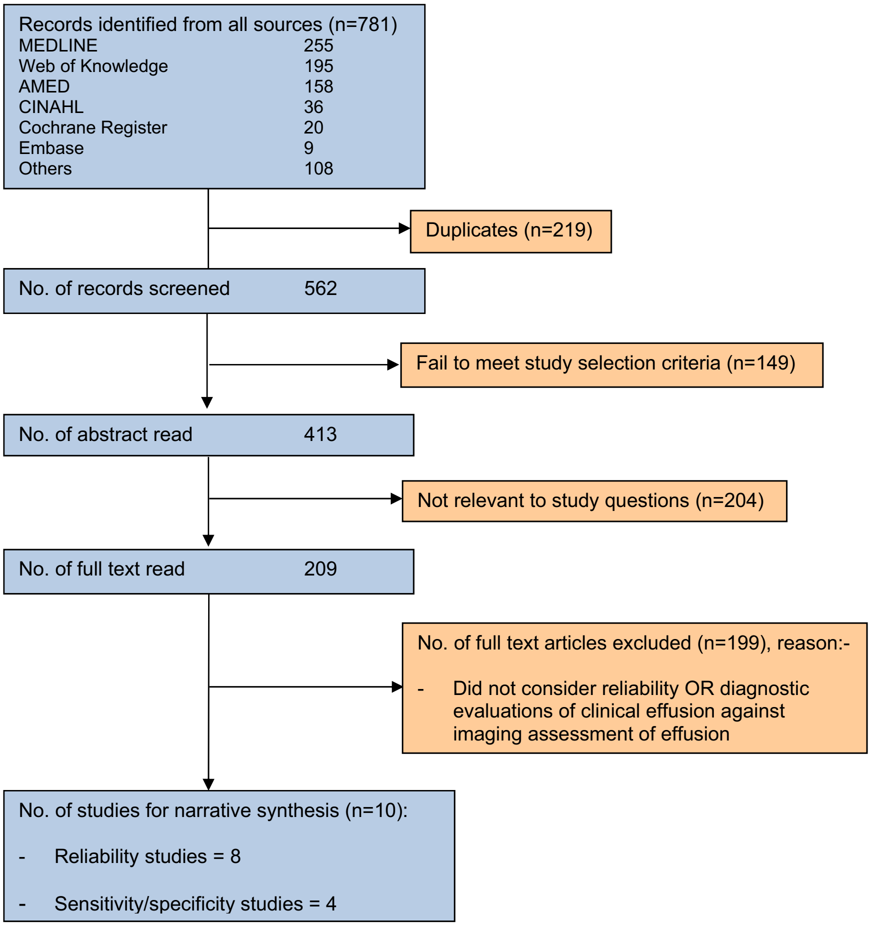

Records identified from all sources (n=781)
MEDLINE 255
Web of Knowledge 195
AMED 158
CINAHL 36
Cochrane Register 20
Embase 9
Others 108
Duplicates (n=219)
No. of records screened 562
Fail to meet study selection criteria (n=149)
No. of abstract read 413
Not relevant to study questions (n=204)
No. of full text read 209
No. of full text articles excluded (n=199), reason:-
- Did not consider reliability OR diagnostic evaluations of clinical effusion against imaging assessment of effusion
No. of studies for narrative synthesis (n=10):
- Reliability studies = 8
- Sensitivity/specificity studies = 4

Fig. 1. Article search and selection process.

also at studies which used combinations of these tests and also when the particular method used was not stated.

Three studies [12,16,21] included assessment of visual inspection. A further study included assessment of visual inspection in combination with other clinical tests [18]. Five studies included assessment of the bulge sign [12,14,16,18,20], with two of these in combination with other tests [18,20]. Two studies included assessment of "ballottement" [16,21], two the "patellar tap" [12,14] and in one ballottement and patellar tap tests were used in combination with bulge sign [20]. Two studies included assessment of "balloon" sign [12,14] where the test procedure was described in one [12]. Hauzeur et al. [18] included assessment of a combination of tests that included inspection, the bulge sign and a "compression" test as described by Moder and Hunder [23] to differentiate joint effusion from synovial thickening. Joint effusion was diagnosed when after compressing the fluid into the suprapatellar pouch, the edge of the resulting bulge disappeared on release of the compression. One other study appeared to include the observation of a bulge sign within its 4-point grading scale [19]. In two studies the clinical test for effusion was described as "non-bony or soft-tissue swelling" without further elaboration on the clinical test(s) used [15,17].

Data quality

All studies on validity of clinical tests satisfied the required quality items though three were unclear in one [11,20,21] and another was unclear in two quality items [18] (Fig. 2). For the reliability studies, all studies scored as either excellent, good, or fair on quality items with the exception of two items which were rated as poor; if adequate sample size was used [14,15] and, when a clinical test was performed, whether the test was repeated more than two times during the reliability testing [12,14,16,17,19] (Fig. 3).

Reliability of clinical tests of knee joint effusion

Visual inspection

One study recorded the kappa agreement for visual inspection of effusion at specific areas around the knee [12]. The highest inter-observer agreement was noted for swelling at the infrapatellar site ($\kappa = 0.65$), with fair reliability over the medial gutter ($\kappa = 0.28$) and poor reliability over the suprapatellar ($\kappa = 0.13$) and prepatellar ($\kappa = -0.02$) sites. Intra-observer agreement was higher particularly at the prepatellar site ($\kappa = 1$), good at the infrapatellar site ($\kappa = 0.66$), moderate at the medial gutter ($\kappa = 0.57$) and fair at the suprapatellar site ($\kappa = 0.37$) [12]. In another study that did not state where the knee visual inspection was carried out, the inter-observer agreement for visible effusion was poor ($\kappa = 0.28$); no intra-observer value was provided [16].

Bulge sign/sweep

Inter-observer agreement was reported as good in one study that used a four-point scale (1 = none, 2 = mild, 3 = moderate, 4 = severe; $R_c = 0.82$) at a pre-standardization stage and excellent ($R_c = 0.97$) when using a dichotomous scale (present/absent) at post-standardization stage [14]. In another study, however,

Table 1
Inter- and intra-observer reliability of assessment of knee effusion

References	Sample	Sample size	No. of observers, time interval between assessments	Clinical tests	Reliability values		
					κ coefficient	95% CI	% Agreement
(A) *Inter-observer reliability of assessment of knee effusion*							
Cibere et al. [14]	OA	6	6 Experienced rheumatologists, same day. Pre- and post-standardization—results presented are post-standardization	Bulge sign	0.97[d]	–	–
				Balloon test	0.82	–	–
				Patellar tap	0.75	–	–
Cushnaghan et al. [15]	OA	8	5 Observers (2 consultants,3 trainees), within 3 h	[a]Non-bony swelling	0.28	0.12–0.44	–
Dervin et al. [16]	OA	115	6 Clinical fellows vs. 7 orthopedic staff, 7–10 days	[a]Visible Effusion	0.28	0.10–0.46	72
				[a]Ballottement	0.19	0.02–0.36	62
				[a]Sweep (bulge sign)	0.14	−0.04–0.32	58
Hart et al. [17]	Population sample including OA	41	2 Observers	[a]Soft Tissue swelling	0.25	–	97.5
Hauzeuer et al. [18]	Knee pain including OA/IA	11 (22 knees)	2 Skilled rheumatologists	Inspection (with other tests)	0.45[b]	0.04–0.85	–
Jones et al. [19]	OA	49 (98 knees)	5 Observers (consultant rheumatologist, consultant geriatrician, 2 rheumatology senior registrars, general medical registrar)	[a]Joint effusion (scale of 0-3 where: 0 = no effusion,1 = positive “bulge,” 2 = moderate effusion, 3 = tense effusion)	0.18 0.22	0–0.41 –	– –
				[a]Joint effusion (dichotomous)			
Ulasli et al. [20]	OA	86 (172 knees)	2 Residents (2 and 4 years’ experience)	Bulge sign, ballottement, patellar tap	0.25	–	–
Wood et al. [12]	Recent presentation to primary care or to rheumatology clinics with knee problems that include OA or on waiting list for total knee replacement	58	2 Out of 3 physiotherapists, 5 minutes apart	Observation of swelling:			
				medial gutter	0.28	–	71.9
				suprapatellar	0.13	–	57.9
				prepatellar	−0.02	–	94.6
				infrapatellar	0.65	–	96.5
				Bulge sign	−0.04	–	57.9
				Balloon sign	−0.11	–	59.7
				Patellar tap	−0.02	–	96.5
				Judgment of effusion (dichotomous)	0.39	–	69.6
				Judgment of effusion (4-point scale)	0.43[b]	–	–
(B) *Intra-observer reliability of assessment of knee effusion*							
Cushnaghan et al. [15]	OA	8	2 Consultants and 3 trainees, at least 30 min and several other patients apart	[a]Non-bony swelling	0.67[c]	0.45–0.89	–
Jones et al. [19]	OA	49 (98 knees)	5 Observers (consultant rheumatologist, consultant geriatrician, 2 rheumatology senior registrars, general medical registrar), within 3 h and after examining other patients in the intervening time	[a]Joint effusion (scale of 0-3 where: 0 = no effusion, 1 = positive “bulge,” 2 = moderate effusion, 3 = tense effusion)	0.78[c] 0.91	0.63–0.93 –	– –
				[a]Joint effusion (dichotomous)			

Wood et al. [12]	Recent presentation to primary care or to rheumatology clinics with knee problems that include OA or on waiting list for total knee replacement	58	1 of 3 physiotherapists, 1 month apart	Observation of swelling:			
				medial gutter	0.57	–	85.1
				suprapatellar	0.37	–	70.2
				prepatellar	1	–	100.0
				infrapatellar	0.66	–	97.9
				Bulge sign	0.47	–	78.7
				Balloon sign	0.37	–	79.2
				Patellar tap	−0.03	–	93.6
				Judgment of effusion (dichotomous)	0.32	–	66.0
				Judgment of effusion (4 point scale)	0.35[b]	–	–

Abbreviations: OA, osteoarthritis; IA, inflammatory arthritis; CI, confidence interval; κ, kappa.

[a] Methods for assessment were not specified.

[b] Weighted kappa.

[c] Average.

[d] Reliability coefficient.

the inter-observer agreement for bulge sign when assessed on a dichotomous scale following post-standardization was found to be poor (κ = −0.04) [12]. Inter-observer agreement was also poor (κ = 0.14) in one other study where the grading was not described [16]. As for intra-observer agreement, when assessed on a dichotomous scale (present/absent) it achieved moderate reliability (κ = 0.47) [12].

Patellar tap/ballottement

In the three studies that evaluated the ballottement or patellar tap test, one reported good inter-observer agreement (κ = 0.75) [14] while the other two reported poor agreement (κ = 0.19, −0.02) [12,16]. Intra-observer agreement investigated in one study was found to be poor (κ = −0.03) [12].

Balloon sign

At post-standardization, the balloon sign scored as absent/present using 6 patients with knee OA was reported to have excellent inter-observer agreement by a group of skilled and experienced rheumatologists (κ = 0.82) [14] but poor (κ = −0.11) by a group of physiotherapists [12]. Wood et al. [12] reported fair intra-observer agreement of the balloon sign performed by a physiotherapist (κ = 0.37).

Combination tests/unknown

Some studies reported reliability using combinations of approaches [12,18,20]. A combination of inspection performed with palpation including the bulge sign and compression test showed moderate inter-observer agreement (κ^{ω} = 0.45) [18] while combination of bulge sign, ballottement and patellar tap tests had fair inter-observer agreement (κ = 0.25) [20]. Intra-observer agreement was not evaluated in these studies. In another study, the bulge sign, balloon test, and inspection over the medial gutter and suprapatellar pouch were used to derive to an overall judgment of effusion and when assessed on a dichotomous scale had fair reliability (κ = 0.39) and on a 4-point scale moderate reliability (κ^{ω} = 0.43) [12]. However, the intra-observer agreement differed little whether a dichotomous (κ = 0.32) or a 4-point scale (κ = 0.35) was used [12]. The inter-observer agreement for non-bony or soft-tissue swelling where the tests used were unclear range from κ = 0.18 to κ = 0.28 while their intra-observer agreement range from κ = 0.67 to κ = 0.91 [15,17,19].

Performance characteristics: Sensitivity and specificity

Due to the ambiguity of the clinical tests used in most studies (Table 2), we could only report on the performance of two individual clinical tests; visual inspection over the suprapatellar area and the ballottement test. In all studies the gold standard was ultrasound effusion appearance. Both tests had poor sensitivity (18.2% and 32.7%) but high specificity (93.3% and 88.9%) for detecting effusion though how these tests were performed and graded was not described [21]. One study performed visual inspection in combination with other clinical tests including the bulge sign and compression and found sensitivity and specificity of detecting effusion increased by effusion size (large effusion 100% and 78% versus moderate effusion 79% and 25%, respectively) [18]. When a combination of bulge sign, ballottement and patellar tap was used, the sensitivity was higher than its specificity [20]. This study also found both sensitivity and specificity of tests increased with seniority of assessor [20]. Another study assessing effusion on

Table 2
Sensitivity and specificity of clinical assessments of knee effusion

References	Sample	Sample size	Assessment	% Sensitivity	% Specificity
D'Agostino et al. [11]	OA	600	[a]Swelling (4-point scale) vs. US (moderate-important vs absent-minimal)	42.5	72.5
Esen et al. [21]	OA	100	Inspection (suprapatellar swelling) vs. US Ballottement vs. US	32.7 18.2	88. 9 93.3
Hauzeur et al. [18]	OA/IA	50 (82 knees)	Inspection (with other tests) vs. US	Moderate effusion: 79 Large effusion: 100	Moderate effusion: 25 Large effusion: 78
Ulasli et al. [20]	OA	86 (172 knees)	Bulge sign, ballottement, patellar tap vs. US	Junior resident: 82.7 Senior resident: 85.7	Junior resident: 35.3 Senior resident: 47.2

Abbreviations: OA, osteoarthritis; IA, inflammatory arthritis; US, ultrasound.

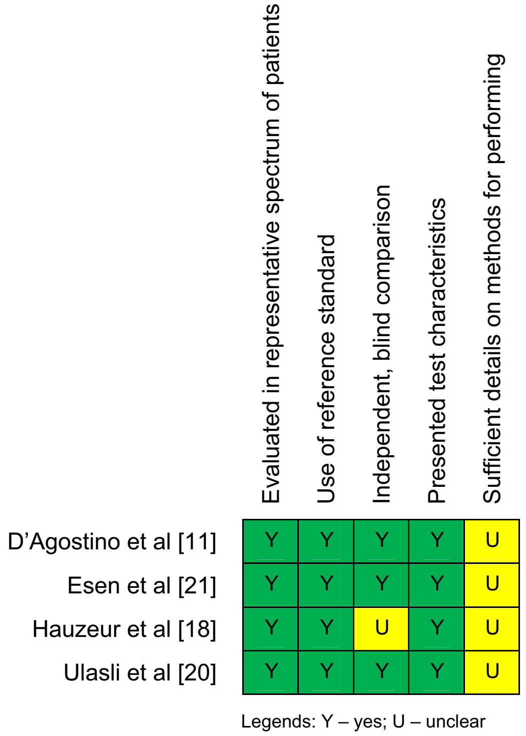
[a] Clinical test/s not specified

a 4-point scale had a sensitivity of 42.5% and specificity of 72.5% [11].

Discussion

This systematic review found a paucity of data surrounding the diagnostic and reliability of clinical assessment of effusion in knee OA. In this systematic review there was a wide range of reliability in the clinical assessment of knee effusion though overall most unstandardized clinical tests had poor to moderate reliability. When compared to effusion as detected by imaging (US), sensitivity and also specificity of clinical assessment varied substantially. As might be expected a combination of clinical tests improved sensitivity. There was some evidence that clinical experience impacted on reliability of the clinical tests with reliability being higher among more skilled clinicians. There was no good evidence based on either reliability or performance characteristics to favor any individual clinical test.

The variability in reliability is perhaps not surprising and likely due to a range of factors including the use of standardized protocols and standardization of the clinical assessment, clinical experience of those testing for effusion, differences in the baseline prevalence of effusion and perhaps also differences in size of the effusion.



	Evaluated in representative spectrum of patients	Use of reference standard	Independent, blind comparison	Presented test characteristics	Sufficient details on methods for performing
D'Agostino et al [11]	Y	Y	Y	Y	U
Esen et al [21]	Y	Y	Y	Y	U
Hauzeur et al [18]	Y	Y	U	Y	U
Ulasli et al [20]	Y	Y	Y	Y	U

Legends: Y – yes; U – unclear

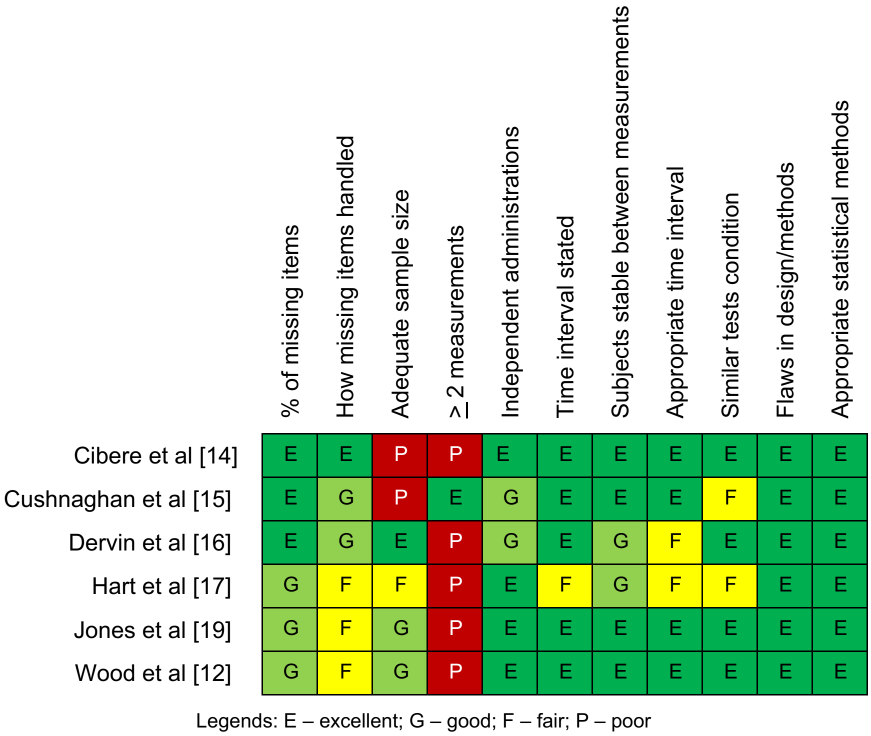
Fig. 2. Methodological quality summary on diagnostic studies.

In terms of standardization, two studies included prior training sessions to standardize the clinical examinations [12,14] though their findings varied. In one [14], changes were made to the examination protocol following formal analyses of reliability and this could have further contributed to the high inter-observer reliability values during the post-standardization stage. The two studies, however, also used different designs and samples that could have led to differences in the findings. Cibere et al. [14] used six rheumatologists who looked at six patients all assessed in a single day, including repeated testing on the same individual by individual assessors raising the possibility of recall bias. Wood et al. [12] assessed pairs of physiotherapists on 58 patients over a 6-month period where the learning effect of standardization may have diminished over time. Additionally, Cibere et al. [14] recruited patients with more severe disease (as determined by the presence of definite radiographic features of OA) than Wood et al. [12] who also included patients who presented to primary care with knee pain (only 14% had OA based on ACR).

Some of the observed variability in reliability assessment may also be due to differences in the baseline frequency of knee effusion. Kappa values are effected by prevalence of the exposure; high or low prevalence in a sample tends to lower the value of kappa despite high percentage total agreement [24]. For example, in the study of Wood et al. [12] the patellar tap had low prevalence of 0.02 with subsequent low kappa value (-0.02) for inter-observer agreement despite a high percentage agreement of 96.5%. Similar findings were also seen for intra-observer agreement for the patellar tap test (prevalence $= 0.03$) where the kappa value was near null ($k = -0.03$) while the percentage agreement was high at 93.6% [12]. Also the differences in the size of the effusion may potentially have impacted on the outcomes—we found evidence from at least one study that performance of the test varied by the size of the underlying effusion with sensitivity being higher for the larger effusion [18].

There was some evidence that clinical experience of the examiners may influence reliability and accuracy. Two studies [14,18] suggested experienced clinicians achieved higher reliability than those which included less experienced assessors [15,16]. One study reported on sensitivity and specificity for assessment of knee effusion for junior and more senior residents; although sensitivity was similar, specificity was higher for the senior resident [20]. These data suggest clinical tests of knee effusion like other physical examination techniques are to some extent dependent on clinical experience.

We had hoped at the outset to determine the optimum approach to clinical characterization of effusion in knee OA; however there were too few studies looking at validity in order to allow determination of the optimal test. A combination of clinical tests appeared to increase sensitivity in detecting effusion



	% of missing items	How missing items handled	Adequate sample size	≥ 2 measurements	Independent administrations	Time interval stated	Subjects stable between measurements	Appropriate time interval	Similar tests condition	Flaws in design/methods	Appropriate statistical methods
Cibere et al [14]	E	E	P	P	E	E	E	E	E	E	E
Cushnaghan et al [15]	E	G	P	E	G	E	E	E	F	E	E
Dervin et al [16]	E	G	E	P	G	E	G	F	E	E	E
Hart et al [17]	G	F	F	P	E	F	G	F	F	E	E
Jones et al [19]	G	F	G	P	E	E	E	E	E	E	E
Wood et al [12]	G	F	G	P	E	E	E	E	E	E	E

Legends: E – excellent; G – good; F – fair; P – poor

Fig. 3. Methodological quality summary on reliability studies.

though with a corresponding reduction in specificity. Presently, therefore, for accurate identification as to the presence or absence of knee effusion imaging is needed. In the clinical setting where for example joint aspiration is being considered in the absence of imaging, it would appear sensible currently to undertake a combination of clinical tests to increase the likelihood of identifying an effusion.

In terms of reliability there was no good evidence that any one test is more reliable than any other—though head to head comparisons were difficult because of the marked difference in design and clinical setting of the studies in which the assessments were undertaken. Further studies using standardized method and comparing different tests are needed. At present where clinical assessment may be important in terms of predicting outcome then it would seem reasonable to suggest that formal assessment of reliability is undertaken in those studies.

Our focus in this review was studies which included subjects with knee OA. In other knee conditions including inflammatory arthritis, we found few studies that have looked at diagnostic accuracy and/or reliability of clinical assessments of knee effusion. In one study of patients with rheumatoid arthritis, the presence of knee effusion assessed by either the presence of fluctuant fluid in the medial or lateral knee compartment or through the presence of patellar tap, was compared with ultrasound assessment; the sensitivity of the clinical assessment was 59% and the specificity was 65% [25]. In another study, the ballottement test was used to assess for the presence of knee effusion following trauma; when compared against MRI effusion the test had a sensitivity of 83% and a specificity of 49% [26]. As for reliability, fluctuation test and patellar tap test when performed on 35 individuals with unilateral knee dysfunction had fair inter-observer agreement of $\kappa = 0.37$ and $\kappa = 0.21$, respectively [27]; however, intra-observer reliability was not assessed in this study.

Our review had several limitations. Many reviewed studies despite scoring well on COSMIN/CEBM had inadequate data and lacked elaboration on how the tests were performed including the scoring system and scales used thereby preventing pooling of their results. Reliability and diagnostic studies are often inadequately indexed in electronic databases and reliability data may also be embedded within other studies and as such not listed within key indexing terms making it difficult to retrieve all published articles. Given this, it is possible that despite an extensive and systematic search we may have missed some publications. Most of the studies included in this review had small samples and used different methods of assessment of reliability including kappa, weighted kappa (ordinal variables) and also reliability coefficients (continuous data) which are not directly comparable and so some caution is needed in interpreting the results. In one study too, the method used to generate the kappa was not described [16]. There were also only four studies that assessed the diagnostic quality of clinical examination of knee joint effusion against imaging such as US; however, only one evaluated a specific dynamic clinical test, that is, ballottement but gave little description on how this test was executed thereby limiting the generalizability of the findings. There was only one study that compared clinical assessment against MRI as the gold standard though there was insufficient data provided for it to be included in our review [22]. Also only a few studies described blinding of observers during reliability testing [12,14,20]. Several studies too were unclear with the time intervals between each set of measurements making it difficult to ascertain if the observed variability may in part be related to true subject change as in the study of Wood et al. [12] where the intra-observer reliability assessments were performed a month apart.

Conclusion

Despite a variety of clinical tests to assess joint effusion in knee OA, the majority of unstandardized tests had relatively low intra- and inter-observer reliability. Reliability and diagnostic accuracy seems to be improved with experience. Compared to individual tests, using a combination of tests for effusion appears to improve sensitivity. Currently there is insufficient evidence to recommend any particular test in clinical practice.

Acknowledgments

This review was supported by a program grant from Arthritis Research UK, (Grant no. 18676). Nasimah Maricar is funded by a Clinical Doctoral Fellowship award from the National Institute for Health Research (NIHR), United Kingdom. We would also like to thank the ROAM team for their assistance.

Appendix A. Supplementary Information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.semarthrit.2015.10.004.

References

[1] Hill CL, Gale DG, Chaisson CE, Skinner K, Kazis L, Gale ME, et al. Knee effusions, popliteal cysts, and synovial thickening: association with knee pain in osteoarthritis. J Rheumatol 2001;28:1330–7.

[2] Isenberg O, Maddison P, Woo P, Glass D, Breedveld. F. Oxford Textbook of Rheumatology. 3rd ed. New York: Oxford University Press; 2004.

[3] Altman RD. Criteria for classification of clinical osteoarthritis. J Rheumatol Suppl 1991;27:10–2.

[4] Davies GJ, Malone T, Bassett FH 3rd. Knee examination. Phys Ther 1980;60: 1565–74.

[5] Sturgill LP, Snyder-Mackler L, Manal TJ, Axe MJ. Interrater reliability of a clinical scale to assess knee joint effusion. J Orthop Sports Phys Ther 2009;39: 845–9.

[6] Currey H, Hull S. Rheumatology for general practitioners. Oxford: Oxford Medical Publications; 1987.

[7] Doherty M, Hazleman B, Hutton C, Maddison P, Perry J. Rheumatology examination and injection techniques. London: W.B. Saunders; 1992.

[8] Altman R, Asch E, Bloch D, Bole G, Borenstein D, Brandt K, et al. Development of criteria for the classification and reporting of osteoarthritis: classification of osteoarthritis of the knee. Arthritis Rheum 1986;29:1039–49.

[9] Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21:651–7.

[10] Landis JR, Koch. GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.

[11] D'Agostino MA, Conaghan P, Le Bars M, Baron G, Grassi W, Martin-Mola E, et al. EULAR report on the use of ultrasonography in painful knee osteoarthritis. Part 1: prevalence of inflammation in osteoarthritis. Ann Rheum Dis 2005;64:1703–9.

[12] Wood L, Peat G, Wilkie R, Hay E, Thomas E, Sim. J. A study of the noninstrumented physical examination of the knee found high observer variability. J Clin Epidemiol 2006;59:512–20.

[13] Conaghan PG, D'Agostino MA, Le Bars M, Baron G, Schmidely N, Wakefield R, et al. Clinical and ultrasonographic predictors of joint replacement for knee osteoarthritis: results from a large, 3-year, prospective EULAR study. Ann Rheum Dis 2010;69:644–7.

[14] Cibere J, Bellamy N, Thorne A, Esdaile JM, McGorm KJ, Chalmers A, et al. Reliability of the knee examination in osteoarthritis: effect of standardization. Arthritis Rheum 2004;50:458–68.

[15] Cushnaghan J, Cooper C, Dieppe P, Kirwan J, McAlindon T, McCrae F. Clinical assessment of osteoarthritis of the knee. Ann Rheum Dis 1990;49:768–70.

[16] Dervin GF, Stiell IG, Wells GA, Rody K, Grabowski J. Physicians' accuracy and interrator reliability for the diagnosis of unstable meniscal tears in patients having osteoarthritis of the knee. Can J Surg 2001;44:267–74.

[17] Hart DJ, Spector TD, Brown P, Wilson P, Doyle DV, Silman AJ. Clinical signs of early osteoarthritis: reproducibility and relation to x ray changes in 541 women in the general population. Ann Rheum Dis 1991;50:467–70.

[18] Hauzeur JP, Mathy L, De Maertelaer V. Comparison between clinical evaluation and ultrasonography in detecting hydrarthrosis of the knee. J Rheumatol 1999;26:2681–3.

[19] Jones A, Hopkinson N, Pattrick M, Berman P, Doherty M. Evaluation of a method for clinically assessing osteoarthritis of the knee. Ann Rheum Dis 1992;51:243–5.

[20] Ulasli AM, Yaman F, Dikici O, Karaman A, Kacar E, Demirdal US. Accuracy in detecting knee effusion with clinical examination and the effect of effusion, the patient's body mass index, and the clinician's experience. Clin Rheumatol 2014;33:1139–43.

[21] Esen S, Akarirmak U, Aydin FY, Unalan H. Clinical evaluation during the acute exacerbation of knee osteoarthritis: the impact of diagnostic ultrasonography. Rheumatol Int 2013;33:711–7.

[22] McAlindon TE, Watt I, McCrae F, Goddard P, Dieppe PA. Magnetic resonance imaging in osteoarthritis of the knee: correlation with radiographic and scintigraphic findings. Ann Rheum Dis 1991;50:14–9.

[23] Moder KG, Hunder GG. Examination of the joints. In: Kelley WN, Ruddy S, Harris ED, Sledge CB, editors. Textbook of rheumatology. Philadelphia: Saunders; 1997. p. 353–70.

[24] Byrt T, Bishop J, Carlin. JB. Bias, prevalence and kappa. J Clin Epidemiol 1993; 46:423–9.

[25] Kane D, Balint PV, Sturrock RD. Ultrasonography is superior to clinical examination in the detection and localization of knee joint effusion in rheumatoid arthritis. J Rheumatol 2003;30:966–71.

[26] Kastelein M, Luijsterburg PA, Wagemakers HP, Bansraj SC, Berger MY, Koes BW, et al. Diagnostic value of history taking and physical examination to assess effusion of the knee in traumatic knee patients in general practice. Arch Phys Med Rehabil 2009;90:82–6.

[27] Fritz JM, Delitto A, Erhard RE, Roman M. An examination of the selective tissue tension scheme, with evidence for the concept of a capsular pattern of the knee. Phys Ther 1998;78:1046–56.