ARTICLES

The Diagnostic Validity of Clinical Tests in Temporomandibular Internal Derangement: A Systematic Review and Meta-analysis

Ève Chaput, BScPT, MClSc, FCAMPT;[*] Anita Gross, BScPT, MSc, Grad.Dip.Manip. Therapy, FCAMPT;[†] Ryan Stewart, BScPT, MClSc, FCAMPT;[*] Gordon Nadeau, BScPT, MClSc, FCAMPT;[*] Charlie H. Goldsmith, PhD[‡§]

ABSTRACT

Purpose: To assess the diagnostic validity of clinical tests for temporomandibular internal derangement relative to magnetic resonance imaging (MRI). ***Methods:*** MEDLINE and Embase were searched from 1994 through 2009. Independent reviewers conducted study selection; risk of bias was assessed using Quality Assessment of studies of Diagnostic Accuracy included in Systematic reviews (QUADAS); $\geq$9/14) and data abstraction. Overall quality of evidence was profiled using Grading of Recommendations Assessment, Development, and Evaluation (GRADE). Agreement was measured using quadratic weighted kappa (κw). Positive (+) or negative (−) likelihood ratios (LR) with 95% CIs were calculated and pooled using the DerSimonian–Laird method and a random-effects model when homogeneous ($I^2 \geq 0.40$, Q-test $p \leq 0.10$). ***Results:*** We selected 8 of 36 studies identified. There is very low quality evidence that deflection (+LR: 6.37 [95% CI, 2.13–19.03]) and crepitation (LR:5.88 [95% CI, 1.95–17.76]) as single tests and crepitation, deflection, pain, and limited mouth opening as a cluster of tests are the most valuable for ruling in internal derangement without reduction (+LR:6.37 [95% CI, 2.13–19.03]), (−LR:0.27 [95% CI, 0.11–0.64]) while the test cluster click, deviation, and pain rules out internal derangement with reduction (−LR: 0.09 [95% CI, 0.01–0.72]). No single test or cluster of tests was conclusive and of significant value for ruling in internal derangement with reduction. ***Conclusions:*** Findings of this review will assist clinicians in deciding which diagnostic tests to use when internal derangement is suspected. The literature search revealed a lack of high-quality studies; further research with adequate description of patient populations, blinded assessments, and both sagittal and coronal MRI planes is therefore recommended.

Key Words: diagnosis; magnetic resonance imaging; predictive value of tests; temporomandibular joint disorders.

RÉSUMÉ

Objectif : Évaluer la validité du diagnostic résultant de tests cliniques en imagerie de résonance magnétique (IRM) pour le dépistage du dérangement interne de l'articulation temporo-mandibulaire (ATM). ***Méthode :*** Une recherche a été effectuée dans les bases de données MEDLINE et Embase pour les années 1994 à 2009. Des examinateurs indépendants ont procédé au choix des études répertoriées ; les risques de biais ont été évalués à l'aide de l'échelle QUADAS (*Quality Assessment of studies of Diagnostic Accuracy included in Systematic reviews*), visant à évaluer la qualité des études sur la fiabilité diagnostique incluses dans les revues systématiques (le pointage obtenu a été $\geq$9/14) et à l'aide d'abstraction des données. Le profil de la qualité globale des preuves a été établi avec l'échelle GRADE (*Grading of Recommendations Assessment, Development, and Evaluation*). Le degré d'accord a été mesuré à l'aide du coefficient quadratique kappa pondéré (κw). Les rapports de vraisemblance (RV) positifs (+) ou négatifs (−) avec intervalles de confiance de 95 % ont été calculés et groupés suivant la méthode de DerSimonian–Laird et à l'aide d'un modèle à effets aléatoires lorsque les données étaient homogènes ($I^2 \geq$ 0,40, test-Q $p \leq$ 0,10). ***Résultats :*** Nous avons retenu 8 des 36 études identifiées. Les preuves voulant que les tests individuels de déviation (RV+ : 6,37 [95 % IC : 2,13–19,03]) et de crépitation ([RV : 5,88 [95 % IC : 1.95–17.76]) soient fiables étaient de très faible qualité, alors qu'un ensemble de tests regroupant crépitation, déviation, douleur et ouverture limitée de la bouche sont les plus valables au moment de déterminer s'il y a effectivement dérangement interne sans réduction (RV+ : 6,37 [95 % IC : 2,13–19,03]), (RV− : 0,27 (95 % IC : 0,11–0,64]), alors que les tests regroupés – claquement, déviation et douleur – déterminent habituellement s'il y a dérangement interne avec réduction (RV− : 0,09 [95 % IC : 0,01–0,72]). Aucun test unique ni aucun groupe de tests ne semblent avoir de valeur considérable ou significative ni concluante pour déterminer avec précision un dérangement interne avec réduction. ***Conclusions :*** Les conclusions de cet examen aideront les cliniciens à décider quels tests diagnostics utiliser

From the: *School of Physiotherapy, University of Western Ontario, London, Ont.; †School of Rehabilitation Science, McMaster University, Hamilton, Ont.; ‡Faculty of Health Science, Simon Fraser University, Burnaby, B.C.; §Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ont.

Correspondence to: Anita Gross, School of Rehabilitation Science, IAHS 4th Fl., McMaster University, 1400 Main St. W., Hamilton, ON L8S 1C7; grossa@mcmaster.ca.

Contributors: All authors designed the study, collected the data, and analyzed and interpreted the data; drafted or critically revised the article; and approved the final draft.

Competing interests: None declared.

Acknowledgements: The authors acknowledge and thank Centric Health Lifemark, our industry partner for HaNSA (Head and Neck Shoulder Arm) research group project financial and thank A. Faghani for his input, and H. Kim, J. Elliott, M. Stott, K. Gorman and C. Leger for their help in earlier stages of this project.

Physiotherapy Canada **2012; 64(2);116–134; doi:10.3138/ptc.2010-54**

lorsqu'ils soupçonnent un dérangement interne. La recherche documentaire a révélé un manque d'études de grande qualité ; des recherches supplémentaires comportant une description adéquate des populations de patients, des évaluations à l'aveugle et des plans sagittaux et coronaires de l'IRM sont donc recommandés.

INTRODUCTION

Temporomandibular joint (TMJ) internal derangement is one of the most common forms of temporomandibular disorders (TMD). The most prevalent signs and symptoms associated with these disorders are tenderness of the masticatory muscles, pain, joint noise, and decreased range of jaw movement.[2] Although these clinical manifestations are common, only a small proportion of those affected (as little as 5%[3] and 7%[4]) seek help.

The natural history of TMD has received minimal study. It is thought that signs and symptoms are transient, self-limiting, and frequently without serious long-term effects.[3] Others have suggested that minor dysfunction may progress to more serious joint disease.[5,6]

The American Academy of Orofacial Pain has estimated that between 40% and 75% of the U.S. population displays at least one sign of TMD, while 33% report at least one symptom.[6] In addition, the 2002 U.S. National Health Institute Survey (NHIS) of 31,000 individuals found an overall prevalence of TMJ and muscle-disorder-type pain of 4.6% (6.3% in women, 2.8% in men).[7] Among the general population, TMD is classified as a condition of young and middle-aged adults rather than of children or the elderly.[8] It was estimated in 1999 that 5.3 million Americans seek treatment for TMD over the course of a 6- to 12-month period, resulting in an estimated $2 billion in direct costs alone.[9]

Numerous imaging modalities are currently in use to assess TMJ disorders, but a gold standard has not been established to identify the "true" clinical diagnosis.[2] Arthrography has been reported to have the best diagnostic outcomes for identifying internal derangement (ID);[10] arthrography is invasive, however, and has poor inter-observer performance relative to magnetic resonance imaging (MRI). Our study used MRI as the reference standard, as it has been found to be the most accurate modality in detecting disk positions and can visualize bilateral joints simultaneously.[1,11]

Clinical diagnosis is the first step in management of TMD, yet information about the validity of TMD tests is very limited. Tests to identify TMD can be either single or clustered; a single test evaluates individual symptoms, such as pain or joint noise, whereas a cluster of tests can assess the cumulative effects of a variety of findings (e.g., TMJ range of motion, pain, and joint noise) to determine a clinical diagnosis. At present, numerous clinical diagnostic tests are administered in clinical settings. The combined ability of these tests to identify ID of the TMJ, however, has not been determined.

Research comparing clinical diagnostic tests with MRI findings to distinguish among TMD types has demonstrated inconsistent results.[11] The purpose of this systematic review of prospective cohort and case-control studies is to determine the diagnostic validity of clinical tests for ID relative to MRI findings. ID includes three categories: internal derangement of any kind (ID), ID with reduction (IDR), and ID without reduction (IDnoR).

METHODS

Review Criteria

Studies were considered for the review if they met the following criteria: case-control or prospective cohort study; participants aged $\geq$14 years with TMD, ID, IDR, or IDnoR; study using tests easily performed in a clinical setting with minimal equipment, either individual (single) or a group of tests (clustered); MRI images including sagittal and coronal planes used.

Search Strategy

We searched MEDLINE and Embase (with English language restrictions) for medical, specialist, and allied health literature from January 1, 1994, to October 1, 2009. MeSH subject headings and keywords used were *temporomandibular joint disorders*, *validity*, *validation*, *diagnosis*, *sensitivity and specificity*, *predictive value of tests*, and *magnetic resonance imagery*. The search was limited to studies involving humans. Our search strategies are outlined in Appendix A.

Three authors independently reviewed the articles and made selections based on the relevance of the title and abstract to the research topic. Calibration was based on the first 50 citation postings for title and abstract screening; for full-text screening, calibration was performed on a training set before article selection. Screening criteria were applied as outlined above.

Assessment of Methodological Quality

Following calibration, the QUADAS quality assessment tool[12] was used by two independent reviewers to assess internal validity (the degree to which the diagnostic tests truly evaluate ID and not other variables; see Table 1). Consensus was reached through discussion. Articles scoring $\geq$9 on the 14-item validity tool were advanced to data abstraction.

To determine the external validity of the diagnostic tests (the degree to which the results are generalizable to other populations or environmental settings), the GRADE system for grading the body of evidence was

Table 1 QUADAS Methodological Quality Data

	QUADAS score														
	Item														
Study	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total (/14)
Barclay et al.[17]	Y	Y	Y	U	Y	Y	Y	Y	Y	U	Y	Y	Y	Y	12
Bertram et al.[21]	Y	Y	Y	U	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	12
Emshoff et al.[22]	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	14
Manfredini et al.[18]	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	14
Marguelles-Bonnet et al.[23]	Y	N	Y	U	Y	Y	Y	U	Y	Y	Y	Y	Y	Y	11
Rudisch et al.[1]	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	14
Taşkaya-Ýýlmaz and Oğütcen-Toller	Y	U	Y	U	Y	Y	Y	Y	Y	U	U	Y	Y	Y	10
Uşümez et al.[19]	Y	Y	Y	U	Y	Y	Y	Y	U	U	U	Y	Y	U	9

QUADAS = Quality Assessment of studies of Diagnostic Accuracy included in Systematic reviews (code item: Y = Yes; N = No; U = unclear)

1. Was the spectrum of patients representative of the patients who will receive the test in practice?
2. Were selection criteria clearly described?
3. Is the reference standard likely to correctly classify the target condition?
4. Is the period between the reference standard (MRI) and index test (clinical test) short enough to be reasonably sure that the target condition did not change between the two tests?
5. Did the whole sample, or a random selection of the sample, receive verification using a reference standard (MRI) of diagnosis?
6. Did patients receive the same reference standard regardless of the index results?
7. Was the reference standard independent of the index test? (i.e., the index test did not form part of the reference standard)
8. Was the execution of the index test described in sufficient detail to permit replication of the test?
9. Was the execution of the reference standard described in sufficient detail to permit its replication?
10. Were the index tests results interpreted without knowledge of the results of the reference standard?
11. Were the reference standard results interpreted without knowledge of the index test?
12. Were the same clinical data available when the tests result was interpreted as would be available when the test is used in practice?
13. Were uninterpreted or intermediate test results reported?
14. Were withdrawals from the study explained?

used. This system, first developed in 2004 by an informal working group with the goal of surmounting the shortcomings of the current grading systems,[13,14] takes into account

- study design;
- limitations (risk of bias using QUADAS);
- indirect outcome;
- indirect patient populations, diagnostic test, comparison test, and comparison;
- imprecise evidence;
- important inconsistency in study results; and
- high probability of publication bias

in determining the quality of evidence for each important outcome.[13,15] Based on this information, GRADE classifies the level of evidence into one of four categories: high, moderate, low, or very low.[13,14] As of 2009, more than 32 medical societies, health regulatory bodies, and health-related branches of government had adopted the GRADE system.[16] It should be noted, however, that a recent review of the GRADE system noted that it has not yet been proven to be reliable or valid, and identified the potential for bias.[16] The GRADE score (reported in Table 2) was determined by two independent reviewers, and consensus was reached through discussion.

Data Abstraction

Three independent reviewers extracted data from eight validated articles. Data were extracted based on MRI type, clinical test, referral pattern, study design, and sample population (prevalence) (see Table 3).

Analysis

Our data analysis included the following five steps: (1) inter-rater reliability testing for study selection and methodological quality assessment; (2) calculation of test accuracy measures such as sensitivity, specificity, and likelihood ratios using Meta-DiSc[24]; (3) testing of heterogeneity; (4) pooling of data; and (5) sub-group analysis.

For article selection and methodological quality assessment, inter-rater reliability between evaluators was calculated using the quadratic weighted kappa statistic (Cicchetti weights [κw with standard deviation (SD)]).[25] For data extraction, raw data were retrieved for 2×2 table construction. Pre-test probability was calculated. The data sets were categorized based on MRI view, clinical test type, and pathological condition. Analysis was based on ID and sub-group pathology of IDR and IDnoR.

Meta-DiSc[24] was used to generate sensitivity, specificity, and likelihood ratios (LR). A 95% confidence interval

Table 2 Summary of Findings across All Diagnostic Tests for TMJ Internal Derangements with MRI as the Gold Standard

		Summary of Findings					Cochrane GRADE Domains							
Study	Tests	Specificity (interval 95%)	Sensitivity (interval 95%)	+LR (interval 95%)	−LR (interval 95%)	Pre-test Probability	Quality of Evidence (GRADE)	Study Design (follow-up period)	Limitations (QUADAS Total Score)	Out-come	Popula-tions, Tests, Indirect Com-parisons	Inconsis-tency (I^2 p-value)	Impreci-sion (Sparse Data; Group Size)	High prob-ability of Publica-tion Bias
Click														
Click vs. MRI for IDR														
Barclay et al.[17]	All clicks	0.82	0.64	2.26	0.28	0.78	Low	—	—	—	−1	NA	−1	NA
$n = 39$	(palpation)	(0.66–0.92)	(0.31–0.89)	(1.02–5.00)	(0.13–0.63)				(12)					
	RC	0.59	0.82	3.24*	0.50*	0.78								
	(palpation)	(0.42–0.74)	(0.48–0.98)	(0.90–11.67)	(0.31–0.80)									
	ARC	0.13	0.91	1.41	0.96	0.78								
	(palpation)	(0.04–0.27)	(0.59–1.00)	(0.18–10.85)	(0.77–1.20)									
	SC	0.10	0.91	1.13	0.99	0.78								
	(palpation)	(0.03–0.24)	(0.59–1.00)	(0.14–9.09)	(0.80–1.22)									
Manfredini et al.[18]	Click	0.46	0.66	1.34	0.82	0.35	Moderate	—	—	—	−1	NA	—	NA
$n = 194$	(palpation)	(0.35–0.56)	(0.58–0.73)	(0.98–1.83)	(0.66–1.03)				(14)					
Uşümez et al.[19]	Click	0.89	0.20	1.12	0.53	0.60	Very low	—	−1	—	−1	NA	−1	NA
$n = 63$	(auscultation)	(0.75–0.97)	(0.07–0.41)	(0.89–1.40)	(0.16–1.77)				(9)					
Click vs. MRI for IDnoR														
Barclay et al.[17]	All clicks	0.56	0.64	1.53	0.70	0.62	Low	—	—	—	−1	NA	−1	NA
$n = 39$	(palpation)	(0.31–0.78)	(0.31–0.89)	(0.63–3.70)	(0.35–1.38)				(12)					
	RC	0.39	0.82	2.14*	0.75	0.62								
	(palpation)	(0.17–0.64)	(0.48–0.98)	(0.54–8.51)	(0.47–1.19)									
	ARC	0.06	0.91	0.61	1.04	0.62								
	(palpation)	(0.00–0.27)	(0.59–1.00)	(0.04–8.81)	(0.84–1.29)									
	SC	0.11	0.91	1.22	0.98	0.62								
	(palpation)	(0.01–0.35)	(0.59–1.00)	(0.12–11.95)	(0.76–1.25)									
Manfredini et al.[18]	Click	0.49	0.66	1.44	0.77	0.45	Moderate	—	—	—	−1	NA	—	NA
$n = 194$	(palpation)	(0.40–0.58)	(0.58–0.73)	(1.09–1.90)	(0.63–0.94)				(14)					
Uşümez et al.[19]	Click	0.29	0.20	0.37	3.53	0.40	Very low	—	−1	—	−1	NA	−1	NA
$n = 63$	(auscultation)	(0.10–0.56)	(0.07–0.41)	(0.17–0.79)	(1.52–8.19)				(9)					
Click vs. MRI for ID														
Barclay et al.[17]	All clicks	0.74	0.64	2.03*	0.41*	0.84	Low	—	—	—	−1	NA	−1	NA
$n = 39$	(palpation)	(0.60–0.84)	(0.31–0.89)	(0.91–4.50)	(0.22–0.77)				(12)					
	RC	0.53	0.82	2.89*	0.58	0.84								
	(palpation)	(0.39–0.66)	(0.48–0.98)	(0.81–10.39)	(0.39–0.86)									
	ARC	0.11	0.91	1.16	0.98	0.84								
	(palpation)	(0.04–0.22)	(0.59–1.00)	(0.15–8.70)	(0.76–1.25)									
	SC	0.11	0.91	1.16	0.98	0.84								
	(palpation)	(0.04–0.22)	(0.59–1.00)	(0.15–8.70)	(0.80–1.21)									
Manfredini et al.[18]	Click	0.48	0.66	1.40	0.79	0.57	Moderate	—	—	—	−1	NA	—	NA
$n = 194$	(palpation)	(0.41–0.54)	(0.58–0.73)	(1.09–1.80)	(0.67–0.94)				(14)					

(Continued)

Table 2 *(Continued)*

		Summary of Findings					Cochrane GRADE Domains							
Study	Tests	Specificity (interval 95%)	Sensitivity (interval 95%)	+LR (interval 95%)	−LR (interval 95%)	Pre-test Probability	Quality of Evidence (GRADE)	Study Design (follow-up period)	Limitations (QUADAS Total Score)	Out-come	Popula-tions, Tests, Indirect Com-parisons	Inconsis-tency (I^2 *p*-value)	Impreci-sion (Sparse Data; Group Size)	High prob-ability of Publica-tion Bias
Taşkaya-Yýlmaz and Oğütcen-Toller[20] $n = 73$	Click (N/A)	0.95 (0.90–0.98)	NE	NE	NE	1	Very low	—	−1 (10)	—	−1	NA	−1	NA
Uşümez et al.[19] $n = 63$	Click (auscultation)	0.71 (0.57–0.82)	0.20 (0.07–0.41)	0.89 (0.68–1.15)	1.45 (0.60–3.53)	0.69	Very low	—	−19)	— 0	−1	NA	−1	NA
Pain														
Pain vs. MRI for IDR														
Bertram et al.[21]† $n = 131$	Pain in TMJ on palpation/function/ opening	0.45 (0.33–0.57)	0.66 (0.55–0.76)	1.32 (0.90–1.95)	0.84 (0.65–1.08)	0.46	Moderate	—	— (12)	—	—	—	—	NA
Rudisch et al.[1]† $n = 41$	Pain in TMJ on palpation/function/ opening	0.53 (0.28–0.77)	0.73 (0.54–0.88)	1.99 (0.94–4.18)	0.64 (0.37–1.11)	0.36	Moderate	—	— (14)	—	—	—	−1	NA
Pooled data†	Pain in TMJ on palpation/function/ opening	0.46 (0.36–0.57)	068 (0.59–0.76)	1.44 (1.02–2.04)	0.80 (0.63–1.00)	—	Moderate	—	—	—	—	$I^2 = 0\%$ $p = 0.343$–0.389	—	—
Taşkaya-Yýlmaz and Oğütcen-Toller[20] $n = 73$	Pain in TMJ	0.60 (0.49–0.71)	NE	NE	NE	1.00	Very low	—	−1 (10)	—	−1	NA	−1	NA
	Pain in ear	0.15 (0.09–0.25)	NE	NE	NE	1.00								
	Tenderness lat pter	0.85 (0.75–0.91)	NE	NE	NE	1.00								
	Tenderness med pter	0.55 (0.44–0.66)	NE	NE	NE	1.00								
Uşümez et al.[19] $n = 40$	Joint and muscle palpation for pain	1.00 (0.91–1.00)	0.20 (0.07–0.41)	NE	NE	0.60	Very low	—	−1 (9)	—	−1	NA	−1	NA
Bertram et al.[21]† $n = 131$	Pain in TMJ on palpation/function/ opening	0.68 (0.58–0.77)	0.66 (0.55–0.76)	2.01* (1.45–2.78)	0.49* (0.25–0.67)	0.54	Moderate	—	— (12)	—	—	—	—	NA
Rudisch et al.[1]† $n = 41$	Pain in TMJ on palpation/function/ opening	0.69 (0.51–0.83)	0.73 (0.54–0.88)	2.57* (1.36–4.85)	0.43* (0.25–0.73)	0.54	Moderate	—	— (14)	—	—	—	−1	NA
Pooled data†	Pain in TMJ on palpation/function/ opening	0.68 (0.59–0.76)	0.68 (0.59–0.76)	2.11* (1.58–2.82)	0.47* (0.36–0.62)	—	Moderate	—	—	—	—	$I^2 = 0\%$ $p = 0.494$–0.682	—	—

(Continued)

Table 2 *(Continued)*

		Summary of Findings					Cochrane GRADE Domains							
Study	Tests	Specificity (interval 95%)	Sensitivity (interval 95%)	+LR (interval 95%)	−LR (interval 95%)	Pre-test Probability	Quality of Evidence (GRADE)	Study Design (follow-up period)	Limitations (QUADAS Total Score)	Out-come	Popula-tions, Tests, Indirect Com-parisons	Inconsis-tency (I^2 *p*-value)	Impreci-sion (Sparse Data; Group Size)	High prob-ability of Publica-tion Bias
Pain vs. MRI for IDnoR														
Taşkaya-Yýlmaz and Oğütcen-Toller[20]	Pain in TMJ	0.85 (0.71–0.94)	NE	NE	NE	1.00	Very low	—	−1 (10)	—	−1	NA	−1	NA
n = 73	Pain in ear	0.36 (0.22–0.51)	NE	NE	NE	1.00								
	Tenderness lat pter	0.84 (0.71–0.94)	NE	NE	NE	1.00								
	Tenderness med pter	0.64 (0.49–0.78)	NE	NE	NE	1.00								
Uşümez et al.[19]	Joint and muscle	1.00	0.20	NE	NE	0.40	Very low	—	−1	—	−1	NA	−1	NA
n = 40	palpation for pain	(0.80–1.00)	(0.07–0.41)						(9)					
Pain vs. MRI for ID														
Bertram et al.[21]†	Pain in TMJ on	0.58	0.66	1.72	0.63	0.67	Moderate	—	—	—	—	—	—	NA
n = 131	palpation/function/ opening	(0.50–0.65)	(0.55–0.76)	(1.25–2.37)	(0.50–0.80)				(12)					
Emshoff et al.[22]†	Pain in TMJ on	0.50	0.67	1.50	0.75	0.65	Moderate	—	—	—	—	—	—	NA
n = 194	palpation/function/ opening	(0.43–0.56)	(0.58–0.75)	(1.15–1.96)	(0.64–0.89)				(14)					
Rudisch et al.[1]†	Pain in TMJ on	0.63	0.73	2.38*	0.50*	0.63	Moderate	—	—	—	—	—	−1	NA
n = 41	palpation/function/ opening	(0.49–0.76)	(0.54–0.88)	(1.27–4.46)	(0.33–0.76)				(14)					
Pooled data†	Pain in TMJ on palpation/function/ opening	0.54 (0.50–0.59)	0.67 (0.61–0.73)	1.65 (1.36–2.01)	0.66 (0.54–0.80)	—	Moderate	—	—	—	—	I^2 = 0–48% *p* = 0.145–0.393	—	—
Uşümez et al.[19]	Joint and muscle	1.00	0.20	NE	NE	0.69	Very low	—	−1	—	—	NA	−1	NA
n = 40	palpation for pain	(0.94–1.00)	(0.07–0.41)						(9)					
Barclay et al.[17]†	ROM/muscle-jt	0.80	0.69	2.60	0.29	0.75	Low	—	—	—	−1	NA	−1	NA
n = 39	palpation/jt sounds	(0.64–0.91)	(0.39–0.91)	(1.13–5.96)	(0.14–0.59)				(12)					
Marguelles-Bonnet et al.[23]†	ROM/muscle-jt	0.64	0.79	3.01	0.46	0.31	Low	—	−1	—	−1	NA	—	NA
n = 242	palpation/jt sounds	(0.55–0.71)	(0.74–0.83)	(0.36–3.82)	(0.37–0.58)				(11)					
Pooled data†	ROM/muscle-joint palpation/joint sounds	0.67 (0.60–0.74)	0.79 (0.74–0.83)	2.97* (2.36–3.74)	0.42* (0.28–0.61)	—	Low	—	—	—	—	I^2 = 0–34% *p* = 0.212–0.740	—	—
Uşümez et al.[19]	click/deviation/pain	0.97	0.28	1.35	0.09‡	0.60	Very low	—	−1	—	−1	NA	−1	NA
n = 40		(0.86–1.00)	(0.12–0.49)	(1.05–1.74)	(0.01–0.72)				(9)					
	crepitation/deflection/ pain/limited mouth opening	0.05 (0.01–0.18)	0.88 (0.69–0.97)	0.44 (0.08–2.44)	1.08 (0.91–1.27)	0.60								

(Continued)

Table 2 *(Continued)*

		Summary of Findings					Cochrane GRADE Domains							
Study	Tests	Specificity (interval 95%)	Sensitivity (interval 95%)	+LR (interval 95%)	−LR (interval 95%)	Pre-test Probability	Quality of Evidence (GRADE)	Study Design (follow-up period)	Limitations (QUADAS Total Score)	Out-come	Popula-tions, Tests, Indirect Com-parisons	Inconsis-tency (I^2 *p*-value)	Impreci-sion (Sparse Data; Group Size)	High prob-ability of Publica-tion Bias
Test Clusters														
Cluster vs. MRI for IDnoR														
Barclay et al.[17]†	ROM/muscle-joint	0.13	1.00	3.33	0.88	0.47	Low	—	0	—	—	−1	−1	NA
n = 39	palpation/joint sounds	(0.00–0.53)	(0.66–1.00)	(0.15–71.90)	(0.63–1.21)				(12)					
Marguelles-Bonnet et al.[23]†	ROM/muscle-jt	0.56	0.86	4.14	0.51	0.39	Low	—	−1	—	—	−1	—	NA
n = 242	palpation/joint sounds	(0.49–0.63)	(0.82–0.90)	(3.02–5.67)	(0.43–0.60)				(11)					
Pooled data†	ROM/muscle-joint	0.54	0.87	4.13*	0.66	—	Low	—	—	—	—	I^2 = 0–90%	—	—
	palpation/joint sounds†	(0.47–0.61)	(0.83–0.90)	(3.02–5.64)	(0.37–1.16)							*p* = 0.020–0.891		
Uşümez et al.[19]	click/deviation/pain	0.06	0.28	0.08	3.36	0.40	Very low	—	−1	—	−1	NA	−1	NA
n = 40		(0.00–0.29)	(0.12–0.49)	(0.01–0.56)	(1.77–6.37)				(9)					
	crepitation/deflection/ pain/limited mouth opening	0.76 (0.50–0.93)	0.88 (0.69–0.97)	6.37‡ (2.13–19.03)	0.27* (0.11–0.64)									
Cluster vs. MRI for ID														
Barclay et al.[17]	ROM/muscle-joint	0.75	0.69	2.44*	0.36*	0.82	Low	—	0	—	−1	NA	−1	NA
n = 39	palpation/joint sounds	(0.62–0.85)	(0.39–0.91)	(1.06–5.58)	(0.20–0.64)				(12)					
Uşümez et al.[19]	click/deviation/pain	0.69	0.28	0.96	1.10	0.69	Very low	—	−1	—	−1	NA	−1	NA
n = 40		(0.55–0.81)	(0.12–0.49)	(0.71–1.30)	(0.53–2.32)				(9)					
	crepitation/deflection/ pain/limited mouth opening	0.27 (0.16–0.41)	0.88 (0.69–0.97)	2.27* (0.72–7.15)	0.83 (0.67–1.03)	0.69								
Crepitation vs. MRI														
Taşkaya-Yýlmaz and Oğütcen-Toller[20]	Crepitus vs folded disc on MRI	0.76 (0.56–0.90)	0.78 (0.68–0.85)	NE	NE	0.23	Low	—	−1 (10)	—	−1	NA	−1	NA
n = 73														
Uşümez et al.[19]	Crepitation for IDR	0.11	0.88	0.88	1.02	0.60	Very low	—	−1	—	−1	NA	−1	NA
n = 40		(0.03–0.25)	(0.69–0.97)	(0.21–3.59)	(0.85–1.22)				(9)					
	Crepitation for IDnoR	0.71 (0.44–0.90)	0.88 (0.69–0.97)	5.88‡ (1.95–17.76)	0.33* (0.16–0.71)	0.40								
	Crepitation for ID	0.29 (0.18–0.43)	0.88 (0.69–0.97)	2.42* (0.78–7.57)	0.81 (0.64–1.01)	0.69								

(Continued)

Table 2 *(Continued)*

		Summary of Findings					Cochrane GRADE Domains							
Study	Tests	Specificity (interval 95%)	Sensitivity (interval 95%)	+LR (interval 95%)	−LR (interval 95%)	Pre-test Probability	Quality of Evidence (GRADE)	Study Design (follow-up period)	Limitations (QUADAS Total Score)	Out-come	Popula-tions, Tests, Indirect Com-parisons	Inconsis-tency (I^2 *p*-value)	Impreci-sion (Sparse Data; Group Size)	High prob-ability of Publica-tion Bias
Deviation vs. MRI														
Uşümez et al.[19]	Deviation for IDR	0.92	0.08	1.00	0.99	0.60	Very low	—	−1	—	−1	NA	−1	NA
$n = 40$		(0.79–0.98)	(0.01–0.26)	(0.86–1.16)	(0.18–5.49)				(9)					
	Deviation for IDnoR	0.35	0.08	0.38	8.09	0.40								
		(0.14–0.62)	(0.01–0.26)	(0.20–0.74)	(2.05–31.99)									
	Deviation for ID	0.75	0.08	0.81	3.18	0.69								
		(0.61–0.85)	(0.01–0.26)	(0.67–0.98)	(0.78–12.96)									
Deflection vs. MRI														
Uşümez et al.[19]	Deflection for IDR	0.11	0.88	0.88	1.02	0.60	Very low	—	−1	—	−1	NA	−1	NA
$n = 40$		(0.03–0.25)	(0.69–0.97)	(0.21–3.59)	(0.85–1.22)				(9)					
	Deflection for IDnoR	0.76	0.88	6.37‡	0.27*	0.40								
		(0.50–0.93)	(0.69–0.97)	(2.13–19.03)	(0.11–0.64)									
	Deflection for ID	0.31	0.88	2.58*	0.79	0.69								
		(0.19–0.45)	(0.69–0.97)	(0.83–8.00)	(0.62–0.99)									
Limited mouth opening vs. MRI														
Uşümez et al.[19]	Limited mouth	0.11	0.84	0.66	1.07	0.60	Very low	—	−1	—	−1	NA	−1	NA
$n = 40$	opening for IDR	(0.03–0.25)	(0.64–0.95)	(0.18–2.39)	(0.87–1.30)				(9)					
	Limited mouth	0.76	0.84	4.78*	0.28	0.40								
	opening for IDnoR	(0.50–0.93)	(0.64–0.95)	(1.87–12.19)	(0.12–0.67)									
	Limited mouth	0.31	0.84	1.93	0.82	0.69								
	opening for ID	(0.19–0.45)	(0.64–0.95)	(0.72–5.15)	(0.64–1.05)									

Clinical value of likelihood ratio (see Table 3 for definition): *Small/sometimes important, ‡Moderate/usually important, †Pooled data

+LR = positive likelihood ratio; −LR = negative likelihood ratio; GRADE = Grading of Recommendations, Assessment, Development, and Evaluation (ratings: high, moderate, low, very low; coding GRADE domains: −1 if domain not met;—if domain is met; NA if domain is not applicable); QUADAS = Quality Assessment of studies of Diagnostic Accuracy included in Systematic reviews (total score: 0–14); IDR = internal derangement with reduction; RC = classical reciprocal click; ARC = atypical reciprocal click: opening click separated from closing click by less than 5 mm; SC = single opening or closing click accompanied by excursion and/or protrusion click; IDnoR = internal derangement without reduction; ID = internal derangement of any kind; NE = not estimable; TMJ = temporomandibular joint; ROM = range of motion.

Table 3 Study Demographics (Prospective Cohort Studies)

Study	Sample population				Clinical test; single/cluster	MRI type; plane	Referral source	Quality; QUADAS score (max 14)
	n	Mean age (range), y*	No. of patients: Female	No. of patients: Male				
Barclay et al.[17]	40	34.9 (21–68)	35	5	Single Cluster	Sagittal & coronal	Tertiary	12
Bertram et al.[21]	131	36.4 (14–79)	112	19	Cluster	Sagittal & coronal	Tertiary	12
Emshoff et al.[22]	194	36 (17–79)	152	42	Single	Sagittal & coronal	Tertiary	14
Manfredini et al.[18]	194	55.3 (18–72)	153	41	Single	Sagittal & coronal	Tertiary	14
Marguelles-Bonnet et al.[23]	242	26.4	198	44	Cluster	Sagittal & coronal	Primary	11
Rudisch et al.[1]	41	39.1 (17–78)	32	9	Cluster	Sagittal & coronal	Tertiary	14
Taşkaya-Yýlmaz and Oğütcen-Toller[20]	70	not reported	53	17	Single	Sagittal & coronal	Unknown	10
Uşümez et al.[19]	40	32.6 (SD 9.3)	27	13	Cluster	Sagittal	Tertiary	9

*Unless otherwise indicated
QUADAS = Quality Assessment of studies of Diagnostic Accuracy included in Systematic reviews.

(CI) was applied. Accuracy can be expressed in terms of sensitivity (the proportion of positive test findings among people with the disorder) and specificity (the proportion of negative test results among people without the disorder) or by LRs; LRs are the most clinically helpful in reporting diagnostic accuracy of clinical tests. The LR incorporates the sensitivity and specificity of a test into a single measure and is independent of the prevalence of the disorder within a given population. A positive LR (+LR) indicates how much more likely a positive test result is in people who have the disorder than in those who do not; the ideal test for *ruling in* a disorder is the one with the largest +LR. A negative LR (−LR) indicates how much more likely a negative test result is in people without the disorder than in people with it; the best test for *ruling out* a disorder is the one with the smallest −LR. Values in Table 4 show the clinical application of LRs in assessing a shift in the probability of the disorder's being present; that is, a $+LR > 10$ indicates a large and often conclusive probability that the condition is present, while a $-LR < 0.10$ suggests a large and often conclusive probability that the condition is not present.[26] A +LR or −LR of 1 means that a positive or negative result, respectively, is equally probable in a participant with and a participant without the disorder. A CI including an LR of 1 should therefore be interpreted with caution. The DerSimonian–Laird method was chosen for estimating the pooled likelihood ratios (LR_p), using a random-effects model.[27]

When studies were deemed clinically similar enough for pooling of results, the heterogeneity of studies was investigated first graphically and then statistically. Cochran's Q-test was calculated, and a p-value ≤ 0.10 was considered to indicate statistical heterogeneity.[28,29] Inconsistency was measured using I^2, a method of quantifying heterogeneity in a meta-analysis.[30] An I^2 of 0% for a trial indicates that all variability in effect estimates is due to sampling error within trials and none is due to heterogeneity; an I^2 of 40% indicates that 40% of variability between the trials can be attributed to study variation, and is considered mildly to moderately inconsistent.

Sub-group analysis was planned a priori for three factors: (1) methodological quality (poorer trial quality indicated by a QUADAS score $\leq 9/14$); (2) pre-test probability levels, that is, the type of population, from general to tertiary care (usually low in general care, intermediate

Table 4 Likelihood Ratios and Clinical Values

Positive likelihood ratio (+LR): sensitivity / (1 - specificity)	Negative likelihood ratio (−LR): (1 - sensitivity) / specificity	Shift in probability when the condition is present
>10	<0.1	Significant/large, often conclusive that condition is present Significant/small, often conclusive that condition is absent
5–10	0.1–0.2	Moderate / usually important
2–5	0.2–0.5	Small / sometimes important
1–2	0.5–1	Very small / rarely important

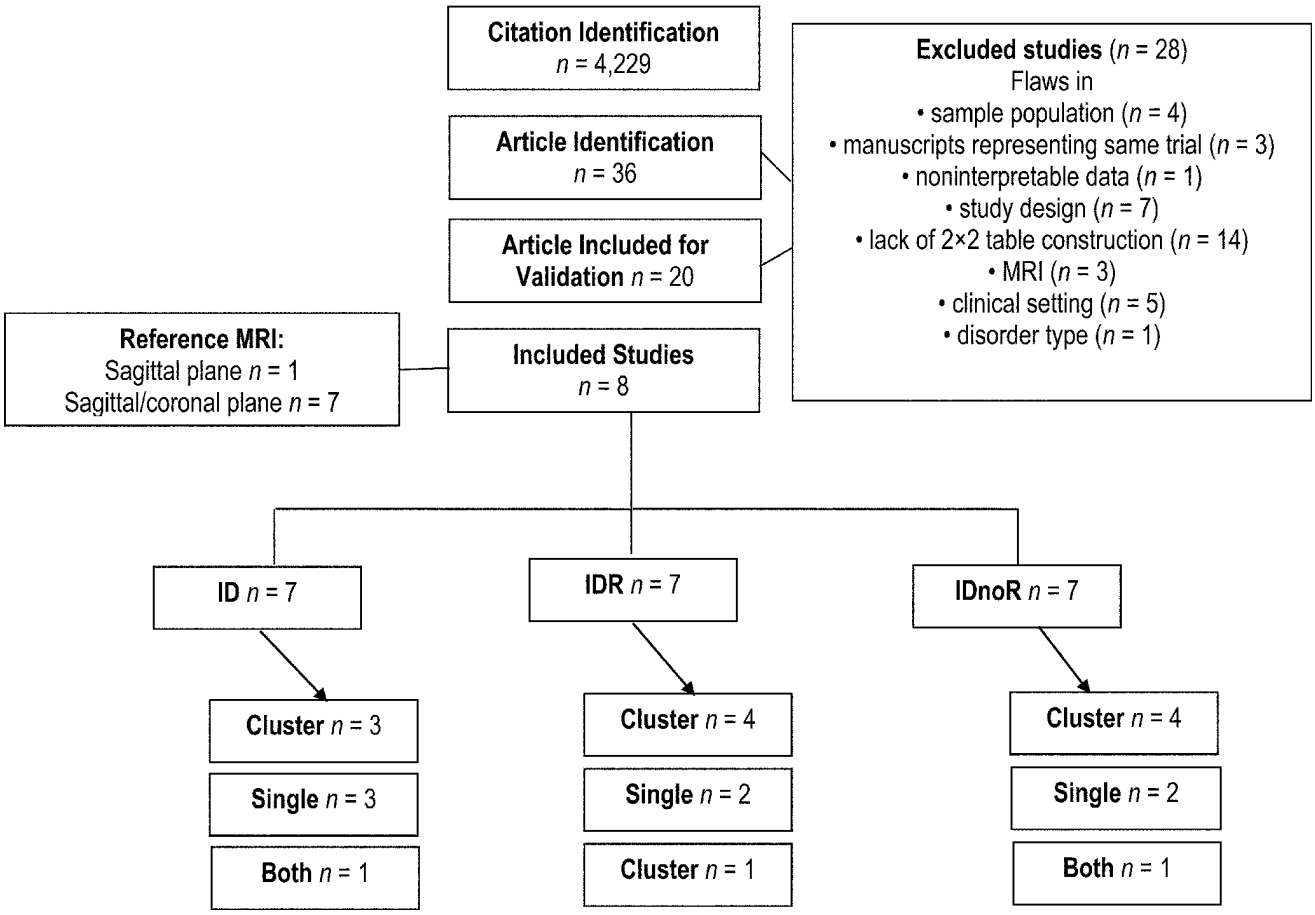

Figure 1 Study-selection flowchart
ID = internal derangement; IDR = internal derangement with reduction; IDnoR = internal derangement without reduction.

in secondary care, and high in tertiary care); and (3) publication bias. This analysis was not conducted, however, because we did not have sufficient data for a meta-regression to be useful.[31]

RESULTS

Selection

Our database search identified 4,229 citation postings between January 1994 and October 2009 (see Figure 1). Based on relevant abstracts and titles, 36 articles were retrieved for full-text screening. The final results were based on 8 primary studies (κw 0.66, SD = 0.24). Studies were excluded for reasons relating to sample population ($n = 4$), clinical testing ($n = 5$), MRI ($n = 3$), 2×2 table construction ($n = 14$), disorder type ($n = 1$), and study design ($n = 7$).

Study Descriptions

A single study[19] examined ID, IDR, and IDnoR from the sagittal plane only, using a cluster of clinical tests (see Table 2). The other seven studies examined the sagittal and coronal views.[1,17,18,20–23] Of the eight selected articles,[1,17–23] seven examined ID[1,17–22] (cluster $n = 3$, single $n = 3$, both $n = 1$) and seven examined both IDR and IDnoR[1,17–21,23] (cluster $n = 4$, single $n = 2$, both $n = 1$). All measures within these clusters must be scored positive for the diagnosis to be made, although reporting of the scoring method was not consistently detailed or transparent. One study was based on primary referral and six on tertiary referral; the remaining study was unclear on this point (see Table 3).

Methodological Quality

All articles selected were of high quality (≥9/14 QUADAS score; see Table 1). Each study consistently administered both the clinical test(s) and MRI to all participants. In addition, the studies explained both reference and clinical test(s) in depth, which ensures reproducibility. Two factors determined poorer quality. First, five studies (62.5%) did not specify the duration between MRI and administration of clinical test(s), which allows for the possibility that natural history changed the condition. Second, blinding of evaluators to the MRI or clinical test(s) was not consistently reported: three studies did not report any blinding, and three studies reported

single blinding. Lack of blinding can bias the evaluator's interpretations of the data.

Clicks

We identified four trials[17–20] that explored the presence or absence of a click through palpation for various types of ID. Although these trials appear to have had the same type of clinical setting, used the same reference tests and clinical tests, and had similar population characteristics, they were heterogeneous in terms of pre-test probability. This suggests that even though the participants were from a tertiary-care population they may have varied along the spectrum of the disorder. We therefore did not pool the results.

There is very low to moderate-quality evidence that the presence of any click is of rare importance in identifying ID (4 trials,[17–20] 369 participants), IDR, or IDnoR (3 trials,[17–19] 369 participants). Similar findings exist for atypical reciprocal click (ARC, opening click separated from closing click by less than 5 mm) and single click (SC, opening or closing click accompanied by excursion and protrusion click).

There is low-quality evidence (1 trial,[17] 39 participants) suggesting that a reciprocal click (RC) has a small and sometimes important probability in ruling in or ruling out IDR (+LR: 3.24 [95% CI, 0.90–11.67]; −LR: 0.50 [95% CI, 0.31–0.80]), IDnoR (+LR: 2.14 [95% CI, 0.54–8.51]), and ID (+LR: 2.89 [95% CI, 0.81–10.39]). The −LRs have very little importance for IDnoR and ID, and are therefore not reported.

Pain

We identified five trials[1,19–22] that explored the presence of pain on palpation, functioning, and opening in various types of ID. Two studies[1,21] had similar population characteristics and pre-test probabilities and were not statistically heterogeneous (for IDR, $I^2 = 0\%$; $p = 0.34$–0.38; for IDnoR, $I^2 = 0\%$; $p = 0.49$–0.68; for ID, $I^2 = 0$–48%; $p = 0.14$–0.39). Pooling of their results for further analysis was therefore possible; Figures 2–4 illustrate the results. The remaining three studies[19,20,22] were heterogeneous because of the differences in how their diagnostic test(s) were defined, which prevented any further meta-analysis. One article[20] did not present sufficient data for us to summarize its findings.

There is moderate-quality evidence (2 trials,[1,21] 172 participants) suggesting that pain on palpation, function, or opening has a very small and rarely important probability of identifying the presence of IDR ($+LR_p$: 1.44 [95% CI, 1.02–2.04]; $-LR_p$: 0.80 [95% CI, 0.63–−1.00]). The same 2 trials[1,21] (172 participants) offer moderate-quality evidence suggesting that pain on palpation, function, or opening is of small importance and can sometimes assist in identifying IDnoR ($+LR_p$: 2.11 [95% CI, 1.58–2.82]; $-LR_p$: 0.47 [95% CI, 0.36–0.62]). There is also moderate-quality evidence from three trials[1,21,22] (366 participants) that pain on palpation, function, or opening is of very small importance and can rarely assist in identifying patients with any ID ($+LR_p$: 1.65 [95% CI, 1.36–2.01]; $-LR_p$: 0.66 [95% CI, 0.54–0.80]).

Diagnostic Test Cluster vs. MRI

We identified three trials[17,19,23] that investigated the use of a cluster of tests including ROM, muscle palpation, joint palpation, and joint sounds for ID. All tests in this cluster need to be positive to identify ID. Although the clinical tests performed were similar, heterogeneity in diagnostic criteria, reference standard (MRI of sagittal vs. coronal view), referral source, and pre-test probability precluded pooling of results, with the exception of IDR and IDnoR findings for two of the articles.[17,23]

There is low-quality evidence (2 trials,[17,23] 281 participants) to suggest that this cluster of tests is a small and sometimes important indicator in identifying patients with IDR ($+LR_p$: 2.97 [95% CI, 2.36–3.74]; $-LR_p$: 0.42 [95% CI, 0.28–0.61]). Although the data showed some variation in pre-test probability, they were deemed clinically similar enough to pool and were not statistically heterogeneous ($I^2 = 0$–34%; $p = 0.21$–0.74). In addition, there is low-quality evidence (2 trials,[17,23] 281 participants) suggesting that the same cluster of tests is a small and sometimes important indicator for ruling in IDnoR in the presence of a positive cluster of tests ($+LR_p$: 4.13 [95% CI, 3.02–5.64]) and has a very small and rarely important probability of ruling out IDnoR in the presence of a negative cluster of tests ($-LR_p$: 0.66 [95% CI, 0.37–1.16]). Although these trials were statistically heterogeneous ($I^2 = 0$–90%; $p = 0.00$–0.89), we decided to pool them because they were deemed to be clinically similar.

One trial[19] with 40 participants examined the effectiveness of clusters of tests—including (a) click, deviation, and pain, and (b) crepitation, deflection, pain, and limited mouth opening—in determining the presence of ID. This trial yielded very low quality evidence that test cluster (a) presence of click, deviation, and pain is often conclusive in ruling out IDR in the presence of a negative cluster of tests (−LR: 0.09 [95% CI, 0.01–0.72]). In addition, there was very low quality evidence that test cluster (b) crepitation, deflection, pain, and limited mouth opening is a moderate and usually important indicator for ruling in IDnoR in the presence of a positive test (+LR: 6.37 [95% CI, 2.13–19.03]) and a small and sometimes important indicator for ruling out IDnoR in the presence of a negative test (−LR: 0.27 [95% CI, 0.11–0.64]). A cluster of tests appears to be better at ruling a condition in or out than a single test.

Furthermore, there was very low quality evidence from this same trial suggesting that another test cluster (c) crepitation, deflection, pain, and limited mouth opening has a small and sometimes important probability of ruling in ID of any kind in the presence of a positive cluster of tests [(+LR: 2.27 [95% CI, 0.72–7.15]) and a very small and rarely important probability of ruling out

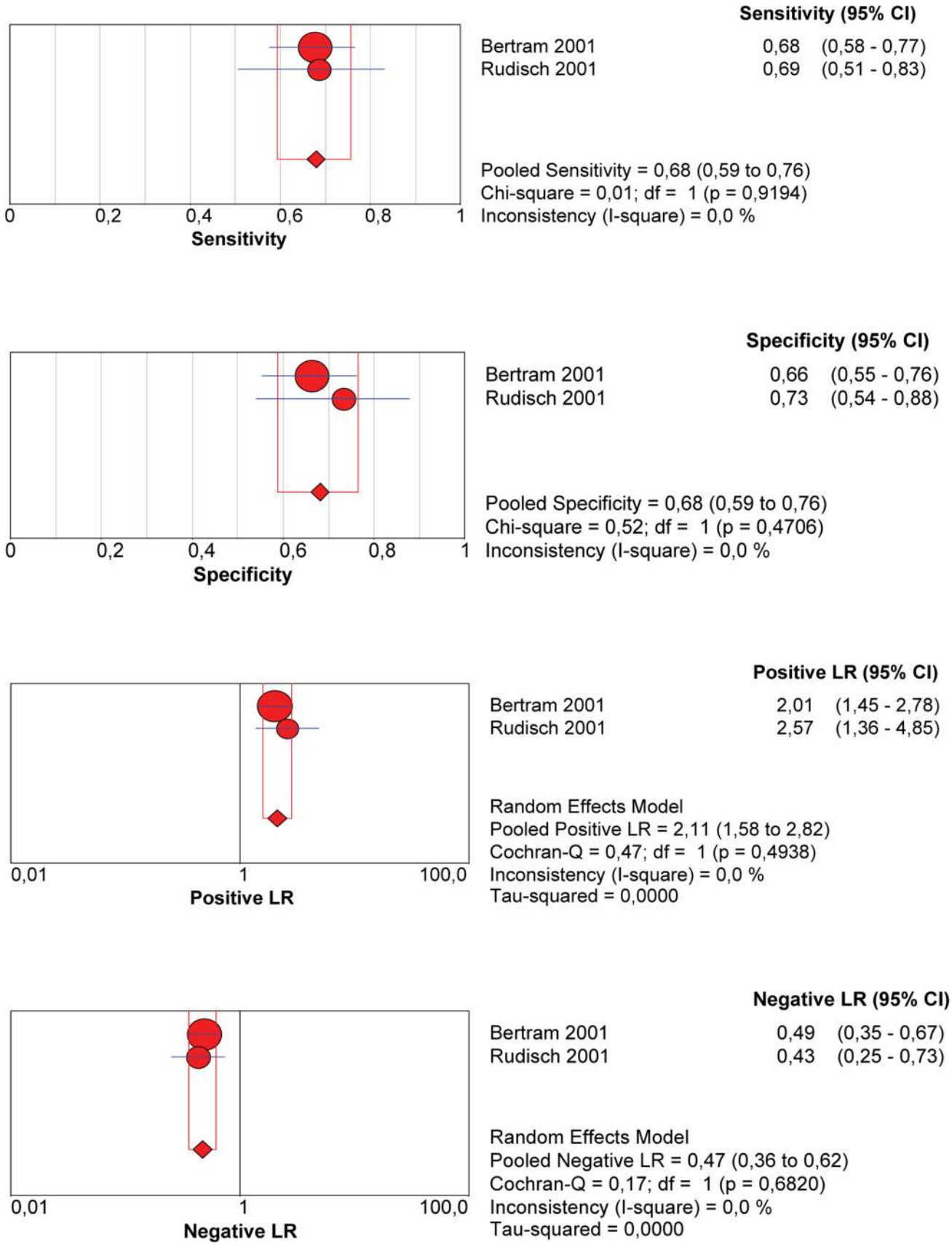

Figure 2 Pain in TMJ on palpation, function, opening vs. MRI for internal derangement without reduction (IDnoR).

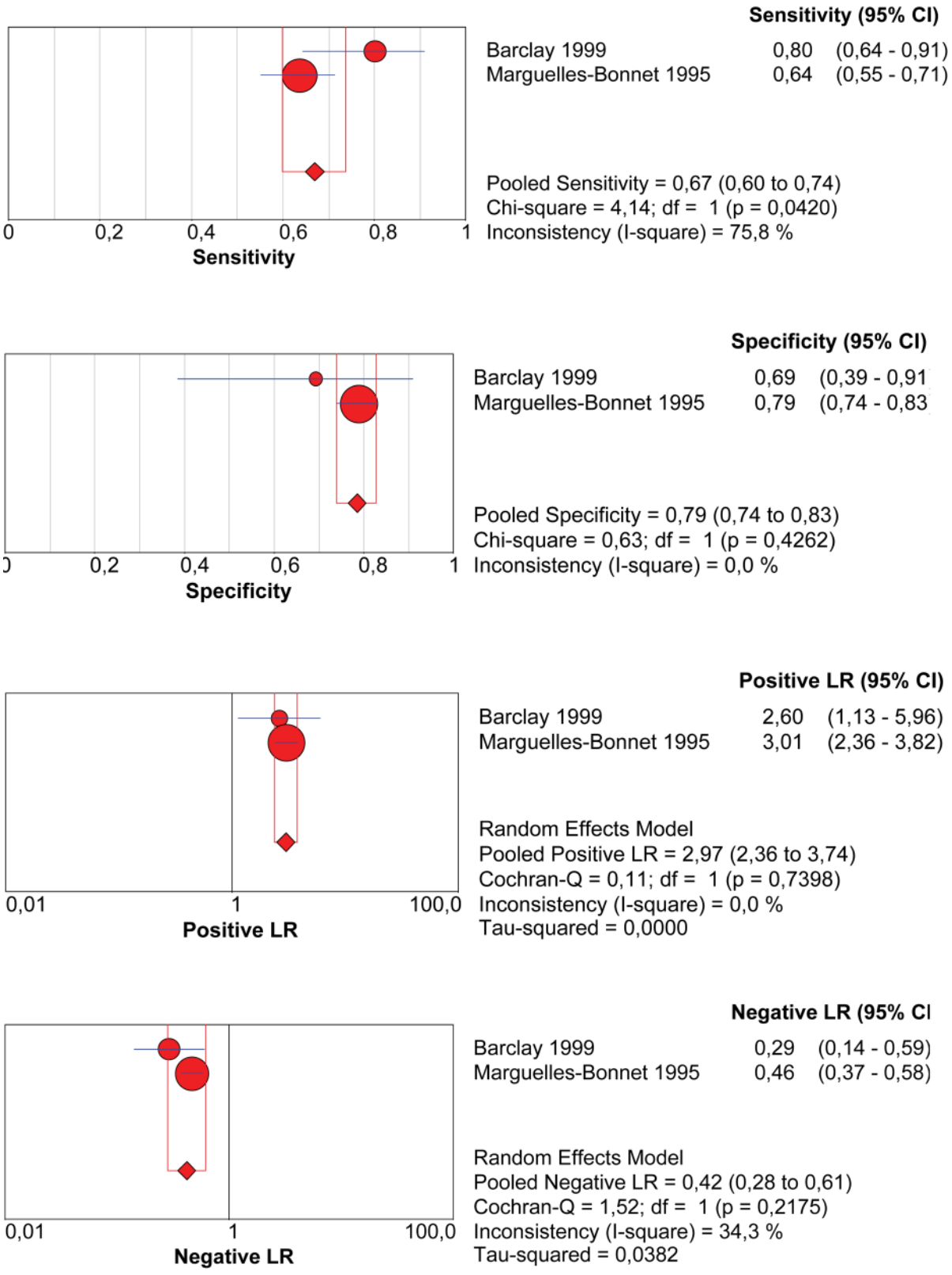

Figure 3 Cluster (ROM / muscle–joint palpation / joint sounds) vs MRI for internal derangement with reduction (IDR).

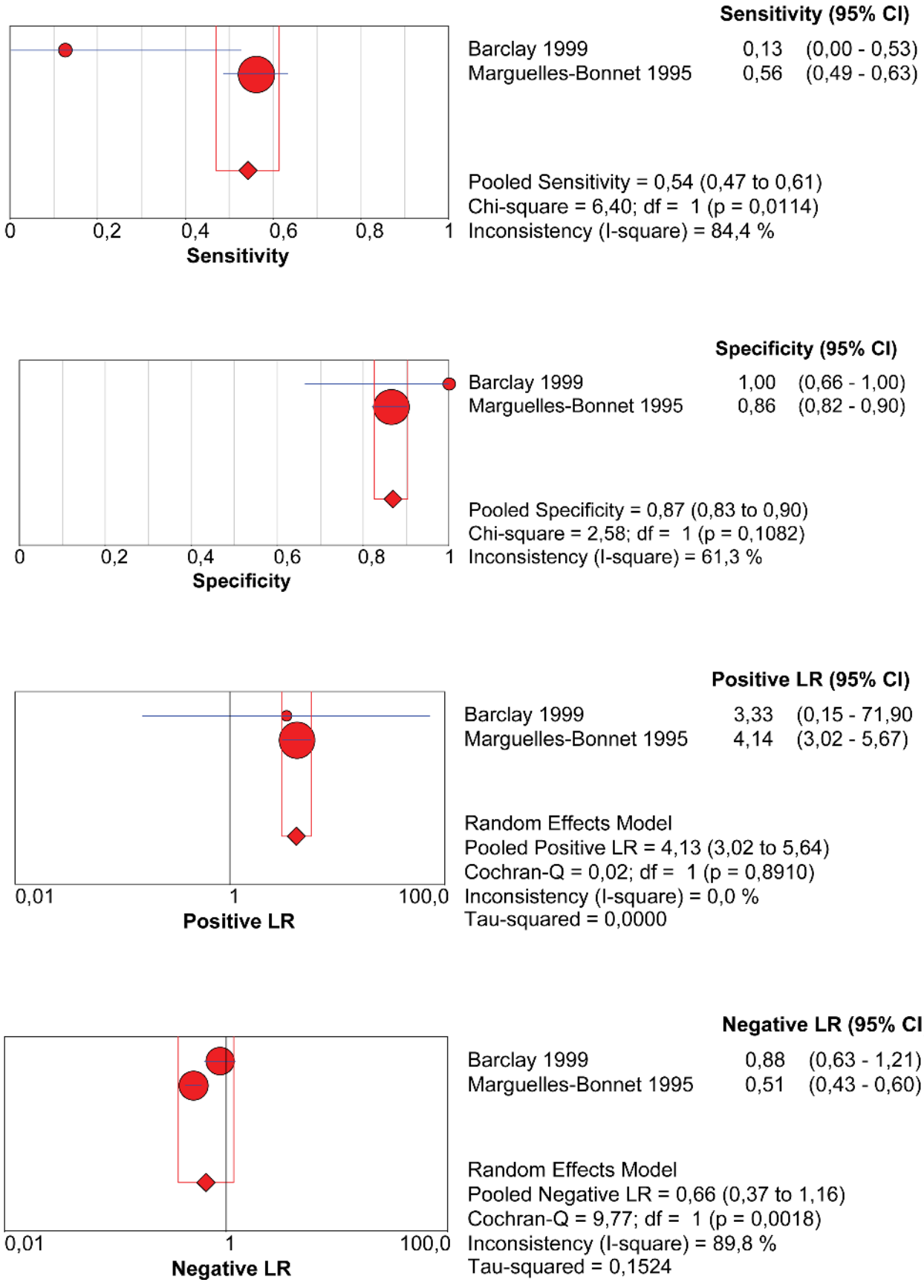

Figure 4 Cluster (ROM / muscle–joint palpation / joint sounds) vs MRI for internal derangement without reduction (IDnoR).

ID of any kind in the presence of a negative cluster of tests (−LR: 0.83 [95% CI, 0.67–1.03]).

Other Tests vs. MRI

Heterogeneity precluded meta-analysis of trials (2 trials,[19,20] 113 participants) indicating crepitation as a test in identifying participants suffering from ID. One trial[19] with 40 participants yielded very low quality evidence suggesting that the sign of crepitation has a moderate and usually important probability that IDnoR is present (+LR: 5.88 [95% CI, 1.95–17.76]). In addition, the absence of crepitation has a small and sometimes important probability of ruling out IDnoR (−LR: 0.33 [95% CI, 0.16–0.71]) in the presence of a negative test finding.

There was very low quality evidence (1 trial,[19] 40 participants) that the sign of crepitation is a small and sometimes important indicator for ruling in ID (+LR: 2.42 [95% CI, 0.78–7.57]). Other clinical tests were suggestive of ID. The presentation of deviation is a very small and rarely important indicator of IDR (+LR: 1.00 [95% CI, 0.86–1.16]), as is the absence of deviation (−LR: 0.99 [95% CI, 0.18–5.49]). LR values for deviation as a predictor for or against IDnoR or for ID were not significant.

Deflection as a positive indicator for IDnoR is considered a moderate and usually important sign (+LR: 6.37 [95% CI, 2.13–19.03]), and deflection as a negative indicator of IDnoR is considered a small and sometimes important sign (−LR: 0.27 [95% CI, 0.11–0.64]). The presence of deflection is a small and sometimes important sign to rule in ID of any kind (+LR: 2.58 [95% CI, 0.83–8.00]).

DISCUSSION

Clinical Relevance of Findings

The goal of this review was to assess the diagnostic validity of clinical tests for TMD classified as IDR or IDnoR relative to MRI. The review provides no evidence to support any one clinical test as a significant and conclusive predictor of the presence or absence of ID relative to MRI; however, there is evidence that certain tests may be of some relevance in helping to diagnose TMD.

In reviewing the extracted data, we decided to base our recommendations on the LRs of the diagnostic tests rather than on their specificity and sensitivity. Although specificity and sensitivity provide useful information, they cannot be used to quantify the shift in probability of the condition given a certain test result.[32] By contrast, the LR combines the information contained in sensitivity and specificity with the pre-test probabilities to determine the shift in probability based on the results of the diagnostic test.

Interpretation of Results

Although there was no high-quality evidence showing strong and conclusive probability of clinical tests' diagnosing the presence or absence of IDR, IDnoR, or ID, there are some tests with moderate and others with small diagnostic properties, as follows (see Table 5):

1. Tests with *moderate probability of ruling in the diagnosis IDnoR* are (a) crepitation, deflection, pain, and limited mouth opening;[19] (b) crepitation;[19] and (c) deflection.[19] The same study also found a *significant, often conclusive probability* that if one click with deviation and pain are absent, then IDR can be ruled out.
2. Tests with *small or sometimes important probability of ruling in the diagnosis ID* are (a) any click;[17] (b) RC;[17] (c) decreased ROM, pain with muscle or joint palpation, and joint sounds;[17] (d) pain on palpation, function, and opening;[1] and (e) crepitation, deflection, pain, and limited mouth opening.[19] The latter study also found that simple crepitation[19] and deflection[19] have small and sometimes important probability as stand-alone tests to rule in ID. ID can be ruled out if any click[17] is absent or if the test cluster (a) ROM and pain with muscle palpation or (b) joint palpation and joint sounds[17] is negative.
3. Test clusters with *small or sometimes important probability of ruling in the diagnosis IDnoR* are (a) crepitation, deflection, pain, and limited mouth opening;[19] (b) pain in TMJ on palpation, function, and opening;[1,21] (c) ROM, muscle–joint palpation, and joint sounds;[17,23] and (d) RC in addition to limited mouth opening.[17] IDnoR can be ruled out if (a) pain in TMJ on palpation, function, and opening;[1,21] (b) crepitation, deflection, pain, and limited mouth opening;[19] (c) crepitation;[19] and (d) deflection[19] are absent.
4. Tests with *small or sometimes important probability of ruling in the diagnosis IDR* are (a) ROM, muscle–joint palpation, and joint sounds[17,23] and (b) RC.[17] The absence of RC[17] is a stand-alone test that helps to rule out IDR.

Our literature search did not find any other current systematic reviews to which we could compare our results. One unpublished thesis[2] reviewed clinical tests compared to several other reference standards. The author determined that one clinical classification scheme (*IDR*: RC, no coarse crepitus, open $\geq$35 mm; *IDnoR*: history of movement limitation, no RC, no coarse crepitus, maximum opening $\leq$35 mm, passive opening stretch $<$40 mm, contralateral movement $<$7 mm, no SD) had high discriminative power ($n = 1$, pre-test probability: 0.85; sensitivity: 0.97 [95% CI, 0.82–1.00]; specificity: 1.00 [95% CI, unable to calculate]); however, the reference standard was arthrotomography, and the methodological

Table 5 Summary of Significant Findings

Probability that the condition is present according to LR	Type of ID	Authors	Tests	+LR (95% CI)	−LR (95% CI)	GRADE
Significant / large / often conclusive	IDR	Uşümez et al.[19]	Click / deviation / pain		0.09 (0.01–0.72)	Very low
Moderate / usually important	IDnoR	Uşümez et al.[19]	Crepitation / deflection / pain / limited mouth opening	6.37 (2.13–19.03)		Very low
			Crepitation	5.88 (1.95–17.76)		
			Deflection	6.37 (2.13–19.03)		
Small / sometimes important	ID	Barclay et al.[17]	All clicks (palpation)	2.03 (0.91–4.50)	0.41 (0.22–0.77)	Low
			RC (palpation)	2.89 (0.81–10.39)		
			Cluster: ROM / muscle–joint palpation / joint sounds	2.44 (1.06–5.58)	0.36 (0.20–0.64)	
		Rudisch et al.[1]	Pain in TMJ on palpation / function/opening	2.38 (1.27–4.46)		Moderate
		Uşümez et al.[19]	Cluster: crepitation / deflection / pain / limited mouth opening	2.27 (0.72–7.15)		Very low
			Crepitation	2.42 (0.78–7.57)		
			Deflection	2.58 (0.83–8.00)		
	IDR	Barclay et al.	RC (palpation)	3.24 (0.90–11.67)	0.50 (0.31–0.80)	Low
		Pooled: Barclay et al.,[17] Marguelles-Bonnet et al.[23]	Cluster: ROM / muscle–joint palpation / joint sounds	2.97 (2.36–3.74)	0.42 (0.28–0.61)	Low
	IDnoR	Barclay et al.	RC (palpation)	2.14 (0.54–8.51)		Low
		Bertram et al.[21]	Pain in TMJ on palpation / function / opening	2.01 (1.45–2.78)	0.49 (0.25–0.67)	Moderate
		Rudisch et al.[1]	Pain in TMJ on palpation / function / opening	2.57 (1.36–4.85)	0.43 (0.25–0.73)	Moderate
		Pooled: Bertram et al.,[21] Rudisch et al.[1]	Pain in TMJ on palpation / function / opening	2.11 (1.58–2.82)	0.47 (0.36–0.62)	Moderate
		Uşümez et al.[19]	Cluster: crepitation / deflection / pain / limited mouth opening		0.27 (0.11–0.64)	Very low
		Pooled: Barclay et al.,[17] Marguelles-Bonnet et al.[23]	Cluster: ROM / muscle–joint palpation / joint sounds	4.13 (3.02–5.64)		Low
		Uşümez et al.[19]	Crepitation		0.33 (0.16–0.71)	Very low
			Deflection		0.27 (0.11–0.64)	
			Limited mouth opening	4.78 (1.87–12.19)	0.28 (0.12–0.67)	

ID = internal derangement of any kind; +LR = positive likelihood ratio; −LR = negative likelihood ratio; GRADE = Grading of Recommendations, Assessment, Development, and Evaluation (ratings: high, moderate, low, very low); IDR = internal derangement with reduction; IDnoR = internal derangement without reduction; ROM = range of motion; RC = classical reciprocal click.

quality of the study was low, not unlike the clusters determined in this review.

LIMITATIONS

Despite efforts to limit bias and ensure good methodological quality, this study has several potential limitations. A comprehensive search for relevant literature was conducted using two online databases (MEDLINE and Embase) to retrieve published literature; our methods did not include a search of the grey literature (e.g., bibliographical lists, unpublished data), which may have introduced selection bias as this potentially relevant information may not have been found. In addition, publication bias may have influenced our results, as only English-language studies were reviewed.[33] Although sensitivity analysis for publication bias was planned, heterogeneity

of studies precluded pooling of results across more than three trials, and funnel plots were therefore not constructed.

Meta-analysis of the data was limited by several issues. Although study participants were mostly examined at tertiary centres, a few studies included patients from primary-care facilities. In addition, clinical tests of the patient population often had dissimilar pre-test probabilities, despite the participants' being from the same referral source (either primary or tertiary). This would lead us to believe that patients were presenting with different stages of the disease. Therefore, despite the similarity of study characteristics, we did not pool data from studies whose pre-test probabilities were substantially different. Large differences in pre-test probabilities also raise concerns about the presence of spectrum bias, which occurs when diagnostic testing is done on individuals who are not representative of the population to which the test is usually applied in practice.[32] Review of the included articles revealed, however, that all were prospective cohort studies with consecutive groups of participants from a clinical population, a factor essential to minimizing spectrum bias;[32] in addition, all studies included symptomatic patients only. The likelihood of spectrum bias was therefore determined to be minimal.

Another issue was diagnostic test definition, which made meta-analysis sometimes impossible. The test identified as ''pain'' was variously defined as pain on palpation, function, and opening; pain on palpation; or even self-reported ear or TMJ pain. ''Click'' was also assessed by different techniques: an audible click, a palpated click, or a click evaluated with a stethoscope. In addition, clusters of tests did not always refer to the same tests. One study,[21] for example, grouped different tests after a mathematical evaluation of the LR. Other authors followed the RDC/TMD (Research Diagnostic Criteria for Temporomandibular Disorders) protocol.[34]

MRI has been shown to be a reliable and valid method for diagnosing ID of the TMJ.[35] Studies that satisfied our inclusion criteria all used MRI as the gold standard. However, one study[19] used only sagittal images for diagnosis of ID, whereas the other seven studies used both sagittal and coronal images. It has been recommended that both sagittal and coronal images be obtained for optimal MRI assessment of the TMJ.[36] Thus, conclusions drawn from the findings of the aforementioned article[19] should be implemented with caution in clinical practice. Although this study did suggest that cluster, crepitation, and deflection could have important predictability in determining ID, methodological quality was deemed to be very low for several reasons, including unclear blinding procedures, small sample size, use of sagittal MRI only, and unclear time period between index and reference test.

Suggestions for Future Research

In designing future diagnostic studies, researchers should focus special consideration on five critical quality domains to avoid systematic error: adequate description of patient population; blinded assessment; a clear definition of the diagnostic test; clear reporting of the duration between clinical and reference tests; and use of both sagittal and coronal MRI views as the reference standard, to ensure higher accuracy and consistency across reference tests. To reduce the risk of random error from imprecision, estimates should be based on an adequate sample size.

CONCLUSION

Our review indicates that there is a lack of high-quality research examining the accuracy of clinical tests in determining ID. Furthermore, the majority of the clinical tests studied were of little importance in determining the presence or absence of ID. Although some tests display a higher LR, and can therefore be considered to have moderate and usually important probability in determining the presence of ID (and one of significant and large importance in ruling out IDR), the quality of evidence was very low. In other words, the evidence for the test that was of usual or large importance to the diagnosis of ID comes from a single small study with moderate or high risk of bias. However, taking into consideration lower-quality evidence and other limitations, as outlined above, we can draw several conclusions. There is no single or cluster of tests of moderate or high quality that has more than small importance in determining IDR. Contrary to clinical reports, click does not demonstrate significant importance in ruling in IDR. The most important findings in ruling in IDR, according to our review, are (a) a test cluster including ROM, muscle–joint palpation, and joint sounds (1 meta-analysis); and (b) reciprocal click as a single test (1 study). Despite their small importance, there is moderate-quality evidence to suggest that both are effective in ruling in IDR.

Deflection and crepitation are the most valuable single tests (1 study), and the test cluster including crepitation, deflection, pain, and limited mouth opening (1 study) is most valuable to rule in IDnoR, while the test cluster click, deviation, and pain helps to rule out IDnoR in the absence of a positive test. However, the evidence to support this statement is of very low quality (1 study) and needs replication. Conversely, there is moderate-quality evidence (1 meta-analysis), with low clinical importance, of pain in TMJ on palpation, opening, and function, and low-quality evidence (1 meta-analysis) that a test cluster of ROM, muscle–joint palpation, and joint sound is helpful in ruling in IDnoR.

KEY MESSAGES

What Is Already Known on This Topic

Internal derangement (ID) is one of the most common causes of TMD. As a result, the diagnostic validity of tests such as pain on palpation, click, deflection, limited mouth opening, and crepitation to identify ID has been studied in the literature. Clusters of tests have also been examined for their diagnostic accuracy. However, methodological quality concerns have been identified that call into question the clinical effectiveness of implementing these tests in practice. Both click and pain have been suggested as good indicators of ID, but no systematic review of the diagnostic accuracy of clinical tests has been published.

What This Study Adds

This systematic review summarizes the results of the current literature on the diagnostic validity of tests to identify ID in patients with TMD. Most importantly, careful attention was paid to including studies of sound methodological quality. In addition, the GRADE system was used to assess external validity and determine the strength of recommendations for the diagnostic tests and strategies. The quality of evidence varied from very low to moderate; this review reflects low-quality evidence that deflection and crepitation are the most valuable single tests to determine IDnoR and that crepitation, deflection, pain, and limited mouth opening is the most valuable cluster of tests to determine IDnoR. A negative result for the test cluster click, deviation, and pain is often conclusive that the condition IDR is absent. Despite moderate-quality evidence, no single test (including click) or cluster of tests was of more than small importance in identifying the presence of IDR.

REFERENCES

1. Rudisch A, Innerhofer K, Bertram S, et al. Magnetic resonance imaging findings of internal derangement and effusion in patients with unilateral temporomandibular joint pain. Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 2001;92(5):566–71. doi:10.1067/moe.2001.116817. Medline:11709695
2. Gross AR. The diagnostic validity of clinical tests in temporomandibular disorders: A systematic overview and meta-analysis. [dissertation]. Hamilton (ON): McMaster University; 1994.
3. Rugh JD, Solberg WK. Oral health status in the United States: temporomandibular disorders. J Dent Educ. 1985;49(6):398–406. Medline:3859518
4. Greene CS, Marbach JJ. Epidemiologic studies of mandibular dysfunction: a critical review. J Prosthet Dent. 1982;48(2):184–90. doi:10.1016/0022-3913(82)90110-X. Medline:7050363
5. Rasmussen OC. Description of population and progress of symptoms in a longitudinal study of temporomandibular arthropathy. Scand J Dent Res. 1981;89(2):196–203. Medline:6943666
6. Lobbezoo F, Drangsholt M, Peck C, et al. Topical review: new insights into the pathology and diagnosis of disorders of the temporomandibular joint. J Orofac Pain. 2004;18(3):181–91. Medline:15508997
7. Isong U, Gansky SA, Plesh O. Temporomandibular joint and muscle disorder-type pain in U.S. adults: the National Health Interview Survey. J Orofac Pain. 2008;22(4):317–22. Medline:19090404
8. LeResche L. Epidemiology of temporomandibular disorders: implications for the investigation of etiologic factors. Crit Rev Oral Biol Med. 1997;8(3):291–305. doi:10.1177/10454411970080030401. Medline:9260045
9. Gatchel RJ, Stowell AW, Wildenstein L, et al. Efficacy of an early intervention for patients with acute temporomandibular disorder-related pain: a one-year outcome study. J Am Dent Assoc. 2006;137(3):339–47. Medline:16570467
10. Liedberg J, Westesson PL. Sideways position of the temporomandibular joint disk: coronal cryosectioning of fresh autopsy specimens. Oral Surg Oral Med Oral Pathol. 1988;66(6):644–9. doi:10.1016/0030-4220(88)90309-X. Medline:3205553
11. Schmitter M, Kress B, Rammelsberg P. Temporomandibular joint pathosis in patients with myofascial pain: a comparative analysis of magnetic resonance imaging and a clinical examination based on a specific set of criteria. Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 2004;97(3):318–24. doi:10.1016/j.tripleo.2003.10.009. Medline:15024353
12. Whiting P, Rutjes AWS, Reitsma JB, et al. The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003;3(25):1–13.
13. Atkins D, Best D, Briss PA, et al.; GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ. 2004;328:1–8.
14. Atkins D, Briss PA, Eccles M, et al, and the Grade Working Group. Systems for grading the quality of evidence and the strength of recommendations II: Pilot study of a new system. BMC Health Serv Res. 2005;5(25):1–12.
15. Guyatt GH, Oxman AD, Vist GE, et al, and the GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924–6. doi:10.1136/bmj.39489.470347.AD. Medline:18436948
16. Kavanagh BP. The GRADE system for rating clinical guidelines. PLoS Med. 2009;6(9):e1000094. doi:10.1371/journal.pmed.1000094. Medline:19753107
17. Barclay P, Hollender LG, Maravilla KR, et al. Comparison of clinical and magnetic resonance imaging diagnosis in patients with disk displacement in the temporomandibular joint. Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 1999;88(1):37–43. doi:10.1016/S1079-2104(99)70191-5. Medline:10442943
18. Manfredini D, Basso D, Salmaso L, et al. Temporomandibular joint click sound and magnetic resonance-depicted disk position: which relationship? J Dent. 2008;36(4):256–60. doi:10.1016/j.jdent.2008.01.002. Medline:18276055
19. Üşümez S, Oz F, Güray E. Comparison of clinical and magnetic resonance imaging diagnoses in patients with TMD history. J Oral Rehabil. 2004;31(1):52–6. doi:10.1111/j.1365-2842.2004.01065.x. Medline:15125597
20. Taşkaya-Yýlmaz N, Oğütcen-Toller M. Clinical correlation of MRI findings of internal derangements of the temporomandibular joints. Br J Oral Maxillofac Surg. 2002;40(4):317–21. doi:10.1016/S0266-4356(02)00140-7. Medline:12175833
21. Bertram S, Rudisch A, Innerhofer K, et al. Diagnosing TMJ internal derangement and osteoarthritis with magnetic resonance imaging. J Am Dent Assoc. 2001;132(6):753–61. Medline:11433854
22. Emshoff R, Innerhofer K, Rudisch A, et al. The biological concept of ''internal derangement and osteoarthrosis'': a diagnostic approach in patients with temporomandibular joint pain? Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 2002;93(1):39–44. doi:10.1067/moe.2002.117451. Medline:11805776
23. Marguelles-Bonnet RE, Carpentier P, Yung JP, et al. Clinical diagnosis compared with findings of magnetic resonance imaging in 242 patients with internal derangement of the TMJ. J Orofac Pain. 1995;9(3):244–53. Medline:8995924
24. Zamora J, Abraira V, Muriel A, et al. Meta-DiSc: A software for meta-analysis of test accuracy data. BMC Med Res Methodol [Internet].

2006[cited 2007 June 11];6:31[about 1 p.]. Available from: http://www.hrc.es/investigacion/metadisc_en.htm
25. Cicchetti DV. Assessing inter-rater reliability for rating scales: resolving some basic issues. Br J Psychiatry. 1976;129(5):452–6. doi:10.1192/bjp.129.5.452. Medline:1033010
26. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA. 1994;271(9):703–7. doi:10.1001/jama.271.9.703. Medline:8309035
27. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177–88. doi:10.1016/0197-2456(86)90046-2. Medline:3802833
28. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol. 2005;58(9):882–93. doi:10.1016/j.jclinepi.2005.01.016. Medline:16085191
29. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. Stat Med. 1991;10(11):1665–77. doi:10.1002/sim.4780101105. Medline:1792461
30. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539–58. doi:10.1002/sim.1186. Medline:12111919
31. Cochrane WG. The combination of estimates from different experiments. Biometrics. 1954;10(1):101–29. doi:10.2307/3001666.
32. Fritz JM, Wainner RS. Examining diagnostic tests: an evidence-based perspective. Phys Ther. 2001;81(9):1546–64. Medline:11688591
33. Sousa MR, Ribeiro AL. Systematic review and meta-analysis of diagnostic and prognostic studies: a tutorial. Arq Bras Cardiol. 2009;92(3):229–38, 235–45. Medline:19390713
34. Dworkin SF, LeResche L. Research diagnostic criteria for temporomandibular disorders: review, criteria, examinations and specifications, critique. J Craniomandib Disord. 1992;6(4):301–55. Medline:1298767
35. Haley DP, Schiffman EL, Lindgren BR, et al. The relationship between clinical and MRI findings in patients with unilateral temporomandibular joint pain. J Am Dent Assoc. 2001;132(4):476–81. Medline:11315378
36. Schwaighofer BW, Tanaka TT, Klein MV, et al. MR imaging of the temporomandibular joint: a cadaver study of the value of coronal images. AJR Am J Roentgenol. 1990;154(6):1245–9. Medline:2110737

APPENDIX A: SEARCH STRATEGIES

MEDLINE

((((tmj)) AND ((''diagnosis''[Subheading] OR ''diagnosis''[All Fields] OR ''diagnosis''[MeSH Terms] OR diagnos*))) AND ((mri))) AND ((sensitivity specificity OR validity OR predictive value of tests OR validation))

Embase

1. exp temporomandibular joint/ or exp temporomandibular joint disorder/ or exp temporomandibular ankylosis/ or tmj.mp.
2. mri.mp. or exp nuclear magnetic resonance imaging/
3. ''diagnosis, measurement and analysis''/ or exp clinical assessment tool/ or exp clinical observation/
4. exp differential diagnosis/ or exp diagnosis/ or diagnosis.mp. or exp qualitative diagnosis/ or exp quantitative diagnosis/
5. (sensitivity and specificity).mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer name]
6. predictive value of tests.mp. or exp ''prediction and forecasting''/
7. (validity or validation).mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer name]
8. diagnos*.mp.
9. 8 or 4 or 3
10. 6 or 7 or 5
11. 1 and 10 and 9 and 2
12. limit 11 to (English language and yr = ''2007–2010'')