[LITERATURE REVIEW]

ERIC J. HEGEDUS, PT, DPT, MHSc, OCS, CSCS[1] • CHAD COOK, PT, PhD, MBA, OCS, FAAOMPT[1] • VICTOR HASSELBLAD, PhD[2]
ADAM GOODE, PT, DPT, CSCS[3] • DOUGLAS C. MCCRORY, MD, MHSc[4]

Physical Examination Tests for Assessing a Torn Meniscus in the Knee: A Systematic Review With Meta-analysis

Knee pain has a lifetime prevalence of up to 45% and as many as 31% of individuals with knee pain will consult a general practitioner.[5] Roughly 5% of these individuals will undergo a tibial meniscectomy,[5] and many more will undergo partial meniscectomy or meniscus repair. Surgery of the meniscus is a common orthopedic procedure, constituting 10% to 20% of surgeries performed in some practices.[37] Primary practitioners must make a decision regarding conservative intervention or referral to a specialist for imaging or surgery. This decision is typically made after a thorough history and physical examination. Unfortunately, literature regarding the ability of the comprehensive examination to detect a torn tibial meniscus is equivocal.[13,23,25,26,35,38] Despite frequent reports that items like "locking" and "giving-way" are common with tibial meniscus tears, history alone is insufficient as a diagnostic tool.[7,23,33] Therefore, historical items are often combined with physical examination procedures, such as range-of-motion and strength testing, in an attempt to improve diagnostic accuracy.[13,23,25,26,35,38] Physical diagnostic tests, sometimes referred to as "special tests," have been an integral part of this process historically. There are numerous special tests purported to diagnose torn tibial menisci, including traditional non–weight-bearing tests like McMurray's[33] test and Apley's[4] test, and newer weight-bearing tests like the Thessaly test.[24] With the introduction of new tests and the continued evaluation of traditional tests, there is value in examining the body of work regarding special tests of the tibial menisci and the ability of those tests to discriminate between patients with and without a torn meniscus. This body of evidence has been examined previously,[22,41,45] with the most extensive review of individual special tests for a torn meniscus being published in 2001.[41] Since the Scholten et

• **STUDY DESIGN:** Systematic review and meta-analysis.

• **OBJECTIVES:** To identify, analyze, and synthesize the literature to determine which physical examination tests, if any, accurately diagnose a torn tibial meniscus.

• **BACKGROUND:** Knee pain has a lifetime prevalence of up to 45%, and as many as 31% of individuals with knee pain will consult a general practitioner. Roughly 5% of these individuals will undergo a tibial meniscectomy and many more will undergo partial meniscectomy or meniscus repair. Determining which of these individuals is appropriate for surgical consult depends on clinical examination findings.

• **METHODS AND MEASURES:** We searched MEDLINE, CINAHL, and SPORTDiscus from1966 to August 2006 and extracted all English- and German-language studies that reported the diagnostic accuracy of individual physical examination tests for a torn meniscus. We retrieved data regarding true positives, false positives, true negatives, and false negatives to create 2-by-2 tables for each article and test. Like tests were then subjected to meta-analysis and subanalysis. Cochran Q test and the I2 statistic were used to examine for the presence of heterogeneity and the extent of the effect of heterogeneity, respectively. A qualitative analysis was also performed using the QUADAS tool.

• **RESULTS:** Eighteen studies qualified for the final analyses. Three physical examination tests (McMurray's, Apley's, and joint line tenderness) were examined in more than 7 studies and had enough data to consider meta-analysis. However, study results were heterogeneous. Pooled sensitivity and specificity were 70% and 71% for McMurray's, 60% and 70% for Apley's, and 63% and 77% for joint line tenderness. Large between-study differences could not be explained by prevalence, study quality, or how well an index test was described.

• **CONCLUSIONS:** No single physical examination test appears to accurately diagnose a torn tibial meniscus and the value of history plus physical examination is unknown. Differences between studies in diagnostic performance remain unexplained, presumably due to local differences in the way the tests are defined, performed, and interpreted. We recommend a more standardized approach to performing and interpreting these tests and the development of a clinical prediction rule to aid clinicians in the diagnosis of a torn tibial meniscus. *J Orthop Sports Phys Ther 2007;39(9):541–550. doi:10.2519/jospt.2007.2560*

• **KEY WORDS:** *Apley's, diagnosis, joint line tenderness, McMurray's, primary care, tibiofemoral joint*

[1]Assistant Professor, Duke University Medical Center, Durham, NC. [2]Research Professor, Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC. [3]Instructor, Duke University Medical Center, Durham, NC. [4]Associate Professor of Medicine, Duke University Medical Center, Center for Clinical Health Policy Research, Durham, NC. Address correspondence to Dr Eric J. Hegedus, Assistant Professor, Duke University, DUMC 3907, Durham, NC 27710. E-mail: eric.hegedus@duke.edu

al[41] review was published, 5 additional articles[2,14,23,24,36] reporting on the diagnostic accuracy of individual special tests have been published, and we found 1 additional article,[27] published in 1999, that was not included in any of the previous meta-analyses. Additionally, the most recently published (2003) review[45] focused on the clinical exam in general and, therefore, cited only 4 studies of individual special tests, all of which were published in the 1980s. For these reasons, this study will provide an updated review of published literature pertaining to the most common and newest meniscal tests.

TABLE 1 Search Strategy

Number	Search History	Results
1	exp Menisci, Tibial/	4066
2	menisc$.ti.	3307
3	1 or 2	4795
4	exp Physical Examination/	490183
5	clinical examination.tw.	16484
6	(physical adj2 exam$).tw.	26131
7	(objective adj2 exam$).tw.	12976
8	(clinical adj2 test$).tw.	12927
9	(special adj2 test$).tw.	812
10	4 or 5 or 6 or 7 or 8 or 9	548782
11	3 and 10	495
12	limit 11 to humans	470

METHODS

Search Strategy

The search strategy included a literature search within the dates of 1966 to August 2006 for the terms tibial menisci and physical examination using the MEDLINE, CINAHL, and SPORTDiscus databases (**TABLE 1**). Results were limited to studies involving humans, published in the English or German languages. To maximize the sensitivity of the search strategy, the generic search strategy reported by Haynes et al[19] was not employed, because many of the older articles may not have used sensitivity, diagnosis, and other terms related to diagnostic accuracy. Recent journals and personal files were hand searched by 2 of the authors (E.H. and C.C.) independently for publications, posters, or abstracts. The reference lists in review articles were cross-checked and all individual names of each special test were queried using MEDLINE.

TABLE 2 Summary of Articles for Torn Menisci: McMurray's Test

Study	Number and Sex of Subjects	Affected Meniscus	Sensitivity/Specificity	+LR/–LR	CS	QUADAS Score*
Karachalios et al 2005[24]	301 M, 109 F	Med	48/94	8.2/0.55	MR	8
		Lat	65/86	4.7/0.41		
Akseki et al 2004[2]	110 M, 40 F	Med	65/71	2.2/0.50	S	11
		Lat	68/90	6.9/0.36		
Jerosch and Reimer 2004[23]	42 M, 22 F	Med, lat	74/11	0.83/2.35	S	11
Pookarnjanamorakat 2004[36]	95 M, 5 F	Med, lat	28/92	3.5/0.78	S	8
Kurosaka et al 1999[27]	83 M, 73 F	Med, lat	37/77	1.6/0.82	S	10
Corea et al 1994[11]	93 M, 0 F	Med	65/93	9.5/0.38	S	9
		Lat	52/94	8.0/0.52		
Grifka et al 1994[18]	61 M, 52 F	Med, lat	66/38	1.1/0.91	S	9
Evans et al 1993[15]	Not stated	Med, lat	24/93	3.5/0.82	S	9
Saengnipanthkul et al 1992[40]	148 M, 42 F	Med	47/94	8.5/0.56	S	8
Boeree and Ackroyd 1991[7]	154 M, 49 F	Med	29/87	2.3/0.81	MR	8
		Lat	25/90	2.4/0.84		
Fowler and Lubliner 1989[17]	106 M, 55 F	Med, lat	29/96	7.8/0.74	S	10
Steinbruck and Wiehmann 1988[46]	205 M, 95 F	Med	34/86	2.4/0.77	S	10
		Lat	15/97	4.9/0.88		
Anderson and Lipscomb 1986[3]	76 M, 24 F	Med, lat	58/29	0.82/1.5	S	10
Noble and Erat 1980[34]	163 M, 37 F	Med, lat	62/57	1.4/0.67	S	9

Abbreviations: CS, criterion standard; F, female; Lat, lateral; +LR, positive likelihood ratio; –LR, negative likelihood ratio; M, male; Med, medial; MR, magnetic resonance imaging; QUADAS, Quality of Diagnostic Accuracy Studies; S, surgery.
** Score indicates number of unequivocal yes responses out of 14 total.*

Study Selection

All abstracts for 470 articles from Medline, 65 articles from CINAHL, 34 articles from SPORTDiscus, and 9 articles from the hand search were reviewed by 2 of the authors (E.H. and C.C.) independently. After the abstracts of the articles from the computer and hand search were read, agreement between the 2 authors as to which articles to read in full was determined by consensus. After independently reading the articles in full and applying the inclusion/exclusion criteria, the 2 authors (E.H. and C.C.) arrived at a final list of articles for inclusion in this paper. If there was disagreement as to the final selection, a third author (D.M.) made the conclusive decision. Articles

TABLE 3 Summary of Articles for Torn Menisci: Joint Line Tenderness Test

Study	Number and Sex of Subjects	Affected Meniscus	Sensitivity/Specificity	+LR/–LR	CS	QUADAS Score*
Karachalios et al 2005[24]	301 M, 109 F	Med	71/87	5.5/0.33	MR	8
		Lat	78/90	7.9/0.24		
Pookarnjanamorakat 2004[36]	95 M, 5 F	Med, lat	27/96	6.7/0.76	S	8
Akseki et al 2004[2]	110 M, 40 F	Med	85/45	1.55/0.34	S	10
		Lat	81/84	5.1/0.23		
Eren 2003[14]	104 M, 0 F	Med	86/67	2.6/0.20	S	10
		Lat	93/97	36.0/0.08		
Kurosaka et al 1999[27]	83 M, 73 F	Med, lat	55/67	1.6/0.68	S	10
Shelbourne et al 1995[44]	118 M, 55 F	Med	58/53	1.2/0.79	S	10
		Lat	38/70	1.3/0.89		
Grifka et al 1994[18]	61 M, 52 F	Med, lat	95/5	0.99/1.1	S	9
Saengnipanthkul et al 1992[40]	148 M, 42 F	Med	58/74	2.2/0.57	S	8
Boeree and Ackroyd 1991[7]	154 M, 49 F	Med	64/69	2.1/0.52	MR	8
		Lat	28/87	2.1/0.84		
Abdon et al 1990[1]	110 M, 35 F	Med	78/69	2.5/0.31	S	9
		Lat	22/98	8.84/0.80		
Fowler and Lubliner 1989[17]	106 M, 55 F	Med, lat	85/30	1.2/0.51	S	10
Steinbruck and Wiehmann 1988[46]	205 M, 95 F	Med	73/62	1.9/0.43	S	10
		Lat	53/91	5.9/0.52		
Barry et al 1983[6]	37 M, 7 F	Med, Lat	86/43	1.5/0.32	S	8
Noble and Erat 1980[34]	163 M, 37 F	Med, lat	72/13	0.83/2.1	S	9

Abbreviations: CS, criterion standard; F, female; Lat, lateral; +LR, positive likelihood ratio; –LR, negative likelihood ratio; M, male; Med, medial; MR, magnetic resonance imaging; QUADAS, Quality of Diagnostic Accuracy Studies; S, surgery.
** Score indicates number of unequivocal yes responses out of 14 total.*

TABLE 4 Summary of Articles for Torn Menisci: Apley's Test

Study	Number and Sex of Subjects	Affected Meniscus	Sensitivity/Specificity	+LR/–LR	CS	QUADAS Score*
Karachalios et al 2005[24]	301 M, 109 F	Med	41/93	5.9/0.63	MR	8
		Lat	41/86	2.9/0.69		
Pookarnjanamorakat 2004[36]	95 M, 5 F	Med, lat	16/100	8.6/0.85	S	8
Jerosch and Reimer 2004[23]	42 M, 22 F	Med, lat	70/33	1.0/0.91	S	11
Kurosaka et al 1999[27]	83 M, 73 F	Med, lat	13/90	1.31/0.97	S	10
Grifka et al 1994[18]	61 M, 52 F	Med, lat	60/70	2.0/0.57	S	9
Fowler and Lubliner 1989[17]	106 M, 55 F	Med, lat	16/80	0.78/1.06	S	10
Steinbruck and Wiehmann 1988[46]	205 M, 95 F	Med	47/82	2.63/0.65	S	10
		Lat	23/99	20.0/0.78		

Abbreviations: CS, criterion standard; F, female; Lat, lateral; +LR, positive likelihood ratio; –LR, negative likelihood ratio; M, male; Med, medial; MR, magnetic resonance imaging; QUADAS, Quality of Diagnostic Accuracy Studies; S, surgery.
** Score indicates number of unequivocal yes responses out of 14 total.*

TABLE 5 Summary of Articles for Torn Menisci: Other Tests

Study	Number and Sex of Subjects	Affected Meniscus	Sensitivity/Specificity	+LR/−LR	CS	QUADAS Score*
Karachalios et al 2005[24] (Thessaly test 20°)	301 M, 109 F	Med Lat	89/97 92/96	29.7/0.11 23.0/0.08	MR	8
Akseki et al 2004[2]	110 M, 40 F	Med Lat	67/81 64/90	3.53/0.41 6.4/0.40	S	11
Pookarnjanamorakat 2004[36] (Merke's sign)	95 M, 5 F	Med, lat	71/83	4.18/0.35	S	8
Pookarnjanamorakat 2004[36] (Steinmann I sign)	95 M, 5 F	Med, lat	27/96	6.8/0.76	S	8
Mariani et al[30] 1996 (dynamic test)	243 M, 162 F	Lat	85/90	8.5/0.17	S	9

Abbreviations: CS, criterion standard; F, female; Lat, lateral; +LR, positive likelihood ratio; –LR, negative likelihood ratio; M, male; Med, medial; MR, magnetic resonance imaging; QUADAS, Quality of Diagnostic Accuracy Studies; S, surgery.
** Score indicates number of unequivocal yes responses out of 14 total.*

were eligible for inclusion if the criterion standard was surgery or magnetic resonance imaging (MRI), at least 1 physical examination test/special test was studied, if the paired statistics of sensitivity and specificity were both reported for an individual test, and if the article was in the English or German languages. Studies were excluded if the special test was performed under anesthesia or in cadavers, if a group of special tests were assigned the status of "composite physical examination," or if the article was a review. The reviewers were familiar with the literature, thus were not blinded to the authors, the date of publication, or the journals in which the manuscripts were published. A summary of the articles pulled for review based on a consensus of the authors is presented in **TABLES 2** through **5**. These tables have been organized by special test.

TABLE 6 Quality Assessment of Diagnostic Accuracy Studies (QUADAS) Tool

Item	Yes	No	Unclear
Was the spectrum of patients representative of the patients who will receive the test in practice?	__	__	__
Were selection criteria clearly described?	__	__	__
Is the reference standard likely to classify the target condition correctly?	__	__	__
Is the period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	__	__	__
Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis?	__	__	__
Did patients receive the same reference standard regardless of the index test result?	__	__	__
Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	__	__	__
Was the execution of the index test described in sufficient detail to permit replication of the test?	__	__	__
Was the execution of the reference standard described in sufficient detail to permit its replication?	__	__	__
Were the index test results interpreted without knowledge of the results of the reference standard?	__	__	__
Were the reference standard results interpreted without knowledge of the results of the index test?	__	__	__
Were the same clinical data available when test results were interpreted as would be available as when the test is used in practice?	__	__	__
Were uninterpretable/intermediate test results reported?	__	__	__
Were withdrawals from the study explained?	__	__	__

** Reproduced with permission of Dr P. Whiting.[49]*

Quality Assessment

After all relevant articles were obtained, their quality was assessed and data were extracted from each article. The quality of each study was determined unmasked by the lead author examining its internal and external validity using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool developed by Whiting et al[49] (**TABLE 6**). QUADAS involves individualized scoring of 14 components. Each of the 14 components is scored as "yes," "no," or "unclear." Individual procedures for scoring each of the 14 items, including operational standards for each question, have been published, although a cumulative methodological score is not advocated.[48] Contrary to the recommendation of Whiting et al,[48] past studies[12,42,43] have arbitrarily used a score of 7 or more yes answers out of 14 to indicate a high-quality diagnostic accuracy study, whereas scores below 7 were indicative of low quality. Based on extensive experience in use of the QUADAS tool,[10] our consensus higher quality score was defined at 10 or more unequivocal yes answers out of 14, whereas a score below 10 was associated with a low-quality study. Our consensus cut-off score of 10 may still be considered arbitrary. However, this score was only used to facilitate 1 subgroup analysis; no studies were excluded based on this categorization.

Statistical Analyses

Meta-analysis was performed using dr-ROC Version 2.00 software (dr2 Consulting, Glenside, PA). Data were eligible for pooling in 3 special tests: McMurray's,[32] Apley's,[4] and joint line tenderness (JLT). Raw data from each individual study for these 3 tests was placed in a 2-by-2 table and summarized by the paired statistics sensitivity and specificity. The dr-ROC software was used to pool sensitivities and specificities using the inverse variance method, which gives greater weight to individual studies with more subjects. The fixed-effects model was used, as outcomes of both fixed-effects and random-effects models were similar. The diagnostic odds ratio and the area under the curve of the summary receiver operating characteristic curve were both calculated as summary statistics, indicating the overall diagnostic power of each of the 3 tests. The Cochran *Q* test was used to test for heterogeneity and the I2 statistic[20] was used to quantify the percentage of variation across the studies that was associated with heterogeneity. Where there was a lack of multiple studies to achieve a pooled estimate of the diagnostic accuracy of a special test for a torn tibial meniscus, the article results were reported along with an assessment of the quality of that study (**TABLE 5**).

Sources of heterogeneity were explored by performing subgroup analyses. When performing the subgroup analyses, the studies for each test (McMurray's, JLT, or Apley's) were dichotomized based on factors established a priori. These pre-established factors included number of unequivocal yes on the QUADAS quality-scoring instrument, prevalence of torn menisci, and description of a positive test finding. These factors were chosen because quality, prevalence, and varying definitions of a positive test all have an effect on estimates of diagnostic accuracy.[29,39] For each subanalysis, pooled sensitivity, pooled specificity, the diagnostic odds ratio, the area under the curve, the Cochran *Q*, and the I2 statistic were recalculated across subgroups. Subgroup analysis was performed using dr-ROC Version 2.00 software (dr2 Consulting).

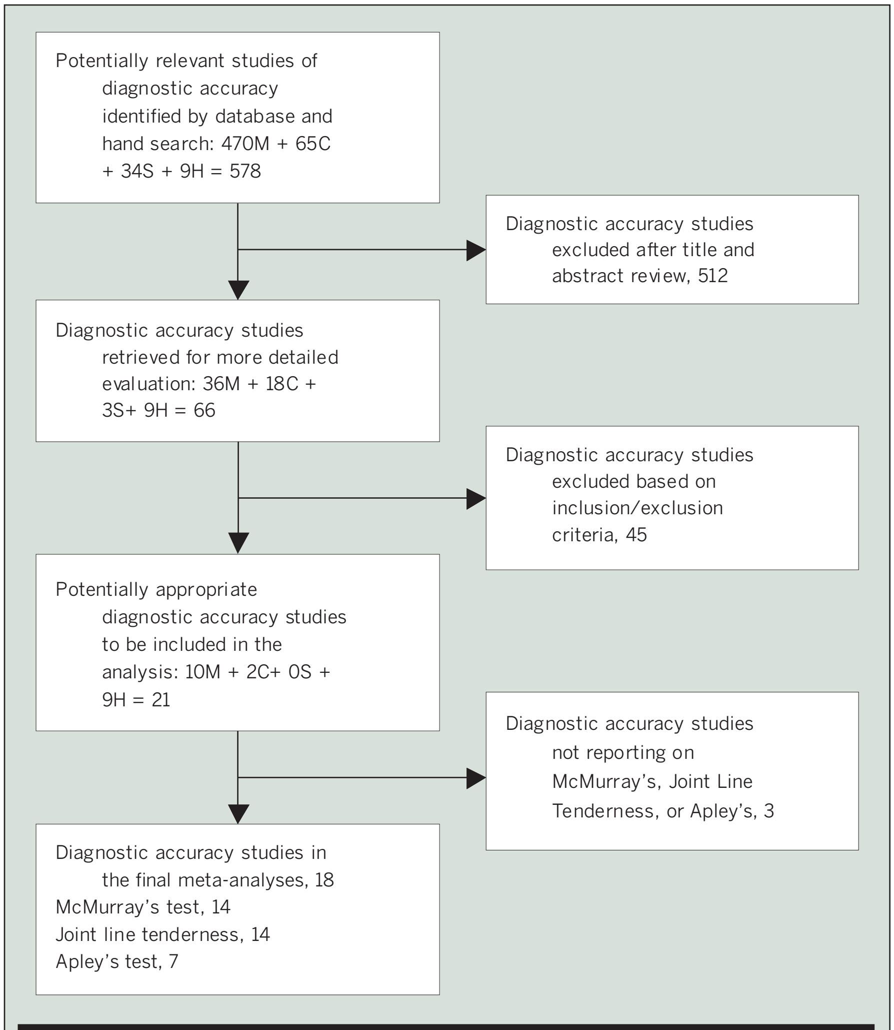

FIGURE. Consort QUOROM flow chart for inclusion criteria within the study. Abbreviations: C, articles from Cinahl; M, articles from MEDLINE; H, articles from the hand search.

RESULTS

The initial database search identified 569 articles, 12 of which were relevant to this study, whereas the hand search revealed 9 additional articles appropriate for the study (**FIGURE**). Of the 21 articles, 18 were considered suitable for summary statistical analysis because they addressed McMurray's, JLT, or Apley's test. A preponderance of the studies were published before 2003.[1,3,6,7,18,23,24,34,40,44,46]

Quality Score Summary

Four studies[11,14,24,36] were judged to have a limited spectrum of subjects, while 2 studies[6,15] lacked a description of subjects

TABLE 7 Summary of Results: Main and Subgroup Analyses

	Pooled SN (95% CI upper limit, lower limit)	Pooled SP (95% CI upper limit, lower limit)	DOR (95% CI upper limit, lower limit)	AUC (95% CI upper limit, lower limit)	Q	P (Q)	I2 (%)
McMurray's test							
Meta-analysis	70.5 (67.4, 73.4)	71.1 (69.3, 72.9)	4.5 (3.7, 5.4)	0.73 (0.71, 0.76)	86	<.01	79
Subanalysis QUADAS							
10+	75.3 (71, 79)	66.5 (64, 69)	3.6 (2.7, 4.9)	0.70 (0.66, 0.75)	32	<.01	77
9–	66.7 (63, 71)	74.2 (72, 76)	5.1 (4.0, 6.6)	0.75 (0.72, 0.78)	53	<.01	81
Subanalysis prevalence							
0.50+	79.2 (75, 83)	37.6 (34, 42)	2.3 (1.7, 3.2)	0.64 (0.59, 0.69)	20	<.01	66
0.49–	60.7 (56, 65)	80.9 (79, 83)	6.8 (5.3, 8.7)	0.79 (0.76, 0.81)	39	<.01	75
Subanalysis definition and test							
Described	73.1 (69, 77)	72.2 (70, 75)	8.5 (6.3,11.3)	0.81 (0.78, 0.84)	26	<.01	65
Not described	67.8 (63, 72)	69.7 (67, 72)	2.3 (1.8, 3.1)	0.64 (0.59, 0.68)	21	<.01	62
Joint line tenderness test							
Meta-analysis	63.3 (60.9, 65.7)	77.4 (75.6, 79.1)	4.5 (3.8, 5.4)	0.73 (0.71, 0.76)	15, 6	<.01	87
Subanalysis QUADAS							
10+	60.5 (57, 64)	76.7 (74, 79)	4.1 (3.3, 5.2)	0.72 (0.69, 0.75)	69	<.01	87
9–	66.4 (63, 70)	78.0 (76, 80)	5.1 (4.0, 6.5)	0.75 (0.72, 0.78)	83	<.01	88
Subanalysis prevalence							
0.50+	68.8 (66, 72)	40.3 (36, 45)	1.8 (1.3, 2.5)	0.60 (0.55, 0.65)	23	<.01	69
0.49–	58.9 (56, 62)	86.1 (84, 88)	6.9 (5.6, 8.4)	0.79 (0.76, 0.81)	90	<.01	87
Subanalysis definition and test							
Described	63.9 (57, 70)	69.4 (63, 75)	7.9 (4.7, 13.3)	0.80 (0.74, 0.85)	25	<.01	88
Not described	63.2 (61, 66)	78.4 (77, 80)	4.2 (3.5, 5.0)	0.72 (0.70, 0.75)	129	<.01	88
Apley's test							
Meta-analysis	60.7 (55.7, 65.5)	70.2 (68, 72.4)	3.4 (2.6, 4.4)	0.69 (0.65, 0.73)	32	<.01	75
Subanalysis QUADAS							
10+	60.7 (53, 68)	64.3 (61, 68)	1.9 (1.3, 2.8)	0.60 (0.54, 0.66)	12	.02	66
9–	60.7 (54, 67)	76.2 (73, 79)	6.1 (4.1, 9.0)	0.77 (0.72, 0.82)	5	.19	37
Subanalysis prevalence							
0.50+	79.5 (72, 86)	28.2 (23, 34)	2.4 (1.4, 4.3)	0.65 (0.55, 0.73)	4	.25	27
0.49–	50.4 (44, 57)	79.9 (78, 82)	3.8 (2.8, 5.1)	0.71 (0.66, 0.75)	25	<.01	84
Subanalysis definition and test							
Described	55.7 (49, 62)	66.6 (64, 69)	3.7 (2.5, 5.3)	0.71 (0.65, 0.75)	27	<.01	85
Not described	66.7 (59, 73)	76.4 (73, 80)	3.0 (2.0, 4.5)	0.68 (0.61, 0.73)	4	.29	21

Abbreviations: AUC, area under the curve; CI, confidence interval; DOR, diagnostic odds ratio; QUADAS, Quality of Diagnostic Accuracy Studies; SN, sensitivity; SP, specificity.

for judgment on spectrum. No study was performed in the primary care setting, nor did any study have an equal ratio of males to females (for all subjects in the meta-analysis, males outnumbered females almost 3 to 1). Selection criteria (both inclusion and exclusion) were clearly described in 9 of 18 articles.[2,3,11,14,23,24,27,34,44] Arthroscopy or arthrotomy were the criterion standards in 16 articles, with MRI serving as the criterion standard in 2 studies.[7,24] Eight of 18 studies reported an adequate period between reference standard and index test results, indicating that the target condition had not changed in the interim.[1,2,15,17,18,23,36,46] All studies confirmed the target condition with the criterion standard in all subjects, with 1 exception.[40] Subjects in all studies received the same reference standard, regardless of the index test result, and that reference standard was independent of the index test in all cases.

Verification bias, which occurs when only subjects with a positive test receive the diagnostic criterion standard, may still have existed in all but 3 studies,[15,23,24] because the special test or tests were used to admit subjects to 15 of the 18 studies. A detailed description of all studied special test or tests was lacking in 11 of 18 studies.[1,3,6,7,18,23,24,34,40,44,46] In 14 of 18 studies, the description of the criterion standard lacked sufficient description.[1-3,6,7,11,15,18,23,24,27,34,36,40,44,46] Blinding of the physician from the results of the special test was reported in 2 of 18 studies.[23,40] Similarly, 2 of 18 studies described the results of the special test as uninterpretable or equivocal.[3,18] Finally, all but 2 studies[17,36] reported no withdrawals or explained those withdrawals sufficiently.

Summary of Analytic Findings

Three special tests—McMurray's,[32] JLT, and Apley's[4]—were included in the meta-analysis. McMurray's[32] test had a pooled sensitivity of 70.5 (95% CI: 67.4 to 73.4) and a pooled specificity of 71.1 (95% CI: 69.3 to 72.9). JLT had a pooled sensitivity of 63.3 (95% CI: 60.9 to 65.7) and a pooled specificity of 77.4 (95% CI: 75.6 to 79.1). Apley's[4] test had a pooled sensitivity of 60.7 (95% CI: 55.7 to 65.5) and a pooled specificity of 70.2 (95% CI: 68.0 to 72.4).

All 3 tests had a significant *P* value for the Cochran *Q* test ($P[Q]<.01$), signifying that statistical heterogeneity was present in the meta-analyses (**TABLE 7**). The I2 value, an indication of how great an effect the heterogeneity had on the meta-analysis of each test,[20] was 79% for the McMurray's meta-analysis, 87% for the JLT meta-analysis, and 75% for the Apley's meta-analysis. These numbers combined with information provided by the summary receiver operating characteristic curves indicate that none of the 3 tests analyzed possess discriminative power in the diagnosis of a torn tibial meniscus secondary to heterogeneity between studies.

Subgroup Analyses

Quality Factors for subgroup analysis were established a priori and included number of unequivocal yes answers on the QUADAS quality-scoring instrument, prevalence of torn menisci, and description of a positive test finding. The effect of case control design was also chosen as an a priori subgroup analysis secondary to recent evidence that this research design heavily biases diagnostic accuracy studies.[29,39] Because only 2 studies[15,24] incorporated this design, subgroup analysis was not attempted. Heterogeneity remained significant among high-quality (QUADAS score of 10 or greater) and low-quality (QUADAS score of less than 10) studies for both the McMurray's[32] and JLT tests. However, for Apley's[4] test, the group of 3 lower-quality studies[18,24,36] showed a $P(Q)$ value of .19, suggesting homogeneity. Nonetheless, studies varied significantly in design (case control, prospective, and retrospective), patient population, and report of the index test or tests. Furthermore, small sample sizes prevented drawing further conclusions about Apley's[4] test.

Prevalence Prevalence of torn menisci allowed dichotomization of studies into groups, with prevalence 0.50 and above and 0.49 and below. Heterogeneity remained among the 2 groups for both the McMurray's[32] and JLT tests. The pooled sensitivity for both tests rose slightly in the higher prevalence group (79% and 69%, respectively) at a greater cost in specificity (38% and 40%, respectively). For Apley's[4] test, the group of 4 studies with higher prevalence[18,23,27,36] showed a $P(Q)$ value of .25, signifying homogeneity. Studies varied significantly in design (case control, prospective, and retrospective), blinding, and report of the index test or tests. Moreover, the small sample size prevented drawing further conclusions about Apley's[4] test.

Definition Subgroup analysis of definition of a positive finding demonstrated comparable findings with the previous 2 subgroup analyses. The studies fell into 3 categories: those with no description of the index test, those that described the index test according to the original author, and those that used a modified description of the original index test. These 3 categories were dichotomized for the subgroup analysis into studies with clear index test description and those without a clear description. Heterogeneity remained among the 2 groups for both the McMurray's[32] and JLT tests. For Apley's[4] test, heterogeneity was not significant ($P[Q]$ = .29) for the 3 studies[18,23,46] in the group where the index test was not described.

DISCUSSION

THE PURPOSES OF THIS ANALYSIS were to summarize the available literature on the diagnostic accuracy of physical examination tests to detect a torn tibial meniscus and to pool the data from original articles to produce an estimate of the clinical utility of these special tests. Available German- and English-language literature produced a sufficient quantity of data for 3 tests: McMurray's,[32] JLT, and Apley's.[4] The diagnostic accuracy of these 3 tests appears to be poor, but this conclusion is tenuous, based on the poor quality of the studies and the amount of heterogeneity in the data. Despite close examination of the data from the main analysis and the subgroup analyses, the source or sources of heterogeneity could not be identified. Therefore, we are left to speculate, based on clinical experience, as to the source of heterogeneity. First, despite a call in 1999 for "large, simple studies of clinical examination,"[31] a study in 2005 by Flahault et al[16] that precisely listed sample size requirements for diagnostic accuracy studies with dichotomous outcomes and the high-profile work of the Standards for Reporting of Diagnostic Accuracy (STARD) initiative[8,9] to help with design and reporting of diagnostic accuracy studies, 0 of 18 studies have a large enough sample size to detect, for example, a sensitivity or specificity of 90% with a lower limit of the 95% confidence interval around that point estimate at or above 85%.[16] Further, despite the STARD initiative[8,9] criteria for the designing and reporting of diagnostic accuracy studies, the design/reporting of all 18 studies is lacking in some fashion. Future research needs to follow the design and reporting recommendations of the STARD initiative,[8,9] with the sample size recommendations of Flauhalt et al,[16] to produce large, well-designed studies of diagnostic accuracy.

In addition to bias and lack of sample size, which limits power, threshold differences may produce heterogeneity. Because McMurray's,[32] JLT, and Apley's[4] tests all involve pain, the interpretation of pain by multiple examiners is likely to vary (a change in the threshold) and, therefore, change sensitivity and specificity values among studies. A more strict interpretation of pain increases specificity, while a less strict interpretation increases sensitivity. That the 3 special tests are performed inconsistently is plausible, because Kappa values, numbers reflecting agreement beyond chance, are poor to fair[13,15,17] according to the scale advocated by Landis and Koch.[28]

The results of our meta-analysis compared to those of 3 previous meta-analyses performed by Scholten et al,[41] Solomon et al,[45] and Jackson et al[22] can be found in **TABLE 8**. Our more extensive meta-analysis showed McMurray's[32] test to be more sensitive and JLT to be more specific than previously reported. Further, we have the first study that incorporates the QUADAS tool and with enough data on Apley's[4] test to perform a meta-analysis. Unfortunately, Apley's[4] test appears to have poorer accuracy than either McMurray's[32] test or JLT. Therefore, despite more extensive data, our conclusion is no different from that of our brethren: individual special tests are of little value in diagnosing a torn tibial meniscus.

TABLE 8 Comparison of Meta-analyses

	Hegedus et al	Scholten et al[41]	Solomon et al[45]	Jackson et al[22]
Number of studies	18	13	9	4
Sample size (n)	2670	2231	1018	424
Mcmurray's[32] test				
Summary statistic	SN, 71; SP, 71	SN, 48; SP, 86	SN, 53; SP, 59	SN, 52; SP, 97
Joint line tenderness test				
Summary statistic	SN, 63; SP, 77	SN, 77; SP, 41	SN, 79; SP, 15	SN, 76; SP, 29
Apley's[4] test				
Summary statistic	SN, 61; SP, 70	Not reported	Not reported	Not reported

Abbreviations: SN, sensitivity; SP, specificity.

Because the McMurray's,[32] JLT, or Apley's[4] tests are not diagnostically accurate when used alone, the clinician can either combine these tests with other components of physical examination or abandon the use of these tests for possibly more promising tests. Several studies have combined special tests with other components of physical examination like patient history and imaging, and physical signs like swelling.[13,23,25,26,35,38] Unfortunately, no conclusions can be made when examining these studies, due to their vast differences. One study[13] examined the ability of the composite physical examination (CPE) to detect an unstable meniscus in patients with primary osteoarthritis of the knee and reported a sensitivity of 88% and a specificity of 20%. These unremarkable numbers, combined with a fair level of agreement (κ = .24), indicate that even the CPE is not accurate for diagnosis in this patient population. Another study[26] examined the accuracy of the CPE in athletic children and found that the CPE modified the posttest probability of detecting a torn meniscus by only a small amount. A third study[35] enrolled only male military personnel and found the CPE to be only slightly more useful in this population than in athletic children. The last 3 studies[23,25,38] advocated the combination of history with 3 to 5 special tests. In 2 of these 3 remaining studies,[25,38] the CPE for a torn medial meniscus was more sensitive than specific, while for a torn lateral meniscus the CPE was more specific than sensitive.

Because the diagnostic performance of the CPE is equivocal at this time, another solution may be to look to alternate physical examination tests. Several recently described weight-bearing tests demonstrated promising diagnostic accuracy but lacked a sufficient number of articles for meta-analysis. The Thessaly test,[24] which is performed at 20° of knee flexion, had sensitivity of 89% and specificity of 97% for the medial meniscus, and sensitivity of 92% and specificity of 96% for the lateral meniscus.[24] Merke's sign involves standing rotation but is performed in full knee extension, and may have value as a positive test to rule in a torn meniscus. Finally, another weight-bearing test that may have value as a positive test to rule in a torn meniscus is the Ege's test, which involves rotation of the lower extremities and squatting, with a specificity of 81% for the medial meniscus and 90% for the lateral meniscus.[2] While these statistics are promising, caution is advised, because each of these tests has only been studied once, and at least 1 of these studies[24] is a case control design that overstates the diagnostic accuracy of the test.[29,39] Finally, for those patients who cannot tolerate weight bearing or squatting, the Dynamic test[30] or Steinmann I sign may be worth further investigation. All of these tests, weight bearing and non-weight bearing alike, are summarized in **TABLE 5**.

The chief limitation of this meta-analysis is that of the others: the quality of the contributing studies limits the conclusions that can be drawn from the synthesis of those studies. There was heterogeneity in the data among studies for the McMurray's,[32] JLT, and Apley's[4] tests. The heterogeneity may have had a number of sources. The patient populations varied from study to study in their ages, ratio of males to females (with females being mostly underrepresented), and in the chronicity of injury. Study designs also varied, including prospective, retrospective, and case control. Both retrospective and case control designs elevate the estimates of diagnostic accuracy.[29,39] Further, there were many sources of bias inherent in the design of the studies that have been shown to vary diagnostic accuracy. Most of the studies lacked a detailed description of the reference standard, and many studies used the index test as part of the admission criteria for the study. Both of these design features decrease the estimate of diagnostic accuracy.[29,39] Also, many studies failed to mask the physician from the results of the special tests and failed to describe the special test. These design features tend to overestimate diagnostic accuracy.[29,39] One final limitation was that the lead author was the only author to perform qualitative assessment of the articles using the QUADAS tool, which may have affected 1 subgroup analysis. Despite the fact that interrater agreement for individual items on the QUADAS tool is low, correlation between the total score assigned by separate raters is high.[21] Therefore, the effect on the subgroup analysis based on quality is probably minimal.

CONCLUSION

The current literature examining the diagnostic accuracy of special tests to detect a torn tibial meniscus shows that Apley's, McMurray's, and joint line tenderness tests are not diagnostic. This conclusion must be tempered by the fact that all of the studies are underpowered,[16] most of the studies possess numerous design or reporting faults,[8,9] and the heterogeneity between studies makes the summary estimates of sensitivity and specificity less than valuable.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contribution of Matt McCree, DPT for his assistance in translating German language articles and Ted Worrell, PT, EdD, SCS, FACSM for his guidance in this line of scientific inquiry. I also acknowledge, as the corresponding author, that I have recognized all individuals that have contributed to this paper and that each of my coauthors has contributed significantly to this work.

REFERENCES

1. Abdon P, Lindstrand A, Thorngren KG. Statistical evaluation of the diagnostic criteria for meniscal tears. *Int Orthop.* 1990;14:341-345.
2. Akseki D, Ozcan O, Boya H, Pinar H. A new weight-bearing meniscal test and a comparison with McMurray's test and joint line tenderness. *Arthroscopy.* 2004;20:951-958.
3. Anderson AF, Lipscomb AB. Clinical diagnosis of meniscal tears. Description of a new manipulative test. *Am J Sports Med.* 1986;14:291-293.
4. Apley AG. The diagnosis of meniscus injuries. *J Bone Joint Surg.* 1947;29:78-84.
5. Baker P, Reading I, Cooper C, Coggon D. Knee disorders in the general population and their relation to occupation. *Occup Environ Med.* 2003;60:794-797.
6. Barry OC, Smith H, McManus F, MacAuley P. Clinical assessment of suspected meniscal tears. *Ir J Med Sci.* 1983;152:149-151.
7. Boeree NR, Ackroyd CE. Assessment of the menisci and cruciate ligaments: an audit of clinical practice. *Injury.* 1991;22:291-294.
8. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med.* 2003;138:40-44.
9. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med.* 2003;138:W1-12.
10. Cook C, Hegedus E. *Orthopedic Physical Examination Tests: An Evidence-Based Approach.* Upper Saddle River, NJ: Prentice-Hall; 2007.
11. Corea JR, Moussa M, al Othman A. McMurray's test tested. *Knee Surg Sports Traumatol Arthrosc.* 1994;2:70-72.
12. de Graaf I, Prak A, Bierma-Zeinstra S, Thomas S, Peul W, Koes B. Diagnosis of lumbar spinal stenosis: a systematic review of the accuracy of diagnostic tests. *Spine.* 2006;31:1168-1176.
13. 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;44:267-274.
14. Eren OT. The accuracy of joint line tenderness by physical examination in the diagnosis of meniscal tears. *Arthroscopy.* 2003;19:850-854.
15. Evans PJ, Bell GD, Frank C. Prospective evaluation of the McMurray test. *Am J Sports Med.* 1993;21:604-608.
16. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol.* 2005;58:859-862.
17. Fowler PJ, Lubliner JA. The predictive value of five clinical signs in the evaluation of meniscal pathology. *Arthroscopy.* 1989;5:184-186.
18. Grifka J, Richter J, Gumtau M. [Clinical and sonographic meniscus diagnosis]. *Orthopade.* 1994;23:102-111.
19. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. *BMJ.* 2004;328:1040.
20. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327:557-560.
21. Hollingworth W, Medina LS, Lenkinski RE, et al. Interrater reliability in assessing quality of diagnostic accuracy studies using the QUADAS tool. A preliminary assessment. *Acad Radiol.* 2006;13:803-810.
22. Jackson JL, O'Malley PG, Kroenke K. Evaluation of acute knee pain in primary care. *Ann Intern Med.* 2003;139:575-588.
23. Jerosch J, Riemer S. [How good are clinical investigative procedures for diagnosing meniscus lesions?]. *Sportverletz Sportschaden.* 2004;18:59-67.
24. Karachalios T, Hantes M, Zibis AH, Zachos V, Karantanas AH, Malizos KN. Diagnostic accuracy of a new clinical test (the Thessaly test) for early detection of meniscal tears. *J Bone Joint Surg Am.* 2005;87:955-962.
25. Kocabey Y, Tetik O, Isbell WM, Atay OA, Johnson DL. The value of clinical examination versus magnetic resonance imaging in the diagnosis of meniscal tears and anterior cruciate ligament rupture. *Arthroscopy.* 2004;20:696-700.
26. Kocher MS, DiCanzio J, Zurakowski D, Micheli LJ. Diagnostic performance of clinical examination and selective magnetic resonance imaging in the evaluation of intraarticular knee disorders in children and adolescents. *Am J Sports Med.* 2001;29:292-296.
27. Kurosaka M, Yagi M, Yoshiya S, Muratsu H, Mizuno K. Efficacy of the axially loaded pivot shift test for the diagnosis of a meniscal tear. *Int Orthop.* 1999;23:271-274.
28. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159-174.
29. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA.* 1999;282:1061-1066.
30. Mariani PP, Adriani E, Maresca G, Mazzola CG. A prospective evaluation of a test for lateral meniscus tears. *Knee Surg Sports Traumatol Arthrosc.* 1996;4:22-26.
31. McAlister FA, Straus SE, Sackett DL. Why we need large, simple studies of the clinical examination: the problem and a proposed solution. CARE-COAD1 group. Clinical Assessment of the Reliability of the Examination-Chronic Obstructive Airways Disease Group. *Lancet.* 1999;354:1721-1724.
32. McMurray TP. The semilunar cartilages. *Br J Surg.* 1942;29:407-414.
33. Miller GK. A prospective study comparing the accuracy of the clinical diagnosis of meniscus tear with magnetic resonance imaging and its effect on clinical outcome. *Arthroscopy.* 1996;12:406-413.
34. Noble J, Erat K. In defence of the meniscus. A prospective study of 200 meniscectomy pa-

tients. *J Bone Joint Surg Br.* 1980;62-B:7-11.
35. O'Shea KJ, Murphy KP, Heekin RD, Herzwurm PJ. The diagnostic accuracy of history, physical examination, and radiographs in the evaluation of traumatic knee disorders. *Am J Sports Med.* 1996;24:164-167.
36. Pookarnjanamorakot C, Korsantirat T, Woratanarat P. Meniscal lesions in the anterior cruciate insufficient knee: the accuracy of clinical evaluation. *J Med Assoc Thai.* 2004;87:618-623.
37. Renstrom P, Johnson RJ. Anatomy and biomechanics of the menisci. *Clin Sports Med.* 1990;9:523-538.
38. Rose NE, Gold SM. A comparison of accuracy between clinical examination and magnetic resonance imaging in the diagnosis of meniscal and anterior cruciate ligament tears. *Arthroscopy.* 1996;12:398-405.
39. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *Cmaj.* 2006;174:469-476.
40. Saengnipanthkul S, Sirichativapee W, Kowsuwon W, Rojviroj S. The effects of medial patellar plica on clinical diagnosis of medial meniscal lesion. *J Med Assoc Thai.* 1992;75:704-708.
41. Scholten RJ, Deville WL, Opstelten W, Bijl D, van der Plas CG, Bouter LM. The accuracy of physical diagnostic tests for assessing meniscal lesions of the knee: a meta-analysis. *J Fam Pract.* 2001;50:938-944.
42. Sehgal N, Shah RV, McKenzie-Brown AM, Everett CR. Diagnostic utility of facet (zygapophysial) joint injections in chronic spinal pain: a systematic review of evidence. *Pain Physician.* 2005;8:211-224.
43. Shah RV, Everett CR, McKenzie-Brown AM, Sehgal N. Discography as a diagnostic test for spinal pain: a systematic and narrative review. *Pain Physician.* 2005;8:187-209.
44. Shelbourne KD, Martini DJ, McCarroll JR, VanMeter CD. Correlation of joint line tenderness and meniscal lesions in patients with acute anterior cruciate ligament tears. *Am J Sports Med.* 1995;23:166-169.
45. Solomon DH, Simel DL, Bates DW, Katz JN, Schaffer JL. The rational clinical examination. Does this patient have a torn meniscus or ligament of the knee? Value of the physical examination. *JAMA.* 2001;286:1610-1620.
46. Steinbruck K, Wiehmann JC. [Examination of the knee joint. The value of clinical findings in arthroscopic control]. *Z Orthop Ihre Grenzgeb.* 1988;126:289-295.
47. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. *N Engl J Med.* 1985;313:793-799.
48. Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Med Res Methodol.* 2005;5:19.
49. Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health Technol Assess.* 2004;8:iii, 1-234.