Diagnostic accuracy of clinical tests for the diagnosis of hip femoroacetabular impingement/labral tear: a systematic review with meta-analysis

M P Reiman,¹,² A P Goode,¹ C E Cook,¹ P Hölmich,³,⁴ K Thorborg³,⁵

ABSTRACT
Background Surgery for hip femoroacetabular impingement/acetabular labral tear (FAI/ALT) is exponentially increasing despite lacking investigation of the accuracy of various diagnostic measures. Useful clinical utility of these measures is necessary to support diagnostic imaging and subsequent surgical decision-making.

Objective Summarise/evaluate the current diagnostic accuracy of various clinical tests germane to hip FAI/ALT pathology.

Methods A computer-assisted literature search of MEDLINE, CINAHL and EMBASE databases using keywords related to diagnostic accuracy of the hip joint, as well as the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were used for the search and reporting phases of the study. Quality assessment of bias and applicability was conducted using the Quality of Diagnostic Accuracy Studies-2 (QUADAS-2). Random effects models were used to summarise sensitivities (SN), specificities (SP), diagnostic odds ratio (DOR) and respective confidence intervals (CI).

Results The employed search strategy revealed 21 potential articles, with one demonstrating high quality. Nine articles qualified for meta-analysis. The meta-analysis demonstrated that flexion-adduction-internal rotation (pooled SN ranging from 0.94 (95% CI 0.90 to 0.97) to 0.99 (95% CI 0.98 to 1.00); DOR 5.71 (95% CI 0.84 to 38.86) to 7.82 (95% CI 1.06 to 57.84)) and flexion-internal rotation (pooled SN 0.96 (95% CI 0.81 to 0.99); DOR 8.36 (95% CI 0.41 to 171.3) tests possess only screening accuracy.

Conclusions Few hip physical examination tests for diagnosing FAI/ALT have been investigated in enough studies of substantial quality to direct clinical decision-making. Further high-quality studies across a wider spectrum of hip pathology patients are recommended to discern the confirmed clinical utility of these tests.

Trials registration number PROSPERO Registration # CRD42014010144.

INTRODUCTION
Femoroacetabular impingement (FAI), an abnormal articulation and abutment of the femoral head against the acetabulum, is suggested to contribute to acetabular labral tear (ALT) and chondrolabral injury.¹ The prevalence of ALT in patients with hip or groin pain ranges from 22% to 55%,²⁻⁵ although higher prevalence would be expected in a hip surgeon’s consultancy.

Surgery for the correction of FAI has increased significantly over the past decade.⁶⁻⁸ In fact, an 18-fold increase between 1999 and 2009 was shown in the USA.⁵ Given that differential diagnosis for the patient presenting with hip or groin pain is still suggested to be a diagnostic challenge,² focus on proper diagnosis would seem warranted. A significant increase in attention has been shown for differential diagnosis of the hip joint and periarticular structures as the primary source for hip-related pain/dysfunction,⁴⁻¹⁰ although the principal focus is on radiographic imaging, and, to a lesser extent, clinical examination.¹⁵⁻¹⁷ The diagnostic accuracy of radiographic imaging for FAI/ALT,¹³⁻¹⁴ findings of radiographic asymmetries in asymptomatic individuals,¹⁸⁻⁴⁸ as well as availability and cost of such imaging merits determination of the clinical utility of the clinical examination for FAI/ALT pathologies.

While the diagnostic accuracy of hip physical examination (HPE) tests for FAI/ALT have previously been described,⁴⁵⁻⁴⁷ the estimation of test probability and, therefore, true clinical utility has not been synthesised. Therefore, the purpose of this study was to conduct a systematic review and meta-analysis of the literature regarding the diagnostic accuracy of HPE tests for FAI/ALT and describe their clinical utility.

METHODS
Registration
The study was registered on 6 August 2014 with the International Prospective Register of Systematic Reviews (PROSPERO# CRD42014010144). PROSPERO is a database of prospectively registered systematic reviews for health and social topics. The study was registered after pilot search and prior to updated data search and extraction.

Data sources
The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were utilised during the search and reporting phase of this review. The PRISMA statement includes a 27-item checklist that is designed to be used as a basis for reporting systematic reviews of randomised trials,⁴⁹ but the checklist can also be applied to multiple forms of research methodologies.⁵⁰ A computer-assisted literature search of MEDLINE, CINAHL and EMBASE databases was performed from inception of each respective database to 10 June 2014. The goal was to optimise the sensitivity our search strategy,¹⁰ to ⁵ and to increase the likelihood that all appropriate studies were identified. The search strategy was developed in collaboration with a medical information specialist and used controlled vocabulary and key words related to...
diagnostic accuracy of the clinical examination measures relative to hip femoroacetabular impingement and labral tear.

Screening filters were initially used during assessment of title, abstract and full text documents. The search was further limited to humans and English-only publications. The search strategy for MEDLINE is listed in online supplementary appendix 1.

**Selection criteria**

Articles examining clinical examination for FAI/ALT pathology were eligible if they met all of the following criteria: (1) included participants with hip pain suspected to be related to hip FAI/ALT; (2) included at least one clinical hip FAI/ALT pathology examination measure, (3) utilised an acceptable reference standard, (4) reported the results in sufficient detail to allow reconstruction of contingency tables to allow calculation of diagnostic accuracy metrics and (5) were written in English.

An article was excluded if: (1) the pathology was associated with a condition located elsewhere (eg, lumbar spine) that referred pain to the hip region, (2) the clinical measures were performed under any form of anaesthesia or on cadavers, (3) specialised instrumentation not readily available to all clinicians were used, (4) physical clinical tests were included as a component of cluster tests, and (5) testing was performed on infants/toddlers.

**Data extraction**

Two reviewers (MPR, KT) independently performed the search. Since computerised search results for diagnostic accuracy data frequently omit many relevant articles, the reference lists of all selected publications were checked to retrieve relevant publications that were not identified in the computerised search. Grey literature was also manually searched and included publications, posters, abstracts or conference proceedings. To identify relevant articles, titles and abstracts of all identified citations were independently screened. Full-text articles were retrieved if the abstract provided insufficient information to establish eligibility or if the article passed the first eligibility screening.

All criteria were independently applied by two reviewers (MPR, KT) to the full text of the articles that passed the first eligibility screening. Disagreements among the reviewers were discussed and resolved by a third reviewer (CEC). We determined which articles to include (for meta-analysis) using clinical and statistical judgement of study heterogeneity. Clinical judgment criteria involved assessment of similarity of populations, assessment context (eg, test performed a priori), study design (eg, case-control vs case based) and method in which specific tests were applied. In addition, after approval using clinical judgment, studies were statistically pooled when ≥2 studies examined the same index test and diagnosis with the same reference standard.

**Risk of bias/quality assessment**

Each of the full text articles was independently reviewed by two reviewers (MPR, CEC) and scored with the Quality Assessment of Diagnostic Accuracy Studies 2 scores (QUADAS-2) tool. Disagreements among the reviewers were discussed and resolved during a consensus meeting. The QUADAS-2 is a quality assessment tool comprised of four domains: patient selection, index test, reference standard, and flow and timing. The risk of bias is assessed in each of the domains, while the first three domains are also assessed for applicability by indicating a ‘low’, ‘high’ or ‘unclear’ rating. Applicability in the QUADAS-2 refers to whether certain aspects of an individual study are matching or not matching the review question. Unlike the QUADAS-1, the QUADAS-2 does not utilise a comprehensive quality score, rather an overall judgement of ‘low’, ‘high’ or ‘unclear’ risk. An overall risk rating of ‘low risk of bias’ or ‘low concern regarding applicability’ requires the study to be ranked as ‘low’ on all relevant domains. A ‘high’ or ‘unclear’ rating in one or more domains may require that the study be rated as an ‘at risk of bias’ or having ‘concerns regarding applicability.’

**Data extraction and analysis**

One reviewer (MPR) independently extracted information and data regarding study population, setting, special test performance, pathology, diagnostic reference-standard and number of true positives, false positives, false negatives and true negatives for calculation of sensitivity (SN), specificity (SP), positive likelihood ratios (+LR) and negative likelihood ratios (–LR) when not provided. SN is defined as the percentage of people who test positive for a specific pathology among a group of people who have the pathology. SP is the percentage of people who test negative for a specific pathology among a group of people who do not have the pathology. A positive likelihood ratio (+LR) is the ratio of a positive test result in people with the pathology to a positive test result in people without the pathology. A +LR identifies the strength of a test in determining the presence of a finding, and is calculated by the formula: SN/(1–SP). A negative likelihood ratio (–LR) is the ratio of a negative test result in people with the pathology to a negative test result in people without the pathology, and is calculated by the formula: (1–SN)/SP. The higher the +LR and lower the –LR the greater the post-test probability is altered. If analysed independently, tests that demonstrate high SN and low –LR are useful in ruling out a condition (screening). In contrast, tests that demonstrate high SP and high +LR assist in ruling in a condition (confirmation).

It has been suggested that post-test probability can be altered to a minimal degree with +LRs of 1–2 or –LRs of 0.5–1, to a small degree with +LRs of 2–5 or –LRs of 0.2–0.5, to a moderate degree with +LRs of 5–10 and –LRs of 0.1–0.2) and to a large and almost conclusive degree with +LRs greater than 10 and –LRs less than 0.1. Pretest probability is defined as the probability of the target pathology before a diagnostic test result is known. It represents the probability that a specific patient, with a specific history, presenting to a specific clinical setting, with a specific symptom complex, has a specific pathology. The diagnostic odds ratio (DOR), defined as +LR/–LR, is a single indicator, independent of prevalence, which represents the ratio of the odds of positivity in disease relative to the odds or positivity in the non-diseased. The values for the DOR range from 0, indicating no test discrimination, to infinity with higher scores indicating better discrimination.

**Meta-analysis**

DerSimonian and Laird random effects models, which incorporate both between and within study heterogeneity, were used to produce summary estimates of SN, SP, +LR, –LR, and diagnostic DOR. We initially attempted to analyse these data with a bivariate/hierarchical receiver operator curve model for those diagnostic tests with at least four studies. However, these models failed to converge and therefore report findings from the random effects models. An I-squared value of >50% and Cochrane’s-Q p value of <0.10 were the criteria to indicate significant between-study heterogeneity, of SN and SP and likelihood ratios respectfully. Publication bias was not formally tested due to low power of the tests with limited included studies. No significant threshold effects were found using Spearman correlation coefficients. When a computational problem of empty cells existed, 0.5 was
added to all four cells as suggested by Cox.\textsuperscript{59} Summary receiver operating characteristic (SROC) curves were produced when ≥4 studies were pooled. These curves summarise the diagnostic performance as a single number; the area under the curve (AUC) with a standard error (SE).\textsuperscript{60} AUC ≥0.90–1.00 were considered excellent, ≥0.80–0.90 considered good, ≥0.70–0.80 considered fair, ≥0.60–0.70 considered poor and ≥0.50–0.60 fail. Another measure, the $Q^*$ index and accompanied SE, is an additional measured produced which is the point on the SROC curve closest to the ideal left top-left corner (where SN and SP meet).\textsuperscript{61} All analyses were conducted by one of the authors (AG), blinded to results of the search, inclusion/exclusion and study quality, in Meta-DiSc V.1.4.\textsuperscript{62}

RESULTS

Selection of studies

The systematic search through MEDLINE, CINAHL and EMBASE netted 340 abstracts, and 10 additional abstracts were identified through an extensive manual search. In total, 350 titles were initially retained after duplicates were removed. Abstract and full text review reduced the acceptable papers to 21 (figure 1 and table 1).

This review included 1335 participants/1398 hips across 21 studies, investigating 11 different clinical special tests (tables 2 and 3). The sample size of the studies ranged from 10\textsuperscript{63} to 241 participants (table 2).\textsuperscript{64} Eight of the studies were retrospective,\textsuperscript{63–73} nine were prospective\textsuperscript{56 47 4\textsuperscript{–79} and the study design was unclear in four of the studies (table 1).\textsuperscript{3 63 80 81}

Quality scores

Table 1 provides the overall risk of bias score, with only one study\textsuperscript{80} demonstrating low risk of bias. It was unclear whether data collection was retrospective or prospective in this study.

RESULTS OF INDIVIDUAL DIAGNOSTIC CLINICAL MEASURES

Thirteen studies investigated the FADDIR test,\textsuperscript{65–69 71–73 75–77 79 81} three studies investigated the Flex-IR test,\textsuperscript{63 64 74}

![Flow diagram for study inclusion. IR, internal rotation; MA, meta-analysis; FADDIR, flexion adduction internal rotation test; MRA, MR arthrogram; RS, reference standard.](http://bjsm.bmj.com/)

one study investigated the Bilateral lower extremity squat test,
three studies investigated the FABER test,
one study investigated the Scour test,
one study investigated the IR-pressure test,
one study investigated the Resisted SLR test,
one study investigated the Thomas test,
one study investigated the THIRD test,
one study investigated the IR-flexion-axial compression test,
and one study investigated the Trochanteric tenderness test.
Table 2 reports which tests each study investigated,
as well as the characteristics of the study participants and
the reference standard utilised. Table 3 reports the tests investigated,
their diagnostic accuracy, reference standard utilised and
the pretest to post-test probability change after use of the tests
in each study for each test investigated.

META-ANALYSIS
Table 4 provides diagnostic properties and total sample sizes of
the nine studies included in the meta-analysis. Four studies
(n=188 participants) investigated the use of the FADDIR test
with MR angiogram (MRA) reference standard, four studies
(n=319 participants) investigated the use of the FADDIR test
with surgery reference standard, and two studies (n=27 participants)
investigated the use of the Flex IR test with surgery reference standard.

Figures 2 and 3 illustrate the SROC curves for FADDIR test
with MRA reference standard and surgery reference standard,
respectively. This figure illustrates the relationship between SN and
1-SP (false-positive rate) for the included studies in pooled analyses.
The AUC was considered fair (AUC=0.76 (SE=0.19))
for the FADDIR with MRA reference and poor (AUC=0.65
(SE=1.0)) for FADDIR with surgery reference.
The point at which SN and SP were equal (Q*) was 0.70
(SE=0.19) the MRA reference standard and 0.61
(SE=0.81) for surgery reference standard.

Figure 4 illustrates pretest to post-test probability changes
utilising the FADDIR test with a surgery reference standard. Pretest
to post-test probability for SN remained unchanged. Pretest to
post-test probability demonstrated a notable shift, although the CIs
were quite large.

DISCUSSION
Our study examined the current literature investigating clinical
examination tests for FAI/ALT. Owing to study heterogeneity
and variable reference standards only two of the 11 different
tests qualified for meta-analysis (FADDIR and Flexion IR tests).
These, as well as most of the other tests, were predominantly
only SN and not SP. Only two tests (in 9 studies) qualified for
meta-analysis, and none of those tests significantly shifted
post-test probability of the diagnosis of FAI/ALT. In fact, none
of the tests investigated in our review is capable of significantly
altering post-test probability. Additionally, the studies investigating
all of the tests in our review, whether performed for the
purpose of determining test diagnostic accuracy or not, were at
risk for bias and low quality. A significant need for improved
quality studies investigating these clinical tests exists.

The purpose of performing clinical testing is to aid with
determination of whether a particular pathology exists or not.
When the clinician is appraising evidence about diagnostic tests
they should consider several key concepts: did the partici-
pating individuals present a diagnostic dilemma? how much will
different levels of the diagnostic test raise or lower the pretest
probability of disease? will the reproducibility of the test result
and its interpretation be satisfactory in my clinical setting? are
the results applicable to the patients in my practice? will the test
results change my management strategy? and will the patients
be better off as a result of the test?

Participants presenting to a clinicians practice with a diagnos-
tic dilemma require utilisation of tests with properties capable
of differentiating those with versus those without the disease. In
order to properly assess these participants it is necessary to
determine if they were drawn from a common group in which it
**Table 2** Demographics and characteristics of included studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Participants/mean duration of symptoms</th>
<th>Pathology</th>
<th>Reference standard</th>
<th>Reference standard SN/SP (95% CI)</th>
<th>Clinical test investigated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aprato et al</td>
<td>65 participants (average age range 23–25 between surgical dislocation and arthroscopy groups; 17 females)</td>
<td>FAI</td>
<td>MRA 1.5 T; Cam: $\alpha$ angle $&gt;$50°; local pincer: (+) crossover sign and (−) posterior wall sign and $\alpha$ angle $&lt;$50°; pincer: (+) crossover and posterior wall signs or coxa profunda and $\alpha$ angle of $&lt;$50°; mixed: local or global pincer associated with a cam</td>
<td>100/100 ($\alpha$ angle) 91 (NR)/86 (NR) (ALT)</td>
<td>▶ FADDIR test</td>
</tr>
<tr>
<td>Ayeni et al</td>
<td>76 consecutive prospectively enrolled participants/78 hips (mean age 38.3 years, 39 females) presenting to adult outpatient clinic over 3 month period.</td>
<td>FAI</td>
<td>MRI or MRA 1.5 T; $\alpha$ angle $&gt;$55° and/or loss of femoral head-neck offset &lt;9.0 mm</td>
<td>NR in study MRI (ALT): 70 (62 to 77)/82 (69 to 94) MRA (ALT): 83 (79 to 88)/57 (44 to 70)</td>
<td>▶ Bilateral LE squat (maximum depth)</td>
</tr>
<tr>
<td>Barton et al</td>
<td>68 participants (age range 17 to 60 years, 37 females) from prospective database of all patients undergoing joint-preserving surgery and all having (+) MRA for ALT.</td>
<td>FAI</td>
<td>MRA 1.5 T; $\alpha$ angle $&gt;$50.5°</td>
<td>NR in study MRA (ALT): 83 (79 to 88)/57 (44 to 70)</td>
<td>▶ FADDIR test</td>
</tr>
<tr>
<td>Beaule et al</td>
<td>30 participants/36 hips (mean age 40.7 years, 36 hips, 13 females) with painful, non-dysplastic hips. &gt;3 months</td>
<td>ALT</td>
<td>MRA 1.5 T</td>
<td>NR in study MRA (ALT): 83 (79 to 88)/57 (44 to 70)</td>
<td>▶ FADDIR test</td>
</tr>
<tr>
<td>Chan et al</td>
<td>30 participants (mean age 41 years, 13 females) with unclear intra-articular hip pain.</td>
<td>ALT, AVN</td>
<td>MRA 1.5 T; Arthroscopy (n=17)</td>
<td>97 (77 to 100)/25 (3 to 80)*</td>
<td>▶ Flexion-IR test</td>
</tr>
<tr>
<td>Domayer et al</td>
<td>49 participants/60 hips (mean age 28±10.2 years, 29 females) with clinical signs and symptoms suggestive of FAI</td>
<td>FAI</td>
<td>MRI 1.5 T; cam FAI $\alpha$ $&gt;$55°</td>
<td>NR in study MRI (ALT): 70 (62 to 77)/82 (69 to 94)</td>
<td>▶ FADDIR test</td>
</tr>
<tr>
<td>Hananouchi et al</td>
<td>69 participants/107 hips (83 painful hips, 24 non-painful hips) prospectively examined (mean age 57.2 years, 54 females) with various pathology</td>
<td>FAI, dysplasia</td>
<td>MRI 3 T; FAI: crossover sign, $\alpha$ angle $&gt;$50°, aspherical femoral head, or CEA $&gt;$40° Dysplasia: CEA $&lt;$25°</td>
<td>NR in study MRI (0.5 to 3 T) (ALT): 66 (59 to 73)/79 (67 to 91)</td>
<td>▶ FADDIR test</td>
</tr>
<tr>
<td>Keeney et al</td>
<td>101 consecutive participants/102 hips (mean age 37.6 years, 71 females) with clinical signs and symptoms suggestive of ALT. 21.6 months</td>
<td>Variable: ALT, chondral defect, synovitis</td>
<td>MRA 1.5 T</td>
<td>NR in study MRA (ALT): 83 (79 to 88)/57 (44 to 70)</td>
<td>▶ FADDIR test</td>
</tr>
<tr>
<td>Laude et al</td>
<td>97 subjects/100 hips (mean age 33.4 years, range 16 to 56 years, 47 females) with persistent hip pain</td>
<td>ALT</td>
<td>Arthroscopy</td>
<td>NA</td>
<td>▶ FADDIR test</td>
</tr>
<tr>
<td>Leunig et al</td>
<td>23 participants (mean age 40±2 years, 14 females) with ARS 3.5 years</td>
<td>Variable: ALT, dysplasia, arthritic changes</td>
<td>Surgery</td>
<td>NA</td>
<td>▶ FADDIR test</td>
</tr>
<tr>
<td>Martin et al</td>
<td>105 participants (mean age 42±15 years, 24 females) potential surgical candidates due to high suspicion of intra-articular pathology (from subjective history) 1.9 years</td>
<td>Variable: ALT, FAI, arthritic changes</td>
<td>Intra-articular Injection with &lt; or &gt;50% pain relief</td>
<td>NR in study</td>
<td>▶ FABER test ▶ FADDIR test ▶ Trochanteric tenderness test</td>
</tr>
<tr>
<td>Maslowski et al</td>
<td>50 participants (mean age 60.2 years, 30 females) with signs/symptoms suggestive of intra-articular pathology</td>
<td>Variable: ALT, FAI, arthritic changes, AVN</td>
<td>Intra-articular injection with $\geq$80% pain relief</td>
<td>NR in study</td>
<td>▶ FABER test ▶ Scour test ▶ IR with overpressure ▶ Resisted SLR test</td>
</tr>
</tbody>
</table>

*Continued*
<table>
<thead>
<tr>
<th>Study</th>
<th>Participants/mean duration of symptoms</th>
<th>Pathology</th>
<th>Reference standard</th>
<th>Reference standard SN/SP (95% CI)</th>
<th>Clinical test investigated</th>
</tr>
</thead>
<tbody>
<tr>
<td>McCarthy and Busconi&lt;sup&gt;3&lt;/sup&gt;</td>
<td>59 participants (mean age 37 years, 32 females) with refractory hip pain NR</td>
<td>ALT, synovitis, loose bodies, arthritic changes</td>
<td>Arthroscopy</td>
<td>NA</td>
<td>▶ Thomas test</td>
</tr>
<tr>
<td>Myrick and Nissen&lt;sup&gt;10&lt;/sup&gt;</td>
<td>100 participants (age range 10–40 years, gender NR) with signs/symptoms suggestive of ALT NR</td>
<td>ALT</td>
<td>Arthroscopy</td>
<td>NA</td>
<td>▶ THIRD test</td>
</tr>
<tr>
<td>Narvani et al&lt;sup&gt;11&lt;/sup&gt;</td>
<td>18 participants (mean age 30.5±8.5 years, 5 females) with groin pain NR</td>
<td>ALT</td>
<td>MRA 1.0 T</td>
<td>NR in study MRA (0.5 to 3 T) (ALT): 87 (84 to 90)/64 (54 to 74)&lt;sup&gt;86&lt;/sup&gt;</td>
<td>▶ IR-flexion-axial compression test</td>
</tr>
<tr>
<td>Nogier et al&lt;sup&gt;14&lt;/sup&gt;</td>
<td>241 participants (mean age 35±10 years, 111 females) with mechanical hip pain All &gt;4 months</td>
<td>Dysplasia, FAI</td>
<td>AP pelvic; false profile and lateral axial radiographs; arthroscopic or arthrogram prescribed at investigator’s discretion</td>
<td>NR in study</td>
<td>▶ Flexion-IR test</td>
</tr>
<tr>
<td>Peters et al&lt;sup&gt;21&lt;/sup&gt;</td>
<td>94 participants/96 hips (mean age 28 years, 40 females) with signs/symptoms suggestive of FAI NR</td>
<td>FAI, articular cartilage lesions/arthritic changes</td>
<td>Open surgical dislocation</td>
<td>NA</td>
<td>▶ FADDIR test</td>
</tr>
<tr>
<td>Petersilge et al&lt;sup&gt;13&lt;/sup&gt;</td>
<td>10 participants (mean age 38.4 years, 5 females) with unexplained intra-articular hip pain NR</td>
<td>ALT, dysplasia, ganglion cyst</td>
<td>Arthroscopy</td>
<td>NA</td>
<td>▶ Flexion-IR test</td>
</tr>
<tr>
<td>Sink et al&lt;sup&gt;22&lt;/sup&gt;</td>
<td>35 participants (mean age 16 years, 30 females) with signs/symptoms suggestive of FAI 3 months to 3 years</td>
<td>Variable: ALT, FAI, cartilage damage</td>
<td>MRI/MRA, AP pelvic radiograph</td>
<td>NR in study MRI (0.5 to 3 T) (ALT): 66 (59 to 73)/79 (67 to 91)&lt;sup&gt;86&lt;/sup&gt;</td>
<td>▶ FADDIR test</td>
</tr>
<tr>
<td>Troelsen et al&lt;sup&gt;79&lt;/sup&gt;</td>
<td>18 participants (age range 32 to 56 years, 16 females) with previous PAO NR</td>
<td>Dysplasia, ALT</td>
<td>MRA</td>
<td>MRA (0.5 to 3 T) (ALT): 87 (84 to 90)/64 (54 to 74)&lt;sup&gt;86&lt;/sup&gt;</td>
<td>▶ FABER test, ▶ FADDIR test</td>
</tr>
<tr>
<td>Wang et al&lt;sup&gt;23&lt;/sup&gt;</td>
<td>21 participants (mean age 37.1 years, range 17 to 65 years, 12 females) with definite pain around hip joint, predominantly in the groin 12.4 months</td>
<td>ALT, FAI</td>
<td>Cross table and frog leg radiographs: cam FAI= α angle &gt;50°; pincer=cova profunda, cross-over sign or CEA &gt;40°</td>
<td>NR in study</td>
<td>▶ FADDIR test</td>
</tr>
</tbody>
</table>

*Values calculated by authors of this study.
α, alpha angle; ARS, acetabular rim syndrome; AVN, avascular necrosis; CEA, center edge angle; Dx, diagnosis; FABER, flexion, abduction, external rotation test; FADDIR, flexion-adduction-internal rotation impingement test, IR, internal rotation; LE, lower extremity; MRA, MR angiogram; NA, not applicable; NR, not reported; PAO, periacetabular osteotomy surgery.
<table>
<thead>
<tr>
<th>Test, authors</th>
<th>Participants</th>
<th>SN/SP (95% CI)</th>
<th>+LR/−LR</th>
<th>PPV/NPV</th>
<th>Reference standard</th>
<th>Pre to post-test probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anterior impingement—labral tear (FADDIR) test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aprato et al</td>
<td>41</td>
<td>99 (88 to 99)/6 (1 to 40)*</td>
<td>1.10/0.22*</td>
<td>82.1</td>
<td>MRA (FAI)</td>
<td>Pretest=80% Post-test=82.1%</td>
</tr>
<tr>
<td>Barton et al</td>
<td>68</td>
<td>99 (90 to 100)/2 (0 to 16)*</td>
<td>1.00/0.59*</td>
<td>63 (51 to 74)/50 (6 to 95)</td>
<td>MRA (α angle &gt;50.5°) for FAI</td>
<td>Pretest=62.1% Post-test=63%</td>
</tr>
<tr>
<td>Beaule et al</td>
<td>30</td>
<td>99 (88 to 100)/25 (3 to 80)* (labral tear)</td>
<td>1.30/0.06* (labral tear)</td>
<td>96 (84 to 99)/50 (5 to 95)</td>
<td>MRA (labral tear)</td>
<td>Pretest=93.4% Post-test=95.9%</td>
</tr>
<tr>
<td>Domayer et al</td>
<td>49/60 hips</td>
<td>99 (90 to 100)/3 (0 to 24)*</td>
<td>1.00/0.35*</td>
<td>75 (62 to 84)/50 (5 to 95)</td>
<td>MRA (labral tear)</td>
<td>Pretest=73.4% Post-test=74.6%</td>
</tr>
<tr>
<td>Hananouchi et al</td>
<td>69/107 hips</td>
<td>56 (37 to 73)/83 (31 to 98)*</td>
<td>3.30/0.53*</td>
<td>97 (75 to 100)</td>
<td>MRI (FAI)</td>
<td>Pretest=92.6% Post-test=96.7%</td>
</tr>
<tr>
<td>Keeney et al</td>
<td>101</td>
<td>99 (94 to 100)/2 (0 to 18)*</td>
<td>1.00/0.3*</td>
<td>76 (66 to 84)/50 (5 to 95)</td>
<td>MRI (labral tear)</td>
<td>Pretest=75.3% Post-test=76.1%</td>
</tr>
<tr>
<td>Laude et al</td>
<td>97/100 hips</td>
<td>99 (95 to 100)/6 (0 to 40)*</td>
<td>1.10/0.09*</td>
<td>93 (86 to 96)/50 (5 to 95)</td>
<td>Arthroscopy (labral tear)</td>
<td>Pretest=91.7% Post-test=92.6%</td>
</tr>
<tr>
<td>Leunig et al</td>
<td>23</td>
<td>97 (79 to 100)/13 (1 to 60)*</td>
<td>1.10/0.23*</td>
<td>84 (64 to 94)/50 (5 to 95)</td>
<td>Surgery (labral tear)</td>
<td>Pretest=80.4% Post-test=84.1%</td>
</tr>
<tr>
<td>Martin et al</td>
<td>105</td>
<td>78 (59 to 89)/10 (3 to 29)</td>
<td>0.86/2.3</td>
<td>53 (38 to 67)/25 (7 to 59)</td>
<td>Fre-arthroscopic injection (FAI and labral tear)</td>
<td>Pretest=42.8% Post-test=52.5%</td>
</tr>
<tr>
<td>Peters et al</td>
<td>94/96 hips</td>
<td>99 (95 to 100)/3 (0 to 25)*</td>
<td>1.00/0.18*</td>
<td>85 (77 to 91)/50 (5 to 95)</td>
<td>Surgery (FAI)</td>
<td>Pretest=84.2% Post-test=85.1%</td>
</tr>
<tr>
<td>Sink et al</td>
<td>35</td>
<td>99 (87 to 100)/17 (2 to 69)* (FAI)</td>
<td>1.20/0.09* (FAI)</td>
<td>57 (39 to 73)/50 (5 to 95)</td>
<td>AP radiograph (pincer FAI)</td>
<td>Pretest=55.0% Post-test=56.9%</td>
</tr>
<tr>
<td>Troelsen et al</td>
<td>18</td>
<td>58 (36 to 78)/75 (20 to 97)*</td>
<td>2.30/0.56*</td>
<td>96 (68 to 100)/17 (4 to 49)</td>
<td>MRI (labral tear)</td>
<td>Pretest=94.4% Post-test=95.3%</td>
</tr>
<tr>
<td>Wang et al</td>
<td>21</td>
<td>96 (73 to 99)/6 (1 to 37)*</td>
<td>1.00/0.7*</td>
<td>61 (41 to 79)/50 (5 to 95)</td>
<td>Radiographs (labral tear)</td>
<td>Pretest=58.7% Post-test=61.4%</td>
</tr>
<tr>
<td>Bilateral LE squat (maximum depth)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ajeni et al</td>
<td>76</td>
<td>75 (58 to 87)/41 (28 to 56)*</td>
<td>1.30/0.61*</td>
<td>47.1 (34.1, 60.5)/70.4 (51.5, 84.1)</td>
<td>MRI or MRA (FAI)</td>
<td>Pretest=41% Post-test=47.1%</td>
</tr>
<tr>
<td>FABER test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maslowski et al</td>
<td>50</td>
<td>52.5 (37.5 to 67.1)/54.5 (28 to 78.7)</td>
<td>1.10/0.72</td>
<td>80.8 (62.1, 91.5)/24 (11.5, 43.4)</td>
<td>Fre-arthroscopic injection (FAI and labral tear)</td>
<td>Pretest=78.0% Post-test=80.8%</td>
</tr>
<tr>
<td>Martin et al</td>
<td>105</td>
<td>60 (41 to 77)/18 (7 to 39)</td>
<td>0.73/2.2</td>
<td>45.5 (29.8, 62.0)/28.6 (11.7, 54.6)</td>
<td>Fre-arthroscopic injection (FAI and labral tear)</td>
<td>Pretest=31.9% Post-test=45.5%</td>
</tr>
<tr>
<td>Troelsen et al</td>
<td>18</td>
<td>42 (22 to 64)/75 (20 to 97)*</td>
<td>1.70/0.78*</td>
<td>93.8 (59.8, 99.3)/12.5 (2.9, 40.2)</td>
<td>MRI (labral tear)</td>
<td>Pretest=94.4% Post-test=93.8%</td>
</tr>
<tr>
<td>Flexion-internal rotation test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chan et al</td>
<td>30</td>
<td>98 (84 to 100)/8 (1 to 48)*</td>
<td>1.10/0.23*</td>
<td>82.3 (65.6, 91.9)</td>
<td>MRI (labral tear)</td>
<td>Pretest=79.7% Post-test=82.3%</td>
</tr>
<tr>
<td>Chan et al</td>
<td>17</td>
<td>97 (77 to 100)/25 (3 to 80)*</td>
<td>1.30/0.12*</td>
<td>91.7 (70.6, 98.1)/50 (5.5, 95.5)</td>
<td>Arthroscopy (labral tear)</td>
<td>Pretest=86.8% Post-test=91.7%</td>
</tr>
<tr>
<td>Nogier et al</td>
<td>241</td>
<td>70 (62 to 77)/44 (34 to 54)</td>
<td>1.30/0.69</td>
<td>67.9 (60.3, 74.8)/46.4 (36.2, 57.0)</td>
<td>Radiographs (FAI)</td>
<td>Pretest=44.2% Post-test=67.9%</td>
</tr>
<tr>
<td>Petersilge et al</td>
<td>10</td>
<td>95 (66 to 100)/25 (3 to 80)</td>
<td>1.30/0.20</td>
<td>86.4 (57.1, 96.8)/50 (5.5, 94.5)</td>
<td>Arthroscopy (labral tear)</td>
<td>Pretest=79.2 Post-test=86.4%</td>
</tr>
</tbody>
</table>

Continued
### Table 3

<table>
<thead>
<tr>
<th>Test, authors</th>
<th>Participants</th>
<th>SN/SP (95% CI)</th>
<th>+LR/−LR</th>
<th>PPV/NPV</th>
<th>Reference standard</th>
<th>Pre to post-test probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal rotation-flexion-axial compression test</td>
<td>Narvani et al&lt;sup&gt;5&lt;/sup&gt;</td>
<td>18</td>
<td>75 (19 to 99)/43 (18 to 72)*</td>
<td>1.3/0.58*</td>
<td>27.3 (9.7, 56.6)/85.7 (48.7, 97.4)</td>
<td>MRA (labral tear)</td>
</tr>
<tr>
<td>Internal rotation with overpressure</td>
<td>Maslowski et al&lt;sup&gt;76&lt;/sup&gt;</td>
<td>50</td>
<td>91 (68 to 99)/18 (5 to 40)</td>
<td>1.1/0.5</td>
<td>45.5 (31.7, 59.9)/71.4 (35.9, 91.8)</td>
<td>Intra-articular injection (FAI and labral tear)</td>
</tr>
<tr>
<td>Resisted straight leg raise test</td>
<td>Maslowski et al&lt;sup&gt;76&lt;/sup&gt;</td>
<td>50</td>
<td>58 (34 to 79)/29 (11 to 54)</td>
<td>0.87/1.28</td>
<td>40.6 (25.5, 57.7)/50 (29, 71)</td>
<td>Intra-articular injection (FAI and labral tear)</td>
</tr>
<tr>
<td>Scour test</td>
<td>Maslowski et al&lt;sup&gt;76&lt;/sup&gt;</td>
<td>50</td>
<td>50 (26 to 74)/29 (12 to 51)</td>
<td>0.70/1.72</td>
<td>35.5 (21.1, 53.1)/42.1 (23.1, 63.7)</td>
<td>Intra-articular injection (FAI and labral tear)</td>
</tr>
<tr>
<td>THIRD test</td>
<td>Myrick and Nissen&lt;sup&gt;70&lt;/sup&gt;</td>
<td>100</td>
<td>98 (93 to 100)/75 (19 to 99)</td>
<td>3.90/0.3*</td>
<td>98.9 (94.3, 99.8)/60 (23.1, 88.2)</td>
<td>Arthroscopy (labral tear)</td>
</tr>
<tr>
<td>Thomas test</td>
<td>McCarthy and Buscon&lt;sup&gt;3&lt;/sup&gt;</td>
<td>59</td>
<td>89 (73 to 97)/92 (75 to 99)*</td>
<td>11.1/0.12*</td>
<td>94 (80 to 99)/86 (67 to 96)*</td>
<td>Arthroscopy (labral tear)</td>
</tr>
<tr>
<td>Trochanteric tenderness (not tender)</td>
<td>Martin et al&lt;sup&gt;77&lt;/sup&gt;</td>
<td>105</td>
<td>57 (39 to 74)/45 (27 to 65)</td>
<td>1.1/0.93</td>
<td>55.6 (37.7, 72.4)/47.6 (28.3, 67.6)</td>
<td>Intra-articular injection (FAI and labral tear)</td>
</tr>
</tbody>
</table>

*Calculations performed by authors of this study.

ARS, acetabular rim syndrome; F, female; FABER, flexion, abduction, external rotation test; FADDIR, flexion-adduction-internal rotation impingement test; LE, lower extremity; +LR, positive likelihood ratio; −LR, negative likelihood ratio; MRA, MR angiogram; NA, not applicable; NPV, negative predictive value; NR, not reported; OST, orthopaedic special tests; PPV, positive predictive value; SN, sensitivity (%); SP, specificity (%); VAS, visual analogue scale.

### Table 4

<table>
<thead>
<tr>
<th>Measure</th>
<th># studies (sample size)</th>
<th>SN (95% CI)</th>
<th>SP (95% CI)</th>
<th>+LR (95% CI)</th>
<th>ue−LR (95% CI)</th>
<th>DOR (95% CI)</th>
<th>PPV (95% CI)</th>
<th>Post-test change</th>
</tr>
</thead>
<tbody>
<tr>
<td>FADDIR (MRA)</td>
<td>4 (65, 68, 79, 81) (n=188)</td>
<td>0.94 (0.90 to 0.97)</td>
<td>0.09 (0.02 to 0.23)</td>
<td>1.02 (0.96 to 1.08)</td>
<td>0.45 (0.19 to 1.09)</td>
<td>5.71 (0.84 to 38.86)</td>
<td>0.833 (0.771 to 0.885)</td>
<td>Pretest=84% Post-test=83%</td>
</tr>
<tr>
<td>FADDIR (surgery)</td>
<td>4 (68, 69, 71, 76) (n=319)</td>
<td>0.99 (0.98 to 1.00)</td>
<td>0.05 (0.01 to 0.18)</td>
<td>1.04 (0.97 to 1.12)</td>
<td>0.14 (0.02 to 0.93)</td>
<td>7.82 (1.06 to 57.84)</td>
<td>0.90 (0.89 to 0.90)</td>
<td>Pretest=90% Post-test=90%</td>
</tr>
<tr>
<td>Flexion IR</td>
<td>2 (63, 74) (n=27)</td>
<td>0.96 (0.81 to 0.99)</td>
<td>0.25 (0.01 to 0.81)</td>
<td>1.28 (0.72 to 2.27)</td>
<td>0.15 (0.01 to 1.99)</td>
<td>8.36 (0.41 to 171.3)</td>
<td>0.90 (0.73 to 0.98)</td>
<td>Pretest=87% Post-test=90%</td>
</tr>
</tbody>
</table>

DerSimonian-Laird random-effects models used throughout.

DOR, diagnostic OR; FADDIR, flexion, adduction, internal rotation test; +LR, positive likelihood ratio; −LR, negative likelihood ratio; SN, sensitivity; SP, specificity.
is not known whether the condition of interest is present or absent. Participants in all 21 studies of this review were of high suspicion for various types of intra-articular joint pathology, most that were of high suspicion for FAI/ALT due to groin pain and/or other subjective symptoms (eg, clicking, catching) that are highly suggestive of these pathologies. The pretest probability in the investigated tests ranged from 17% to 94%, but was higher than 55% in all but five studies involving 10 of the 28 tests listed in table 3. Certainly, a high pretest probability can influence the likelihood of reporting a positive finding on the tests and this may be a reason we found high SN in the tests that were eligible for meta-analysis.

The single largest increase in post-test probability (in studies of low bias) after use of a particular test was only 6.1% when utilising the painful squat test by Ayeni et al. Most other tests investigated in this review provided very small percentage increases in post-test probability. In fact, in meta-analysis calculations (table 4), there were minimal to no increases in post-test probabilities, ranging from 1% decline in probability when using the FADDIR test (MRA reference standard), no change using the FADDIR test (surgical reference standard; figure 2), and 3% increase using the Flexion IR test with MRA reference standard. While the post-test probability could change significantly with, for example, the FADDIR test to rule out the pathology existing, the wide range CIs limit the accuracy of interpretation of these findings (figure 4). Additionally, having these test performed on participants of high suspicion of pathology results in scarce true negative findings in these studies.

Ascertaining results of the tests investigated in this review strongly suggests limitations in their clinical utility. The results of these tests appear to minimally change the clinician’s treatment strategy, and the participants do not appear to be better off as a result of performing these tests. The current scope of literature investigating these tests is narrow in its focus (examining...
participants with high suspicion of pathology prior to test performance) and, therefore, results in test results of high SN, poor SP and limited post-test probability in determining the diagnosis of FAI/ALT. In fact, a clinician practising in an orthopaedic/sports clinic seeing participants with high suspicion of FAI/ALT would benefit minimally by performing these tests.

An additional concern regarding diagnostic accuracy studies is the potential for bias. Only one of the included studies in this review had low risk of bias. This study investigated a newly described test (bilateral lower extremity squat to maximum depth). Using a sample size of 76 participants, a SN of 75 and a SP of 41, this test was found to also provide the greatest shift in pretest to post-test probability (6.1%). As mentioned previously, the large majority of these tests had much greater pretest probabilities prior to test implementation than this test. This study utilised the more potentially biased reference standard of MRI and/or MRA.

Although MRI/MRA is rightfully not considered the ‘gold standard’ for these studies, the use of surgery alone leads to biased participant sampling. Determination of the clinical utility of these tests requires their implementation in future studies across the spectrum of participants with and without undetermined hip pathology. The reference standard for these studies would thus be the more imprecise MRI/MRA. Although the diagnostic accuracy of MRI/MRA currently is limited, limiting image reading to those with specialist training, utilisation of precise, tailored protocols and improving imaging technology could afford suitability for MRI/MRA as acceptable reference standards for these participants. Implementation of these tests across this broad continuum of hip participants is paramount to determine their actual clinical utility.

Limitations

Limitations of this study include: limiting the search strategy to only those articles written in English, one author pulling the data points, lack of comparison of participant inclusion and exclusion across the studies and studies performed in settings of high pretest probability. Limiting to English language could potentially miss some studies to be included in our review. Only one author pulling the data points increases the risk of potential error, although this author (MPR) has pulled many of these data points for a similar review of both intra-articular and extra-articular hip pathology. Since many of the included studies were published prior to prospective guidelines for diagnostic accuracy studies, the robustness of reporting inclusion/exclusion criteria was highly variable among the studies and very difficult to comprehensively describe in this review. Having these studies performed in hip surgeon clinical settings is a limitation of the current literature that is uncontrollable by the authors of this review.

CONCLUSION

Owing to the low quality and biased sampling of patients with high probability of disease, hip physical examination tests do not appear to currently provide the clinician any significant value in altering probability of disease with their use. Currently, only the FADDIR and Flex-IR tests are supported by the data as valuable screening tests for FAI/LT pathology. Further studies involving high quality designs across a wider spectrum of hip pathology patients are necessary to discern the confirmed clinical utility of these tests.

Acknowledgements

The authors would like to thank Leila Ledbetter, MLIS for assisting with the literature search for this study.

Contributors

MPR developed the idea of manuscript. All the authors contributed in writing, editing and approval of final manuscript.

Competing interests

None.

Provenance and peer review

Not commissioned; externally peer reviewed.

REFERENCES


Diagnostic accuracy of clinical tests for the diagnosis of hip femoroacetabular impingement/labral tear: a systematic review with meta-analysis

M P Reiman, A P Goode, C E Cook, P Hölmich and K Thorborg

doi: 10.1136/bjsports-2014-094302

Updated information and services can be found at:
http://bjsm.bmj.com/content/49/12/811

These include:

Supplementary Material
Supplementary material can be found at:
http://bjsm.bmj.com/content/suppl/2014/12/16/bjsports-2014-094302.DC1.html

References
This article cites 82 articles, 17 of which you can access for free at:
http://bjsm.bmj.com/content/49/12/811#BIBL

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections
Articles on similar topics can be found in the following collections

BJSM Reviews with MCQs (97)
Labral tears (12)

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/