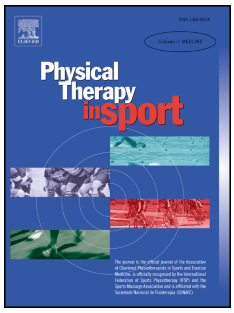
Accepted Manuscript



Physical Examination Tests For The Diagnosis Of Femoroacetabular Impingement. A Systematic Review

Aitana Pacheco-Carrillo, Ivan Medina-Porqueres, PT, RN, MSc., Physical Therapy Section

PII: S1466-853X(16)00003-1

DOI: 10.1016/j.ptsp.2016.01.002

Reference: YPTSP 701

To appear in: *Physical Therapy in Sport*

Received Date: 19 February 2015

Revised Date: 16 December 2015

Accepted Date: 14 January 2016

Please cite this article as: Pacheco-Carrillo, A., Medina-Porqueres, I., Physical Examination Tests For The Diagnosis Of Femoroacetabular Impingement. A Systematic Review, *Physical Therapy in Sports* (2016), doi: 10.1016/j.ptsp.2016.01.002.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

1. **Physical Examination Tests For The Diagnosis Of Femoroacetabular Impingement. A Systematic Review**

2. Aitana Pacheco-Carrillo and Ivan Medina-Porqueres, PT, RN, MSc. Physical Therapy Section.

3. Faculty of Health Sciences, University of Malaga. C/ Arquitecto Francisco Peñalosa. Ampliación Campus Teatinos 29071 Málaga (Spain)

4. Faculty of Health Sciences, University of Malaga. C/ Arquitecto Francisco Peñalosa. Ampliación Campus Teatinos 29071 Málaga (Spain)

5. Aitana Pacheco Carrillo, 657743671, aitanapacheco@gmail.com (it can be published)

6. 3794 words

7. 3 tables, 2 supplemental file and 1 figure.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Physical Examination Tests For The Diagnosis Of Femoroacetabular Impingement. A Systematic Review

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

ABSTRACT

Numerous clinical tests have been proposed to diagnose FAI, but little is known about their diagnostic accuracy.

Objectives: To summarize and evaluate research on the accuracy of physical examination tests for diagnosis of FAI.

Methods: A search of the PubMed, SPORTDiscus and CINAHL databases was performed. Studies were considered eligible if they compared the results of physical examination tests to those of a reference standard. Methodological quality and internal validity assessment was performed by two independent reviewers using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool.

Results: The systematic search strategy revealed 298 potential articles, five of which articles met the inclusion criteria. After assessment using the QUADAS score, four of the five articles were of high quality. Clinical tests included were Impingement sign, IROP test (Internal Rotation Over Pressure), FABER test (Flexion-Abduction-External Rotation), Stinchfield/RSRL (Resisted Straight Leg Raise) test, Scour test, Maximal squat test, and the Anterior Impingement test. IROP test, impingement sign, and FABER test showed the most sensitive values to identify FAI.

Conclusions: The diagnostic accuracy of physical examination tests to assess FAI is limited due to its heterogenecity. There is a strong need for sound research of high methodological quality in this area.

Key words: Femoroacetabular impingement; Diagnostic accuracy; Hip pain; Physical examination.

ACCEPTED MANUSCRIPT

INTRODUCTION

Femoroacetabular Impingement (FAI) is an abnormal anatomical relationship between the femoral head and/or femoral neck and the acetabulum. It produces a premature contact between both structures during coxofemoral joint movement which may lead to early degeneration of the labrum and adjacent cartilage. This continued damage over time may modify the lubrication and normal behaviour of the hip joint and, consequently, may alter the function of the sealed joint. Initially introduced by Ganz et al. (2003), who classified the FAI into two groups: first type, called Cam, in which the presence of a bony prominence (hump) in the femoral head-neck alters the femoral head sphericity. It correlates directly with early osteoarthritis of the young-adult male. The second type. known as Pincer, has a normal femoral neck junction but an over covered acetabular rim. It is more common in middle-aged active women (Banerjee & Mclean, 2011). There is also a combination of both types with a slight predominance of one of them (70 % of the cases) (Marín et al., 2008).

Interest on FAI has increased recently given that its existence was unknown until a few years ago (Ganz et al., 2003; Leunig et al., 2009). Recent studies have underlined FAI as one of the most important causes of labral tear and it is a recognized factor of hip early osteoarthritis, especially in young, adult active patients (Beck et al., 2005; Leunig et al., 2009). Besides, FAI has been recognized as a common cause of hip pain in this type of patients (Byrd, 2006), with 72% of men and 50% of women showing some evidence of radiographic hip abnormality consistent with FAI (Gerhardt et al., 2012), where a proper diagnosis is essential. According to Martin et al. (2010) and Tibor and Sekiya (2008) patient history and a thorough physical examination play a very important role when identifying hip pain of intraarticular origin. In short, an early diagnosis is necessary for early and effective treatment and to avoid the need of a hip replacement implant in a young adult patient.

Accurate clinical testing should facilitate timely and appropriate intervention for patients suffering from hip pain and suspected FAI. Thus, a lack of consensus on diagnostic criteria and concordance in clinical assessment may difficult the choice of intervention. Many academic texts, narrative reviews, and on-line material exist to describe examination techniques, including special tests, specifically conceived for detecting hip pathology. Despite the acceptable number of clinical tests proposed to

diagnose FAI, these tests have not been compared for their accuracy and, consequently, no single test have been identified as superior to another. Previous reviews have sought to assess the diagnostic test accuracy of hip physical examination tests as a whole(Rahman et al., 2013; Reiman, Goode, Hegedus, Cook, & Wright, 2013), but no one has specifically focused on the diagnosis of FAI. The purpose of this systematic review was to determine the diagnostic accuracy of selected clinical tests for FAI and investigate the quality of the studies that have examined these values.

METHODS

Study Design

This systematic review used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines during the search and reporting phase of the research process. The PRISMA statement includes a 27-item checklist designed to improve reporting of systematic reviews and meta-analyses (Liberati et al., 2009). The PRISMA guidelines were created for use in summarizing randomized controlled trials but can be used for multiple forms of research methodologies (Swartz, 2011).

Search Strategy

A systematic search of relevant literature was conducted on March 1, 2014, and the search strategy results were monitored until December 1, 2014. A comprehensive search was conducted using three databases: PubMed, SPORT Discus and CINAHL (through EBSCO), and hand-searching reference lists of retrieved articles. The search terms and resultant hits were summarized in supplemental file 1.

Study Selection

A total of 298 studies were reviewed by the two authors (APC and IMP) independently to identify studies that addressed the diagnostic accuracy of clinical tests. Full-text articles were retrieved if the abstract provided insufficient information to establish eligibility or if the article passed the first eligibility screening. Any discrepancies were settled by further discussion and consensus.

Elegibility criteria

Diagnostic studies were eligible if they included: 1) a description of a clinical test or tests used for diagnosing FAI (including a test that was combined with another finding such as patient history); 2) a report or allow computation of the diagnostic accuracy of the tests (e.g., sensitivity and specificity, positive and negative predictive values, and positive and negative likelihood ratios; 3) an acceptable reference standard for comparison; 4) a goal (among others) to specifically investigate which clinical diagnostic tests are available for diagnosis of FAI and the diagnostic accuracy of these tests.

If a paper failed to provide any of the next criteria: 1) clinical measure and description; 2) report of diagnostic accuracy; and 3) appropriate reference standard, that paper was not included in this review. Further any studies were excluded if they involved imaging without clinical diagnostic tests, reported the value of a non-specific clinical examination, or did not use adequate reference standard.

Data Extraction

Data extraction was performed by one reviewer (APC) and verified by a second (IMP). Data extracted included: 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, specificity, positive predictive values, negative predictive values, positive likelihood ratios, and negative likelihood ratios, when these were not provided. *Sensitivity* is defined as the percentage of people who test positive for a specific disease among a group of people who have the disease, whereas *specificity* is the percentage of people who test negative for a specific disease among a group of people who do not have the diagnosis/disorder. The *Positive Predictive Value* (PPV) is the percentage of patients with a positive test who actually have the disease. The *Negative Predictive Value* (NPV) is the percentage of patients with a negative test who do not have the disease (Jaeschke et al., 1994; Parikh et al., 2008). 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 *Sensitivity/(1 – Specificity).* 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 – Sensitivity)/Specificity*. Therefore, a diagnostic test will be most useful if the LR+ value is greater and the LR- value is smaller (Grimes & Schulz, 2005).

Sensitivity, specificity, and predictive values of at least 80% were considered to be sufficient. Also, a LR+ of 10 and a LR- of 0.1 were considered to be sufficient (Guyatt et al., 2008). More specifically, results of the LR+ with values of 1-2 are considered useless to confirm a diagnosis; 2-10 is considered a moderate test, being a good test when it presents values between 10-50 and excellent test when it is greater than 50. However, results of the LR- with values of 1-0.5 are considered useless to exclude a diagnostic test; between 0.5-0.1 is considered a moderate test, being a good test when it presents values between 0.1 to 0.02 and excellent test when it is less than 0.02 (Cook et al., 2007; Jaeschke et al., 1994).

Quality assessment

Quality Assessment of Diagnostic Accuracy Studies

All articles meeting inclusion criteria for selection were reviewed and assessed for quality and risk of bias. Each article was independently assessed by both authors (APC and IMP) using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool. Disagreements among the reviewers were discussed and resolved through consensus. The QUADAS consists of 14 items, each with response categories of “yes”, “no”, or “unclear”. A “yes” score indicates sufficient information, with bias considered unlikely; a “no” score indicates sufficient information, but with potential bias from inadequate design or conduct; and an “unclear” score indicates that the article or methodology provided insufficient information or the methodology was unclear. The total score was a count of all of the criteria that scored “yes” (valued as 1, whereas “no” and “unclear” scores were valued as zero), with a maximum attainable score of 14. The methodological quality of each of the studies was assessed by the review. Qualitatively, studies that exhibit higher QUADAS values are associated with less risk of design bias than those with lower values. Similar to previously published reviews, the studies were stratified as “high quality/low risk of bias” if their QUADAS score was 10 or greater or as “low quality/high risk of bias” if their QUADAS score was less than 10(Whiting, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2003).

We measured agreement among reviewers using a weighted kappa with 95% confidence intervals for QUADAS scoring. Conventionally, a kappa of <0.2 is considered poor agreement, 0.21–0.4 fair, 0.41–0.6 moderate, 0.61–0.8 strong, and more than 0.8 near complete agreement(Shrout & Fleiss, 1979).

Strength of Evidence

The level of evidence was evaluated using the following definitions(van Tulder, Furlan, Bombardier, Bouter, & Editorial Board of the Cochrane Collaboration Back Review Group, 2003): strong (consistent findings among multiple high-quality randomised trials), moderate (consistent findings among multiple low-quality randomised trials and/or controlled clinical trials and/or one high-quality randomised trial), limited (one low-quality randomised trial and/or controlled clinical trial), conflicting (inconsistent findings among multiple, randomised and/or controlled clinical trials), and no evidence from trials (no randomised or controlled clinical trial).

RESULTS

Selection of Studies

The systematic search through PubMed, SPORT Discus and CINAHL netted 293 abstracts, and 5 additional papers were identified through an extensive hand search. In total, 298 titles were initially retained after duplicates were removed. Abstract and full-text review reduced the acceptable papers to five (Figure 1).

Description of Included Studies

Details on study characteristics are provided in Table 1. Clinical tests investigated, participants, reference standard used, and diagnoses made by authors were included. A total of seven different clinical special tests across five papers were included in this review, assessing 609 hips/505 participants. Only two studies (Hananouchi et al., 2012; Troelsen et al., 2009***)*** evaluated the diagnostic accuracy of clinical tests in hospital setting, defined here as a referral to an orthopaedic surgery department or presentation to an accident and emergency department. The other three

studies evaluated the accuracy of clinical tests in outpatient orthopaedic clinics (Ayeni et al., 2014; Maslowski et al., 2010; Nogier et al., 2010).

The reference standard varied throughout the five studies. In one study the reference standard was MRI-Arthrogram (Troelsen et al., 2009), two studies applied radiography, joined to physical examination (Nogier et al., 2010) or MRI (Hananouchi et al., 2012), and one study applied a 80% improvement of pain on 10-cm VAS after intra-articular hip injection or 80% pain relief (Maslowski et al., 2010). One study applied MRI or MRI-A indistinctly (Ayeni et al., 2014). Both MRI and MRI-A have shown to be reliable tools for diagnosing CAM deformity, with a specificity of 100% and a sensitivity ranging from 79 to 100% (Aprato et al., 2013; González Gil, Llombart Blanco, & Díaz de Rada, 2015). However, there is no previously reported data on VAS improvement regarding FAI.

The size of the study population varied widely, ranging from 18 to 292 subjects. There was also variability between participants in the included studies with respect to average age (35–60.2 years), and proportion of males (11%-62%).

Two studies investigated the Anterior impingement test (Hananouchi et al., 2012; Troelsen et al., 2009), one study evaluated the Impingement sign (Nogier et al., 2010), two studies investigated the FABER test (Maslowski et al., 2010; Troelsen et al., 2009), two studies evaluated the Stinchfield test –also known as RSLR test– (Maslowski et al., 2010; Troelsen et al., 2009), one study assessed the IROP test (Maslowski et al., 2010), one study investigated the Scour test (Maslowski et al., 2010), and one study evaluated the Maximal Squat test (Ayeni et al., 2014).

Diagnostic studies are primarily cross-sectional studies and were labeled as case-control-type or cohort-type accuracy studies to avoid confusion regarding the epidemiologic definitions of case-control or cohort studies, respectively (Rutjes et al., 2005). Of the five studies included, all of them were considered cohort-type accuracy studies.

Diagnostic Accuracy

Data on the diagnostic accuracy of individual clinical tests for FAI in terms of sensitivity, specificity, predictive values, and likelihood ratios are presented in table 3. In all cases 95% confidence intervals were calculated. Sensitivity ranged from 0.2 to

0.91 and specificity ranged from 0.17 to 1.00. Positive Predictive Values ranged from 0.36 to 1.00, whereas Negative Predictive Values ranged from 0.08 to 0.71. The FABER test was the only one that showed sufficient values in terms of sensitivity, specificity and PPV, being greater than 0.8. The Anterior Impingement test could only be considered sufficient in terms of specificity and PPV. In addition, IROP test and impingement sign test showed, respectively, a sensitivity and a specificity greater than 0.8. None of the eight tests included in the review showed a NPV greater than 0.8.

Likelihood ratios were calculated from extracted data. Positive likelihood ratios ranged from 0 to 1.55, and, in most cases, the confidence intervals were very large. Negative likelihood ratios ranged from 0.11 to 1.72. Thus, none of the eight tests showed LR+ greater than 2.00 and only two tests presented a LR- less than 0.50. According to LR- data, only the Anterior Impingement test and the IROP test could be considered a moderate test. However, these values are considered too poor to confirm the diagnosis of clinical tests evaluated.

Quality Scores

Quality of Diagnostic Accuracy Studies

The assessment of the five articles retained for this review indicated that only one article was of low quality/high risk of bias. The remaining four articles had QUADAS score between 10 and 12 out of 14 points. Quality scores for each of the studies are synthesized in Table 2. Using the previously established stratification of the QUADAS, Nogier et al. article was considered of low quality/high risk of bias, with a score of 6 points, whereas Ayeni et al. (2014), Malowski et al. (2010), Troelsen et al. (2009), and Hananouchi et al. (2012) articles had a QUADAS score of 10-12 points, suggesting high quality/low risk of bias. The most poorly scored items of the QUADAS were item 1 (spectrum representative of those in clinical practice), item 3 (reference standard likely to classify the target condition correctly), item 4 (the period of time between the reference standard and index test is acceptable), item 10 (index test results interpretation without knowledge of reference standard), and item 11 (reference standard interpretation without knowledge of index test results). The agreement for the independent scoring of the QUADAS yielded a weighted kappa of 0.811 (95% CI: 0.7, 0.9), with a strength of agreement considered to be near completion.

Strength of Evidence

Based on the review of all the available studies, there is limited evidence to support the use of FAI tests as stand-alone clinical assessment for the diagnosis of this hip condition.

DISCUSSION

Systematic reviews of evaluation of tests are undertaken for the same reasons as the systematic reviews of treatment intervention to produce estimates of test performance and impact based on all available evidence, in order to evaluate the quality of published studies and to account for variation in the findings between studies (Deeks, 2001). This study investigated the diagnostic accuracy and quality of seven selected clinical tests for FAI. There were only five studies illustrating tests that included both sensitivity and specificity values. When reported PPV, NPV, and likelihood ratios were also presented. Remarkably, there was only one instance in which one study investigated a single test.

Some tests had moderate-to-high sensitivity and others moderate-to-high specificity, but overall no FAI physical examination test appeared to be of value in terms of modifying post-test probability and enabling the diagnostic process. Based on our results, the IROP test and the FABER test seem to be the most sensitive tests to be need to help in the diagnosis of FAI. None of the remaining tests reached a 0.8 sensitivity or specificity, which suggests that their diagnostic utility for use in clinical practice is questionable.

The confidence intervals for the sensitivity, specificity, PPV, and NPV, when reported, were very wide in nearly all studies suggesting a lack of precision in the findings. It is also worth noting that in some cases, the 95% confidence interval was close to 1.00, which indicates that the result of the given physical examination test is no better than chance. This variability in reporting and lack of consistent diagnostic accuracy values between studies suggest that a meta-analysis should be performed, though this action was beyond the scope of this review.

We performed our electronic literature search in three databases (PubMed, SPORTDiscus and CINAHL), consistent with recommendations to search in more than just one database for diagnostic test accuracy studies (Whiting et al., 2008). We decided to exclude PEDro database, in contrast to authors of other diagnostic test accuracy reviews (Cook et al., 2012), because this database contains few diagnostic studies (Sherrington et al., 2000). According to this, we are confident that all relevant articles on this issue have been identified.

The quality of the reviewed papers varied widely with scores ranging from 6 to 12 of 14 (Table 2). Of the five articles, four met our prestudy definition of “high quality (Cook et al., 2007), which was a QUADAS score of ten or greater. Studies with QUADAS scores below ten have been suggested to provide biased results and potentially inflated or deflated diagnostic accuracy values (Cook et al., 2012). It should be noted that the majority of studies failed to report if the period of time between the reference standard and index test was acceptable.

Studies of test accuracy should ideally compare the test results between groups of patients with and without the target disease, each of whom undergoes the experimental test as well as the reference gold standard. An ideal study testing a predictor of FAI should have, among other requirements, a well-defined population, prospective and consecutive recruitment, blinding of those involved in assessing the test results and outcomes, adequate test description, defining normal and abnormal test before starting the study and comparing it with the gold standard of FAI. The description of a test within a study should be sufficient to enable replication of the test by researchers. The description should include the exact details of the test’s application and the criteria used to determine positive and negative findings (Fritz & Wainner, 2001). Accordingly, if the test is performed and/or interpreted differently, the demonstrated accuracy of the test cannot be evaluated against any other test for comparison. This review detected certain discrepancies in procedure description. A positive finding in FABER test was described as provoked groin pain (Troelsen et al., 2009) or a recreation of subject´s pain (Maslowski et al., 2010), what do not necessarily have the same clinical meaning. The RSLR has also been described as Stinchfield maneuver, as referred, which could lead to misinterpretation. In accordance, the impingement test is described by Troelsen et al. (2009) as a passive move of the hip

joint in flexion, internal rotation, and adduction; this combination of movements is considered by Hananouchi et al. (2012) to be the anterior impingement test.

Some physical tests were the objective of previous studies, but results show that there was a lack of diagnostic accuracy parameters or these parameters had poor values. This was supported by the finding that, based on the QUADAS score, four of five diagnostic accuracy studies were of good quality. These four studies investigated the IROP test, the FABER test, the Stinchfield/RSLR test, the Scour test, the maximal squat test, and the Anterior Impingement test (Ayeni et al., 2014; Hananouchi et al., 2012; Maslowski et al., 2010; Troelsen et al., 2009). However, because of several methodological problems, none of these tests are appropriate to reliably confirm or discard the diagnosis of FAI. The first methodological issue is that in each of the five studies, there were some flaws that resulted in a lower strength of evidence. The number of subjects per study differed from 18 to 76 (Ayeni et al., 2014; Bellamy et al., 1984; Maslowski et al., 2010; Troelsen et al., 2009). Sample size in the study by Troelsen et al. (2009) is much smaller than it is in the other four studies (Ayeni et al., 2014; Hananouchi et al., 2012; Maslowski et al., 2010; Nogier et al., 2010). Although Nogier et al. (2010) investigated a fairly large sample of 292 patients, the study scored the lowest (6/14) of the five cases on the QUADAS. In contrast, the QUADAS scores for the Troelsen et al. (2009), the Maslowski et al. (2010), Ayeni et al. (2014), and Hananouchi et al. (2012) studies were higher, potentially resulting in a lower risk of bias. It is a group of subjects is too small to reliably interpret diagnostic accuracy. Furthermore, all 4 studies used a study population, in which there was a high suspicion of intra-articular hip pathology, increasing the risk of spectrum bias. These two flaws led to difficulties in interpretation of the diagnostic accuracy figures.

The strengths of this review lie in its systematic and comprehensive nature. A process of systematically identifying, screening and critically appraising the studies helps to ensure that the review process is transparent and replicable. In addition, this review includes the use of inclusion criteria to ensure that the study settings reflected clinical practice, and the application of likelihood ratios as they are the preferred approach to report estimates of diagnostic accuracy (Grimes & Schulz, 2005).

Results of this study should however be interpreted with consideration to a number of limitations. Firstly, this review was limited by the small number of studies

included (n=5), a relative common finding for studies that investigate diagnostic accuracy (Doustet al., 2005), which may have introduced bias. There were also language limitations, as foreign studies were not included which could again introduce bias (Song et al., 2010). This brings an obvious risk that studies were not identified through our search process. In addition, one study (Troelsen et al., 2009) did not investigate diagnostic accuracy values of the physical examination test of interest (RSLR) which disenabled the authors of the current study to estimate its usefulness. Lastly, quality and internal validity assessment of retrieved articles is considered to be an essential component of most systematic reviews (Deeks, 2001). The QUADAS tool was used to assess the quality of articles in this study; however, this and other well-designed tools are suggested to have a number of associated limitations. These include the possibility that even well-conducted studies may score poorly if the methods and results of the study are not reported in sufficient detail (Whiting et al., 2003).

This systematic review of clinical tests for FAI incorporates the most recent knowledge of diagnostic test accuracy methods. The question arises about the relevance of the results of our review and whether they can contribute to the improvement of the current practice. We can affirm that there are a limited number of studies and, therefore, tests that investigate the diagnostic accuracy of FAI and the diagnostic accuracy of these tests is quite variable. Most diagnostic studies on this topic contain methodological flaws which can overestimate the information obtained. The small sample size of most of the studies and the observed heterogeneity make generalisable conclusion difficult. Uniformity in test executions is desirable and these should be thoroughly investigated for diagnostic accuracy. Due to the inherent limitations of the cohort studies presented, health professionals must be cautious in interpreting the research results for use in clinical practice. There is a strong need for sound research of high methodological quality in this area.

ACCEPTED MANUSCRIPT

REFERENCES

Aprato, A., Massè, A., Faletti, C., Valente, A., Atzori, F., Stratta, M., & Jayasekera, N. (2013). Magnetic resonance arthrography for femoroacetabular impingement surgery: is it reliable? *Journal of Orthopaedics and Traumatology: Official Journal of the Italian Society of Orthopaedics and Traumatology*, *14*(3), 201-206.

Ayeni, O., Chu, R., Hetaimish, B., Nur, L., Simunovic, N., Farrokhyar, F., Bhandari, M. (2014). A painful squat test provides limited diagnostic utility in CAM-type femoroacetabular impingement. *Knee Surgery, Sports Traumatology, Arthroscopy: Official Journal of the ESSKA*, *22*(4), 806-811.

Banerjee, P., & Mclean, C. R. (2011). Femoroacetabular impingement: a review of diagnosis and management. *Current Reviews in Musculoskeletal Medicine*, *4*(1), 23-32.

Beck, M., Kalhor, M., Leunig, M., & Ganz, R. (2005). Hip morphology influences the pattern of damage to the acetabular cartilage: femoroacetabular impingement as a cause of early osteoarthritis of the hip. *The Journal of Bone and Joint Surgery. British Volume*, *87*(7), 1012-1018.

Bellamy, N., Newhook, L., Rooney, P. J., Brooks, P. M., Cockshott, W. P., Thompson, G. T., & Buchanan, W. W. (1984). Perception--a problem in the grading of sacro-iliac joint radiographs. *Scand J Rheumatol*, *13*(2), 113-20.

Byrd, J. W. T. (2006). Hip arthroscopy: surgical indications. *Arthroscopy: The Journal of Arthroscopic & Related Surgery: Official Publication of the Arthroscopy Association of North America and the International Arthroscopy Association*, *22*(12), 1260-1262.

Cook, C., Cleland, J., & Huijbregts, P. (2007). Creation and Critique of Studies of Diagnostic Accuracy: Use of the STARD and QUADAS Methodological Quality Assessment Tools. *The Journal of Manual & Manipulative Therapy*, *15*(2), 93-102.

Cook, C., Mabry, L., Reiman, M. P., & Hegedus, E. J. (2012). Best tests/clinical findings for screening and diagnosis of patellofemoral pain syndrome: a systematic review. *Physiotherapy*, *98*(2), 93-100.

Deeks, J. J. (2001). Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ (Clinical Research Ed.)*, *323*(7305), 157-162.

Doust, J. A., Pietrzak, E., Sanders, S., & Glasziou, P. P. (2005). Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. *Journal of Clinical Epidemiology*, *58*(5), 444-449.

Fritz, J. M., & Wainner, R. S. (2001). Examining diagnostic tests: an evidence-based perspective. *Phys Ther*, *81*(9), 1546-64.

Ganz, R., Parvizi, J., Beck, M., Leunig, M., Nötzli, H., & Siebenrock, K. A. (2003). Femoroacetabular impingement: a cause for osteoarthritis of the hip. *Clinical Orthopaedics and Related Research*, (417), 112-120.

Gerhardt, M. B., Romero, A. A., Silvers, H. J., Harris, D. J., Watanabe, D., & Mandelbaum, B. R. (2012). The prevalence of radiographic hip abnormalities in elite soccer players. *The American Journal of Sports Medicine*, *40*(3), 584-588.

González Gil, A. B., Llombart Blanco, R., & Díaz de Rada, P. (2015). Validity of magnetic resonance arthrography as a diagnostic tool in femoroacetabular

impingement syndrome. *Revista Española De Cirugía Ortopédica Y Traumatología*, *59*(4), 281-286.

Grimes, D. A., & Schulz, K. F. (2005). Refining clinical diagnosis with likelihood ratios. *Lancet*, *365*(9469), 1500-1505.

Guyatt G, Rennie D, Meade M, Cook M. Users' Guides to the Medical Literature. Chicago, IL: McGraw-Hill Professional; 2008. (s. f.).

Hananouchi, T., Yasui, Y., Yamamoto, K., Toritsuka, Y., & Ohzono, K. (2012). Anterior impingement test for labral lesions has high positive predictive value. *Clinical Orthopaedics and Related Research*, *470*(12), 3524-3529.

Jaeschke, R., Guyatt, G., & Sackett, D. L. (1994). Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA*, *271*(5), 389-391.

Leunig, M., Beaulé, P. E., & Ganz, R. (2009). The concept of femoroacetabular impingement: current status and future perspectives. *Clinical Orthopaedics and Related Research*, *467*(3), 616-622.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of Clinical Epidemiology*, *62*(10), e1-34.

Marín, O., Ribas, M., Ledesma, R., Tey, M., Valles, A., & Vilarubias, J. M. (2008). Atrapamiento o choque femoroacetabular: Concepto Diagnóstico y Tratamiento. Parte 1. Arch Med Dep 2008; XXV (124): 128-33.

Martin, H. D., Shears, S. A., & Palmer, I. J. (2010). Evaluation of the hip. *Sports Medicine and Arthroscopy Review*, *18*(2), 63-75.

Maslowski, E., Sullivan, W., Forster Harwood, J., Gonzalez, P., Kaufman, M., Vidal, A., & Akuthota, V. (2010). The diagnostic validity of hip provocation maneuvers to detect intra-articular hip pathology. *PM & R: The Journal of Injury, Function, and Rehabilitation, 2*(3), 174-181.

Nogier, A., Bonin, N., May, O., Gedouin, J.-E., Bellaiche, L., Boyer, T., French Arthroscopy Society. (2010). Descriptive epidemiology of mechanical hip pathology in adults under 50 years of age. Prospective series of 292 cases: Clinical and radiological aspects and physiopathological review. *Orthopaedics & Traumatology, Surgery & Research: OTSR*, *96*(8 Suppl), S53-58.

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, *56*(1), 45-50.

Rahman, L. A., Adie, S., Naylor, J. M., Mittal, R., So, S., & Harris, I. A. (2013). A systematic review of the diagnostic performance of orthopedic physical examination tests of the hip. *BMC Musculoskeletal Disorders*, *14*, 257.

Reiman, M. P., Goode, A. P., Hegedus, E. J., Cook, C. E., & Wright, A. A. (2013). Diagnostic accuracy of clinical tests of the hip: a systematic review with meta-analysis. *British Journal of Sports Medicine*, *47*(14), 893-902.

Rutjes, A. W. S., Reitsma, J. B., Vandenbroucke, J. P., Glas, A. S., & Bossuyt, P. M. M. (2005). Case-control and two-gate designs in diagnostic accuracy studies. *Clinical Chemistry*, *51*(8), 1335-1341.

Sherrington, C., Herbert, R. D., Maher, C. G., & Moseley, A. M. (2000). PEDro. A database of randomized trials and systematic reviews in physiotherapy. *Manual Therapy*, *5*(4), 223-226.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*, *86*(2), 420-8.

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J.,Harvey, I. (2010). Dissemination and publication of research findings: an updated review of related biases. *Health Technology Assessment (Winchester, England)*, *14*(8), iii, ix-xi, 1-193.

Swartz, M. K. (2011). The PRISMA statement: a guideline for systematic reviews and meta-analyses. *Journal of Pediatric Health Care: Official Publication of National Association of Pediatric Nurse Associates & Practitioners*, *25*(1), 1-2.

Tibor, L. M., & Sekiya, J. K. (2008). Differential diagnosis of pain around the hip joint. *Arthroscopy*, *24*(12), 1407-21.

Troelsen, A., Mechlenburg, I., Gelineck, J., Bolvig, L., Jacobsen, S., & Søballe, K. (2009). What is the role of clinical tests and ultrasound in acetabular labral tear diagnostics? *Acta Orthopaedica*, *80*(3), 314-318.

van Tulder, M., Furlan, A., Bombardier, C., Bouter, L., & Editorial Board of the Cochrane Collaboration Back Review Group. (2003). Updated method guidelines for systematic reviews in the cochrane collaboration back review group. *Spine*, *28*(12), 1290-1299.

Whiting, P., Rutjes, A. W. S., Reitsma, J. B., Bossuyt, P. M. M., & Kleijnen, J. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*, *3*, 25.

Whiting, P., Westwood, M., Burke, M., Sterne, J., & Glanville, J. (2008). Systematic reviews of test accuracy should search a range of databases to identify primary studies. *Journal of Clinical Epidemiology*, *61*(4), 357-364.

Table 1. Characteristics of included studies.

Author	Test evaluated	Subjects (gender, age)	Diagnoses made by Authors	Reference Standard
Nogier et al. (2010)	Impingement sign	n = 292 (111 females/181 males) mean age, 35 yr (range, 16-50 yr)	FAI	Complete physical examination with radiography (PINCER-type→ crossover sign or acetabular protrusion; CAM-type→ femoral head bump, anterosuperior neck flatness or ovoid head -on AP or lateral axial view-).
Maslowski et al. (2010)	IROP test FABER test Scour test Stinchfield test*	n = 50 (30 females/20 males) mean age, 60.2 yr (range, 22-84 yr)	Labral tear, FAI	80% improvement of pain on 10-cm VAS after intra-articular hip injection or 80% pain relief
Troelsen et al. (2009)	Anterior Impingement test FABER test RSLR test*	n = 18 (16 females/2 males) mean age, 43 yr (range, 32-56 yr)	Labral pathology	MRI-A
Hananouchi et al. (2012)	Anterior Impingement test	n = 69 (54 females/15 males) mean age, 57.3 yr (range, 27-81 yr)	Labral lesions (FAI)	MRI and radiography (presence of crossover sign, α angle >50°, asphericity femoral head -pistol grip deformity-, or center-edge angle >40°)
Ayeni et al. (2014)	Maximal Squat test	n = 76 (39 females/37males) mean age, 38.3 yr	CAM FAI Isolated labral tear Minimal OA	MRI-A and MRI (CAM-type→ α angle >55.0° on axial oblique sequences and/or loss of femoral head–neck offset <9.0 mm)

Abbreviations: *MRI-A: Magnetic Resonance Imaging-Arthrogram; VAS: Visual Analog Scale; IROP test: Internal Rotation Over Pressure Test; RSLR test: Resisted Straight Leg Raise Test; *Stinchfield test: RSLR test.*

ACCEPTED MANUSCRIPT

Table 2. Quality Assessment of the Studies Included in the Review, according to QUADAS tool (Whiting, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2003).

Author (Year of Publication)	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Final Score
Nogier et al. (2010)	N	Y	N	U	Y	Y	N	N	N	Y	Y	Y	N	N	6
Maslowski et al. (2010)	N	Y	N	U	Y	Y	Y	Y	Y	U	U	Y	Y	Y	10
Troelsen et al. (2009)	N	Y	N	U	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	11
Hananouchi et al. (2012)	Y	Y	Y	U	Y	Y	Y	Y	Y	N	N	Y	Y	Y	11
Ayeni et al. (2014)	N	Y	Y	U	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	12

Abbreviations: *N, no; U, unclear; Y, yes.*

***Item 1:** was the spectrum of patients representative of those in clinical practice? **Item 2:** were selection criteria clearly described? **Item 3:** is the reference standard likely to classify the target condition correctly? **Item 4:** is the period of time between the reference standard and index test acceptable? **Item 5:** did the whole sample of patients receive verification using the reference standard? **Item 6:** did patients receive the same reference standard regardless of the index test result? **Item 7:** was the reference standard independent of the index test? **Item 8:** was the execution of the index test described in sufficient detail for replication? **Item 9:** was the execution of the reference standard described in sufficient detail for replication? **Item 10:** were the index test results interpreted without knowledge of the reference standard? **Item 11:** was the reference standard interpreted without knowledge of the results of the index test? **Item 12:** were the same clinical criteria available when test results were interpreted as would be in clinical practice? **Item 13:** were uninterpretable/intermediate test results reported? **Item 14:** were withdrawals from the study explained?

ACCEPTED MANUSCRIPT

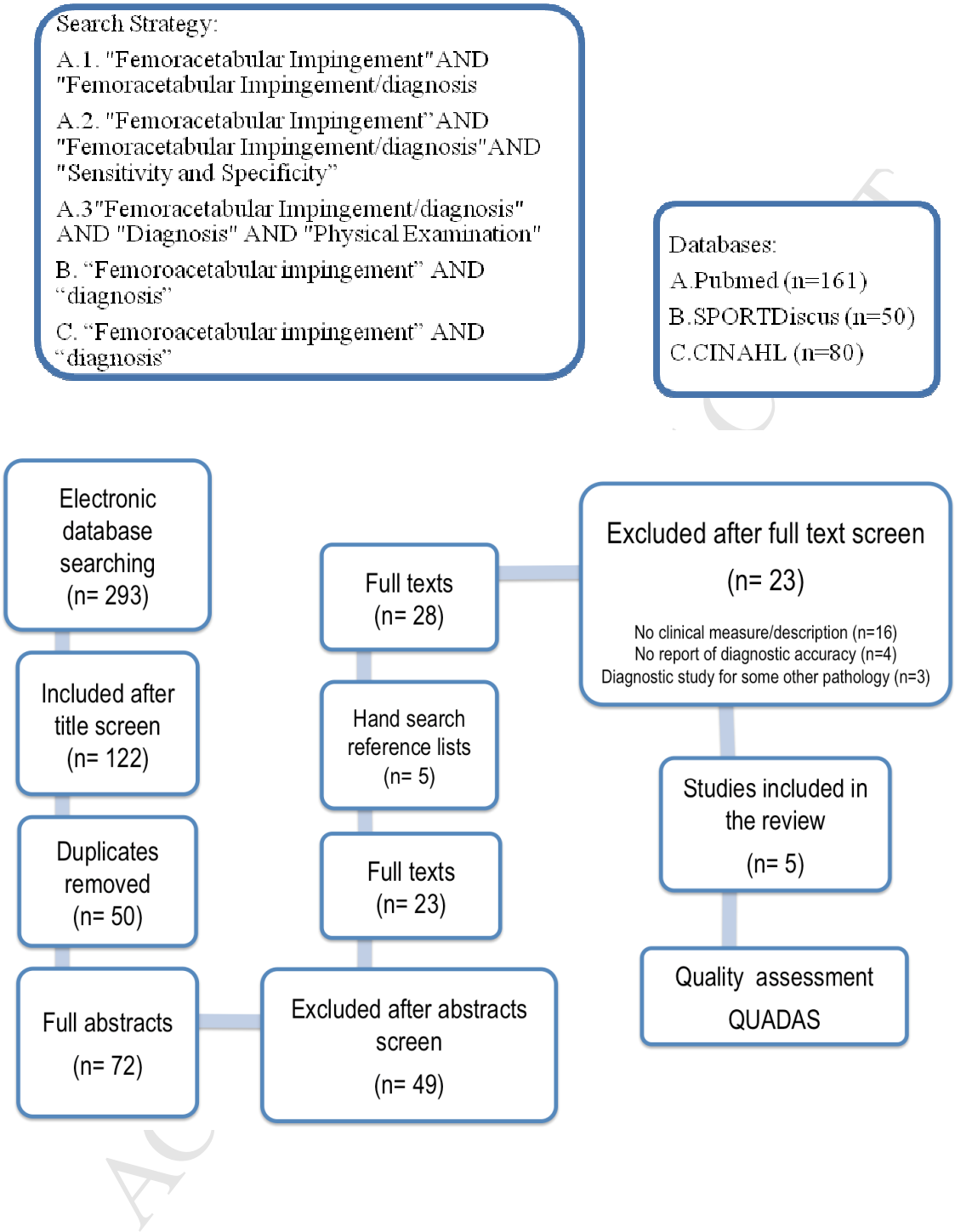
Table 3. Diagnostic accuracy of physical examination tests for femoroacetabular impingement. PPV, Positive Predictive Value; NPV, Negative Predictive Value; LR+, Positive Likelihood Ratio; LR-, Negative Likelihood Ratio.

Author	Test	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	LR+ (95% CI)	LR- (95% CI)	QUADAS score
Nogier et al.	Impingement sign: pain predominating in flexion/internal rotation	0.20	0.44	0.63	0.53	0.36	1.82	6
Nogier et al.	Impingement sign: pain exclusively in flexion/internal rotation	0.20	0.86	0.67	0.44	1.43	0.93	6
Nogier et al.	Impingement sign: reduced pain-free flexion amplitude under internal rotation	0.51	0.67	0.67	0.51	1.55	0.73	6
Maslowski et al.	IROP test (at least 80% VAS Relief as reference standard)	0.91 (.68-.99)	0.18 (.05-.40)	0.47 (.29-.64)	0.71 (.25-.98)	1.11 -	0.11 -	10
Maslowski et al.	IROP test (at least 80% Perceived Relief as reference standard)	0.88 (.67-.98)	0.17 (.04-.40)	0.54 (.36-.71)	0.57 (.15-.92)	0.05 -	0.71 -	10
Maslowski et al.	FABER test (at least 80% VAS Relief as reference standard)	0.82 (.57-.96)	0.25 (.09-.48)	0.46 (.28-.65)	0.64 (.27-.91)	1.09 -	0.72 -	10
Maslowski et al.	FABER test (at least 80% Perceived Relief as reference standard)	0.81 (.58-.95)	0.25 (.08-.50)	0.54 (.35-.72)	0.55 (.20-.86)	1.08 -	0.76 -	10
Maslowski et al.	Scour test (at least 80% VAS Relief as reference standard)	0.50 (.26-.74)	0.29 (.12-.51)	0.36 (.17-.57)	0.42 (.18-.69)	0.82 -	1.72 -	10
Maslowski et al.	Scour test (at least 80% Perceived Relief as reference standard)	0.62 (.38-.82)	0.38 (.17-.62)	0.52 (.31-.72)	0.47 (.22-.74)	1.00 -	1.00 -	10
Maslowski et al.	Stinchfield test (at least 80% VAS Relief as reference standard)	0.59	0.32	0.41	0.50	0.87	1.28	10

ACCEPTED MANUSCRIPT

		(.34-.82)	(.14-.55)	(.22-.62)	(.23-.77)	-	-	
Maslowski et al.	Stinchfield test (at least 80% Perceived Relief as reference standard)	0.58	0.29	0.47	0.39	0.82	1.45	10
		(.34-.79)	(.11-.54)	(.27-.68)	(.15-.67)	-	-	
Troelsen et al.	Anterior Impingement test	0.59	1	1	0.13	0	0.41	11
Troelsen et al.	FABER test	0.41	1.00	1.00	0.08	0	0.59	11
Troelsen et al.	RSLR test	-	-	-	-	-	-	11
Hananouchi et al.	Anterior Impingement test	0.56	1.00	1.00	0.15	0	0.44	10
Ayeni et al.	Maximal Squat test	0.75	0.41	0.47	0.70	1.27	0.61	12
		(.56–.88)	(.27–.57)	(.90–1.7)	(.30–1.2)	-	-	

Figure 1. Flow diagram of search strategy.



ACCEPTED MANUSCRIPT

HIGHLIGHTS

- Clinicians are encouraged to use clinical examination to identify FAI.
- A comprehensive literature search about physical examination tests for FAI was done.
- Diagnostic accuracy of FAI clinical tests is limited due to its heterogenecity.
- More methodologically sound studies on this topic are needed.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Acknowledgements: The authors express their appreciation to Mrs. Maria A. Raya, Librarian, for her assistance in gathering background material and designing the searching method.

ACCEPTED MANUSCRIPT