Diagnostic accuracy of clinical tests assessing ligamentous injury of the talocrural and subtalar joints: a systematic review with meta-analysis

Word count: 4265 words

ABSTRACT

CONTEXT

Ankle sprains are the most common acute musculoskeletal injury. Clinical tests represent the first opportunity to assess the sprain's severity, but no systematic review has compared these tests to contemporary reference standards.

OBJECTIVE

To determine the diagnostic accuracy of clinical tests assessing the talocrural and subtalar joint ligaments after ankle sprain.

DATA SOURCES

CINAHL, Embase, MEDLINE, hand-searching and PubMed related article searches (inception to November 18 2020).

STUDY SELECTION

Eligible diagnostic studies compared clinical examination (palpation, joint laxity) against imaging or surgery. Studies at a high risk of bias or with high concerns regarding applicability on Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) were excluded from the meta-analysis.

STUDY DESIGN

Systematic review and meta-analysis

LEVEL OF EVIDENCE

3a

DATA EXTRACTION

True positive, false negative, false positive and true negative findings were extracted to calculate sensitivity, specificity, and likelihood ratios. If ordinal data was reported, these were extracted to calculate Cohen’s kappa.

RESULTS

14 studies met inclusion criteria (6302 observations; nine clinical tests). No test had both sensitivity and specificity exceeding 90%. Palpation of the anterior talofibular ligament is highly sensitive (sensitivity 95–100%; specificity 0–32%; min-max; n = 6) but less so for the calcaneofibular ligament (sensitivity 49–100%; specificity 26–79%; min-max; n = 6). Pooled data from six studies (885 observations) found a low sensitivity (54%; 95% confidence interval 35–71%) but high specificity (87%; 95% confidence interval 63–96%) for the anterior drawer test.

CONCLUSION

The anterior talofibular ligament is best assessed using a cluster of palpation (rule out), and anterior drawer testing (rule in). The talar tilt test can rule in injury to the calcaneofibular ligament, but a sensitive clinical test for the ligament is lacking. It is unclear if ligamentous injury grading can be done beyond the binary (injured vs uninjured), and clinical tests of the subtalar joint ligaments are not well-researched. The generalisability of our findings is limited by insufficient reporting on blinding and poor study quality.

FUNDING

None.

REGISTRATION

Prospero ID: CRD42020187848

KEYWORDS
Diagnosis; Ankle; Examination; Ligament; Meta-analysis

INTRODUCTION

Each year, over 300,000 people present to UK Emergency Departments with ankle sprain (~800 per day).[5] Many occur during sporting or recreational activity, due to excessive inversion and internal rotation of the ankle at high velocity.[28] Ankle sprains are often regarded as innocuous injuries, but up to 70% of patients develop chronic ankle instability; characterised by mechanical laxity, subjective feelings of giving way, persistent pain and reinjury.[28] In the UK, the total average cost associated with a lateral ankle sprain is estimated at £940.[10] The high incidence of chronic symptoms, risk of recurrence, and long term risk of developing post-traumatic osteoarthritis, further contribute to the significant socioeconomic burden of lateral ankle sprains.[28]

Limited data inform the causality of chronic ankle instability.[4] An emerging hypothesis is that poor prognosis after ankle sprain is mediated by inadequate clinical examination. The primary concerns are that existing clinical tests often fail to identify microinstabilities of the ankle joint complex; which consists of the anterior talofibular ligament (ATFL), calcaneofibular ligament (CFL), and the posterior talofibular ligament (PTFL).[23] Also, few tests target the primary stabilisers of the subtalar joint; consisting of the interosseous talocalcaneal ligament (ITCL), cervical ligament (CL), and the anterior capsular ligament (ACL). Recommendations for clinical examination of suspected lateral ligamentous injury continue to be underpinned by palpation and manual stress tests (eg. anterior drawer and talar tilt).[13] However, only two reviews[56,57] have systematically reported their diagnostic accuracy. The most recent review[56] included just five studies, with the majority limited to arthrographic (stress radiography) reference standards.

We must re-examine the diagnostic utility of clinical examination techniques in this field, by also including contemporary reference standards (ultrasound, MRI, and arthroscopy).[7] Diagnostic accuracy may be optimised through test clustering, and through the inclusion of new index tests (such as modified drawer tests), but this has not been systematically examined. A key part of clinical examination should be to differentiate isolated vs combined injuries of the talocrural and subtalar joints, and use this to determine prognosis, or guide management decisions. MRI and arthroscopy can consistently identify concomitant damage to primary stabilisers of the subtalar joint, but it is unclear if clinical tests have comparable diagnostic utility.

METHODS

Protocol and registration

We used the Preferred Reporting Items for Systematic Reviews and Meta-Analysis of Diagnostic Test Accuracy Studies (PRISMA-DTA)[47] for our review.

We prospectively drafted our study protocol to PROSPERO on May 20 2020, registration ID: CRD42020187848

Eligibility criteria

We assessed original research for eligibility using the criteria presented in Table 1, with no restrictions on the language of the article nor the publication year. Most criteria were decided on a priori, as part of the PROSPERO protocol. However, arthroscopy as an inclusion criterion was extended to include other surgical techniques as well, and avulsion fractures as an exclusion criterion were omitted; to broaden the eligibility criteria.

Table 1. PICOTS criteria for inclusion and exclusion of studies

Parameter	Inclusion criteria	Exclusion criteria
Population	Ankle sprain	Fractures
Index test	Any clinical test aiming to reproduce symptoms or assess joint stability	Surgical or imagery stress tests, testing delivered under anaesthesia
Comparator	Arthrogram, Arthroscopy, Magnetic Resonance Imaging, Stress X-ray, Surgery or Ultrasound	
Outcome measure	Ascertain the presence or absence of ligamentous ankle injury	Studies with insufficient information to compute a 2x2 contingency table to calculate sensitivity and specificity
Type of study	Prospective cohort, diagnostic case-control studies or retrospective studies.	Cadaveric studies, case series, systematic reviews
Setting	Any setting	

Search

We conducted electronic database searching of EBSCOhost and Ovid: searching CINAHL, Embase and MEDLINE from inception to November 18 2020. We used the same search terms for all three databases. We also performed PubMed related article searches for all studies meeting inclusion criteria from the previous database searches. Finally, we examined the references of our included studies and previous systematic reviews. Our search strategy and the number of hits for MEDLINE can be seen in Figure 1.

[FIGURE 1]

Study selection

Two reviewers (???, ??) independently screened the title and abstract of every identified record. Afterwards, both reviewers presented their respective articles, and both reviewers examined the full-text versions separately. If full-text articles contained insufficient information to decide eligibility, we contacted the corresponding authors for additional details. Disagreements regarding final inclusion were fully resolved through consensus (??, ??) without the need for a third reviewer (??). After inclusion criteria had been met for our systematic review, we also considered each article for meta-analysis. We excluded retrospective and case-controlled studies from the meta-analysis, due to the risk of these study designs to overestimate diagnostic accuracy. We also excluded studies at a high risk of bias or with high concerns regarding applicability from the meta-analysis.

Risk of bias in individual studies

Two reviewers (???, ??) performed an independent methodological assessment of the included studies, using the Quality Assessment for Diagnostic Accuracy Studies version 2 (QUADAS-2)[68] tool. There are four domains to QUADAS-2: 1) Patient selection. Ideally, all eligible patients should be consecutively enrolled, all with a suspected injury relevant to the research question. Convenience sampling, case-control designs and inappropriate exclusions risk introducing bias in the form of overestimated measures of diagnostic accuracy, as the patient spectrum is not representative of clinical practice. 2) Index test. To minimise the risk of bias, index testing should be interpreted without knowledge of reference test results. Also, the conduct of the index test should be sufficiently described to permit replication, as deviations in execution could affect the generalisability of the findings. 3) Reference standard. Since estimates of diagnostic test accuracy are based on the presumption that the discriminatory properties of the reference standard are perfect, the sensitivity and specificity of the reference standard must be sufficient to correctly diagnose the presence or absence of the injury in question. The reference standard should also be interpreted without prior knowledge of the index test. 4) Flow and timing. Both the index test and the reference standard should be delivered as close in time to each other as possible. A prolonged time-span risk introducing confounding effects from intermediate interventions or regression to the mean, thus leading to non-valid study findings.[55,65] After we had performed independent quality assessments, a consensus meeting followed, during which we reached full agreement.

Data items

Information regarding study setting (e.g. private, public, sports, primary care, emergency department); study design (prospective, retrospective, case-control); population demographics (age, gender, level of sporting participation, time since injury); details of index tests and reference standards (testing protocol, the definition of a positive test outcome, flow and timing) were extracted independently and in duplicate into a predefined form by two reviewers (???, ??). The extracted information was then reviewed and confirmed by a third reviewer (??), who compared the completed forms to each other and the original research reports.

Synthesis of results

We produced 2x2 contingency tables based on the true positive, false positive, true negative, and false negative findings of the included studies. With this information, we used Review Manager 5.4 software[9] to compute sensitivity and specificity values and their respective 95% confidence intervals (CI). Sensitivity values are representative of the proportion of those with injury correctly classified as injured, whilst specificity values are representative of the proportion of those without injury correctly classified as healthy.

All contingency table data kept in Review Manager 5.4 was also exported and analysed in OpenMetaAnalyst, to produce I^2 statistics and assess between-study variability in sensitivity and specificity.

If ordinal level data were reported, these were extracted and analysed to see if clinical tests can accurately grade the degree of injury. We calculated the inter-rater agreement between index test and reference test with weighted Cohen's kappa (linear weighting), using an online calculator.[26] According to McHugh,[46] kappa values for agreement are to be interpreted as: 0–20 = none; 21–39 = minimal; 40–59 = weak; 60–79 = moderate; 80–90 = strong; > 90 = almost perfect

All data extracted into Review Manager 5.4 was done independently and in duplicate by two reviewers (???, ??). A third reviewer (??) verified the extracted data by comparing the results between the two reviewers (???, ??) and by cross-referencing against the original

research reports. If discrepancies were noticed between the two reviewers responsible for data extraction, the third reviewer decided what data to present. The primary author (???) then performed all statistical analyses.

Meta-analysis

We performed HSROC and bivariate meta-analyses with MetaDTA 2.0 software.[18,50] We calculated pooled summary estimates of test sensitivity, specificity, and positive and negative likelihood ratios (LR), each with 95% CI. Likelihood ratios are considered a useful diagnostic metric and represent the prevalence of positive tests in those with injury versus those without (LR+) and the prevalence of negative tests in those that are healthy versus those that are not (LR-).[12] We plotted the pooled LRs in Fagan's nomogram,[17] to examine the change in pre to post-test probability after positive and negative tests. We estimated the pretest probability through the median disease prevalence of studies eligible for meta-analysis. To determine heterogeneity, we used the Cochran Q test ($p < 0.05$ indicating presence of heterogeneity) and the *I*-squared statistic. *I*-squared values of 0-40%, 30-60%, 50-90%, and 75-100% were considered non-important, moderate, substantial, and significant levels of heterogeneity, respectively.[30] This univariate analysis of heterogeneity was done with OpenMetaAnalyst software.[66] We also considered the correlation between sensitivity and specificity during bivariate modelling, the distance between each study and the HSROC curve, and the width of the prediction ellipse. Since some amount of heterogeneity is to be expected in studies on diagnostic test accuracy, we used random-effects modelling for all analyses.[45]

Additional analyses

We had prespecified subgroup analyses planned as part of our PROSPERO protocol, using the clinician's experience and the time since injury as covariates. However, due to the low number of studies meeting methodological criteria for meta-analysis, we deemed this inappropriate.

Counting inconclusive findings

According to Simel et al,[59] inconclusive findings can either be termed "uninterpretable", "intermediate", or "indeterminate". Uninterpretable results are when the patient, for whatever reason, cannot adequately undergo the intended test. Intermediate test results raise the disease's probability above what is deemed "healthy", but not enough to be considered "diseased". Indeterminate results add no additional value to the original probability of disease. It is often prudent to include inconclusive findings in the primary analysis to not risk overestimating the test's diagnostic accuracy.[58] For both the primary analysis and the meta-analysis, we grouped "uninterpretable" test results as injury positive, and "intermediate" test results as injury negative. The uninterpretable results were either due to excessive pain or swelling.[14,31,51,52] We believe that counting these patients as injury positive reflects what would have been done in the clinical setting; since clinicians would intuitively raise their suspicion of ligamentous damage if the patient presented with excessive levels of the aforementioned clinical signs. We grouped intermediate findings[31,51] (i.e tests were the clinician could not decide whether the patient had enough laxity to be determined injured vs uninjured) as disease negative; since the positivity criteria for stress testing is the definitive presence of increased joint laxity. We encountered no "indeterminate" tests results in the included studies. Supplementary 1 contains the inconclusive index test findings and the diagnostic yield as a percentage of manual stress tests used for diagnosis versus the number of patients intended to diagnose.

Patient and public involvement

Patients were not involved in the development of the research question or its outcome measures, the conduct of the research, or preparation of the manuscript. Dissemination of results to these groups is not applicable.

RESULTS

Study selection

Our search yielded 4786 records. After the initial title and abstract screening, we assessed 38 full-text articles for final eligibility. We excluded 24 articles due to: insufficient data[19,36,61] (n = 2); not a diagnostic test accuracy study[1,32,39,48] (n = 4); no clinical test[2,3,21,25,33–35,38,43,54] (n = 10); no or inaccurate reference test[15,29,44,49,53] (n = 5); case series[6,63] (n = 2); testing delivered under anaesthesia[69] (n = 1). We contacted three authors to help clarify details related to their data,[24,60,61] with none responding. In total, 14 articles met the inclusion criteria of our systematic review, with six of them contributing to meta-analysis. Figure 2 contains a flow chart of the study selection process.

[FIGURE 2]

Study characteristics and results

Supplementary 2 provides detailed information on study characteristics. Two studies were retrospective reviews,[8,27] the rest being diagnostic case-control,[24] clinical trials,[31] or prospective cohort studies (n = 10).[11,14,16,20,22,40,42,51,52,60] Studies included an aggregate of 2391 participants. The proportion of females within each study ranged from 23 to 51%. Seven studies were conducted in emergency departments[14,16,20,31,42,51,52] and seven in outpatient clinics.[8,11,22,24,27,40,60] Eleven out of 14 studies included sporting populations.[11,16,20,22,24,27,31,40,42,51,52] Only Gremeaux et al[27] and van der Ent[16] specified the level of play; the majority of which were recreational practitioners (85%) and amateur competitors (46%), respectively. Most studies included participants with recent (≤ 7 days) ankle injuries,[14,16,20,27,31,42,51,52,60] with the remainder enrolling participants with either chronic ankle instability,[8,24,40] or a mixture of both.[11] In addition to the binary classification of injury status, two out of the 14 studies also assessed the level of agreement for ordinal injury grading between index and reference testing.[8,22]

The reference standards used were: arthrography[14,16,20,31,51,52] (n = 6); arthroscopy or surgery[8,43] (n = 2); MRI[24,60] (n = 2); and ultrasound[11,22,27,40] (n = 4). Two out of six studies using arthrography as the reference standard did not aim to differentiate between the affected ligaments during reference testing, counting any ligament sprain as a positive finding.[14,20] One study[31] provided detailed information for arthrography criteria, but insufficient information in cross-reference to the index test results to differentiate between what ligaments were involved beyond the ATFL. Two out of four ultrasonographic studies defined a positive reference test as a partial to complete ATFL rupture.[11,40] Croy et al[11] was the only study that numerically quantified the degree of laxity during the ultrasound examination, and defined a positive finding as anterior talar displacement of ≥ 3.7mm, which constituted twice the standard deviation of the values from the healthy control group. George et al[22] and Gremeaux et al,[27] also using ultrasound as the reference standard, differentiated between ATFL and CFL tearing. De Simoni et al[60] also differentiated between injury of the two ligaments, but via MRI. Gomes et al[24] was the only study that did not disclose any details on what defined a positive finding during reference testing.

Five studies explicitly stated that they received financial aids through non-commercial research grants.[11,20,31,40,42] One study[24] noted that no grants whatsoever were received, and another two made clear that no commercial grants that would put the authors at a conflict of interest were received.[14,22] Six studies did not state any details on funding.[8,16,27,51,52,60]

Supplementary 3 has details of index test execution and positive test interpretation. The index test most commonly studied was the anterior drawer test[8,11,14,20,22,24,31,40,51,52] (n = 10) followed by palpation of the ATFL and the CFL (both n = 6).[14,16,20,27,42,60] Other stress tests used were the reverse anterior drawer[40,42] (n = 2), the anterolateral drawer[40] (n = 1), heel adduction[20] (n = 1), talar tilt[20,22,31,51] (n = 4), and supination test[20,42] (n = 2). The anterior drawer test was performed at varying degrees of plantar flexion, ranging from neutral[11,52] to 60°.[31,51] Most studies described a knee flexed test position,[8,11,14,20,22,24,40] either lying supine or seated. Positive test interpretation differed and was based on either increased laxity[8,11,20,22,24,31,40,51,52] or the presence of a dimple sign.[14] One author[42] stated that they had applied an anterior drawer test and a talar tilt test; however, the test description and images seem to align more with the reverse anterolateral drawer test[40] and the supination test.[20]

Details on test execution were scarce for studies examining palpation: most studies failed to report the exact point for palpation across the ligaments, and the amount of force applied. Only one study[16] stated that the entirety of the ligament was palpated for the pain punctum maximum; another[14] that the ATFL was palpated both by the tip of the fibula and over the talus.

Risk of bias within studies

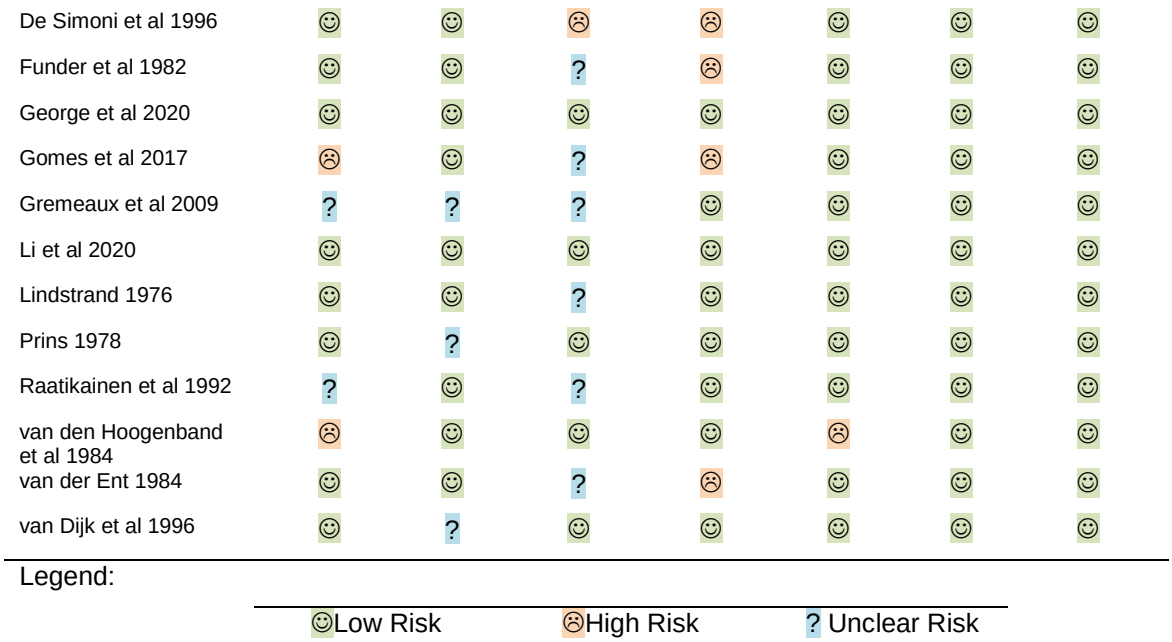
Table 2 summarises our QUADAS-2 assessment. Three studies; Croy et al,[11] George et al,[22] and Li et al,[40] completed all QUADAS-2 domains with a low risk of bias and with low concerns regarding applicability. Most studies had a low risk of bias regarding patient selection and index testing. Only Gomes et al,[24] using a case-control design, did not disclose patient enrollment and exclusion criteria.

There was an unclear risk of bias for test interpretation in nine of the included studies. Prins[51] performed reference testing before index testing, and Gremeaux et al[27] provided insufficient details to determine test order. Van Dijk et al[14] mentioned that a positive anterior drawer test was sometimes unwittingly interpreted based on pain response instead of increased laxity. Still, it is unclear how many patients were deemed injured based on the unintended pain criteria. In a further seven studies, it was unclear if the reference test was interpreted without knowledge of the results of the previous index tests.[8,16,20,24,27,42,52]

For study flow and timing, four studies carried a high risk of bias.[16,20,24,60] De Simoni et al[60] employed an inappropriate time interval between index testing and reference testing (mean delay 9.4 days). As the included patients were examined acutely (0–19 days following injury), each day of delay represents a relatively larger proportional discrepancy in study flow and timing, when compared to more prolonged periods of injury. Both Funder et al[20] and van der Ent[16] limited their reference standard examination to patients with high clinical suspicion and positive index tests, resulting in verification bias. Van der Ent's[16] cohort was further stratified based on the arthrographic findings for the subsequent treatment intervention. However, in the strata serving as the control group, insufficient information regarding the affected structures made it impossible to discern the diagnostic accuracy of the different palpation tests for this subset of patients. The control group in Gomes et al.[24] did not receive the reference standard, and it is unclear whether or not their data was used to calculate the sensitivity and specificity values of the studied clinical tests.

Table 2. QUADAS-2 Summary of Findings

Author [ref] and year	RISK OF BIAS				APPLICABILITY CONCERNS		
	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD	FLOW AND TIMING	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD
Cho et al 2016	?	☺	?	☺	☹	☺	☺
Croy et al 2013	☺	☺	☺	☺	☺	☺	☺



[FIGURE 3]

Results of individual studies

Figure 3 presents the diagnostic accuracy of each test from the individual studies. In total, 6302 observations from 14 studies spread over nine clinical tests contributed to the narrative synthesis.

Manual stress tests

The drawer test has higher specificity than sensitivity for diagnosing injury to the ATFL[8,22,24,31,40,42,51,52], any lateral ligamentous injury,[14,20] or excessive joint instability.[11] This was typically observed, regardless of the technique employed: anterior drawer test[8,11,14,20,22,24,31,40,51,52] (sensitivity range 12–80%, specificity range 67–100%); anterolateral drawer test[40] (47% sensitivity and 99% specificity); reverse anterolateral drawer test[40,42] (sensitivity range 83–89%, specificity range 70–90%). The talar tilt test[20,22,51] and the heel adduction test[20] were also more specific than sensitive for diagnosing any lateral ligamentous injury[20,31] or injury to the CFL[22,51] displaying 17–66% sensitivity with 82–100% specificity, and 35% sensitivity with 77% specificity, respectively. Conversely, the supination test[20,42] proved more sensitive (73–98%) than specific (4–23%) for diagnosing ATFL injury[42] or any lateral ligamentous injury.[20]

Palpation

Palpation is more sensitive than specific. Anterolateral talar palpation[24] displayed a perfect sensitivity (100%) and 80% specificity for diagnosing injury to the ATFL. Direct palpation of the ATFL[14,20,27,42,60] consistently showed high sensitivity (95–100%) across six studies but low (0–32%) specificity when diagnosing ATFL rupture[16,27,42,60] or any affected lateral collateral ligament.[14,20] Palpation of the CFL[14,16,20,27,60] had worse sensitivity, ranging between 49–100%, whilst specificity ranged between 26–79% for diagnosing partial to total tearing of the CFL[16,27,42,60] or any lateral ligamentous tear.[14,20]

No diagnostic test accuracy study examining clinical tests for the subtalar joint met our inclusion criteria.

Meta-analysis

Six studies (885 observations) examining the anterior drawer test were included in our meta-analysis.[11,14,22,40,51,52] Using a bivariate model, the pooled metrics for the anterior drawer test were: sensitivity 54% (95% CI 35 to 71%); specificity 87% (95% CI 63 to 96%); LR+ 3.97 (95% CI 1.50 to 10.47); and LR- 0.54 (95% CI 0.39 to 0.75) (n = 6). Sensitivity and specificity were negatively correlated (-0.73). When modelled independently, sensitivity displayed significant heterogeneity (I-square = 94.17%, Cochran's Q p-value < 0.001) and specificity displayed substantial heterogeneity (I-square = 62.083%, Cochran's Q p-value = 0.022). It is plausible that a threshold effect in test interpretation (i.e. the amount of laxity required during translation for the clinician to say that the patient is injured) explains some of the between-study variations in sensitivity and specificity.[64] A threshold effect is further supported by the distance of the studies from the summary curve and the prediction ellipse (Figure 4).[45]

[FIGURE 4]

The median prevalence for any lateral ankle ligament injury was 65% (36–76% min-max) in the studies underdoing meta-analysis. Using this percentage as the pretest probability of injury for Fagan's nomogram, a positive anterior drawer test (LR+ 3.97) increases the clinical likelihood of lateral ligamentous injury to 88%. A negative test result (LR- 0.54) is associated with a smaller drop in probability to 50% (Figure 5).

[FIGURE 5]

Assessing the degree of ligamentous injury

Cho et al[8] investigated the discriminatory capabilities of the anterior drawer test in comparison to arthroscopic grading of perceived joint laxity on a three-point ordinal scale (subtle/moderate/severe laxity; grade I/II/III). Although 77% agreement was observed between the clinical grading and arthroscopic grading, this was no greater than chance agreement [(Index test: 0, 6, 20) (Reference test: 0, 0, 26) (κ = 0, weighted Cohen's kappa)], implicating limited use of the clinical test in differentiating between moderate and severe cases of joint laxity.

George et al[22] used a similar clinical grading scale (no/some/gross laxity; grade I/II/III) and cross-referenced the findings with stress ultrasound examination (intact/partially torn/completely torn ATFL ligament; grade I/II/III). However, George et al[22] included a larger sample and patients of varying injury severity. In this study, the grading of perceived laxity during anterior drawer testing and the amount of ATFL tearing found during stress ultrasound examination reached moderate agreement [(Index test: 10, 12, 13) (Reference test: 8, 5, 22) (κ = 0.53, weighted Cohen's kappa)].

George et al[22] also examined the agreement between clinical grading during the talar tilt test and the degree of CFL rupture during dynamic ultrasonography. The proportion of unaffected ankles were greater (15 versus 8) for the CFL in comparison to the ATFL, and tears were evenly distributed between partial (n = 5), and total (n = 5) ruptures. Still, the inter-rater agreement between clinical and ultrasound grading of CFL status was almost identical to that of the anterior drawer test and ultrasound ATFL grading, displaying moderate agreement [(Index test: 16, 14, 5) (Reference test: 15, 10, 10) (κ = 0.52, weighted Cohen's kappa)].

DISCUSSION

Principal findings

Lateral ankle sprains are the most common acute musculoskeletal injury. They can result in damage to any of the primary lateral ligaments spanning the talocrural (ATFL, CFL, PTFL) and subtalar joints (ITCL, CL, ACL). Diagnosis and prognosis post-sprain should be informed by the number of ligaments damaged and the severity of the tear. This review suggests accurate clinical diagnosis is limited to one ligament in the ankle complex; the ATFL. Diagnosis of injury to the ATFL achieves maximum accuracy through clustering of ligament palpation (highly sensitive) and anterior drawer testing (highly specific). The talar tilt test can help rule in injury to the CFL, but sensitive tests aimed at the ligament is lacking. There is limited and conflicting evidence that clinical tests can provide an accurate assessment of injury severity. Studies examining the diagnostic accuracy of clinical tests aimed at the subtalar ligaments are lacking.

Explanations and implications for clinicians

Ligamentous injury to the ankle typically follows a hierarchical pattern. The ATFL is the weakest lateral ligament and is involved in ~80% of ankle sprains.[42] The evidence suggests that clinical assessment of the ATFL necessitates a combination of palpation and anterior drawer testing to differentiate between injured and uninjured patients accurately. Although palpation techniques were poorly described, we would suggest that the entire ligament is examined, with tenderness at any point indicating a positive finding. The accuracy of the anterior drawer test may be moderated by the test setup, the positivity threshold, and the timing of the test. Traditionally, this test involves moving the heel anteriorly on the tibia. High accuracy was also achieved using a reverse drawer technique,[40,42] whereby the tibia was pushed posteriorly on a fixed heel. A common feature of both methods was that patients were positioned in knee flexion and plantarflexion. Biomechanical studies corroborate these joint positions, ensuring minimal tension at the triceps surae and maximal recruitment of the ATFL.[35,37]

The positive predictive value of the anterior drawer test may be enhanced further by adopting a high threshold for positivity. This includes interpreting subtle laxities[11,22] and intermediate results[31,51] as negative. Three studies[14,16,51] validate the notion that the accuracy of clinical examination is maximised when undertaken in a delayed (2–7 days) versus acute (< 48h) setting. The CFL is the only ligament in the lateral collateral complex that crosses both the talocrural and subtalar joints,[23] and therefore plays an essential role in the lateral stability of the ankle.[67] Given that peroneal tendons and sheaths cover the majority of the CFL,[23] it is unsurprising that palpating the ligament provides limited diagnostic value. Although we found consistent evidence that the talar tilt test has excellent specificity, and is useful for ruling in injury to the CFL,[20,22,51] caution is required when interpreting a negative test. This finding supports the hypothesis that some instabilities of the lateral ligament complex are occult to clinical examination, which may mediate the risk of inadequate management and development of chronic ankle instability.[4] A related limitation is that we cannot present any clinical tests that are suitable for diagnosing injury to the subtalar ligaments (ITCL, CL, ACL). This is a critical gap in the current evidence base, as differentiating between an isolated vs combined injury of the talocrural and subtalar joints are fundamental for accurate prognostication and clinical management decisions.

Strength and limitations

Our study is the first meta-analysis examining the accuracy of clinical testing commonly used for diagnosing ankle sprains. Others have reviewed the evidence in this field,[56,57] but trial numbers were limited (n = 5), with the majority limited to radiographic reference standards. The current review includes data from 6302 observations across 14 trials, including higher quality, contemporary reference standards (ultrasound, magnetic

resonance imaging, and arthroscopy). Although only two studies incorporated the current gold standard reference (arthroscopy or surgery), a previous meta-analysis show that high diagnostic accuracy is possible using MRI, ultrasound or stress radiography (81–99% sensitivity and 79–91% specificity).[7] Still, as these reference standards are not perfect (and showcase variability), the diagnostic accuracy of the clinical tests of many of our included studies should be interpreted accordingly. Only three out of the 14 studies that we included had a low risk of bias across all QUADAS-2 domains. Verification bias was the most frequent, either due to improper time frames between the index and reference test or selective criteria. The generalisation of our findings is also affected by poor reporting of test interpretation: being commonly ambiguous and presenting with an unclear risk of bias. Only one study made direct comparisons between modified techniques for routine stress tests,[40] and just two studies incorporated an ordinal scale to grade injury severity.[8,22] As their results were contradictory, it is unclear if clinical tests of the talocrural joint can grade ligament damage beyond the binary. This review focuses on lateral ligament injuries, but we acknowledge that ankle sprains can also involve the ankle syndesmosis. Injuries to the syndesmosis will often have a different injuring mechanism,[41] and are assessed through alternative clinical tests featured in previous diagnostic reviews.[62] Although our meta-analysis excluded studies at a high risk of bias, the generalisability of our reported pooled diagnostic estimates to any specific setting might still be limited by reported differences in test technique, time since injury, reference standard used, and potential differences in remittance time. Lastly, our proposed diagnostic algorithm of performing palpation and anterior drawer testing of the ATFL for accurate diagnosis has not yet been validated with patient paired data.

What is already known

- Lateral ankle sprains are the most common musculoskeletal injury and can incur damage to some or all the six major ligaments spanning the ankle and subtalar joints
- Diagnosis should aim to differentiate and grade isolated vs combined injuries of the talocrural and subtalar joints, in order to determine prognosis and management choice (surgical vs conservative)
- Evidence syntheses of diagnostic clinical tests including contemporary reference standards is currently lacking

What are the new findings

- There are risk of bias concerns in most diagnostic research of clinical examination for lateral ankle sprains
- Generalisation of results is primarily affected by insufficient information regarding test interpretation and verification bias
- Clinical examination can accurately assess one major ligament spanning the ankle joint (anterior talofibular ligament), based on a cluster of palpation and anterior drawer testing
- We found limited and contradicting evidence for clinical injury grading beyond the binary for the ankle joint, and evidence for stress tests of the subtalar ligaments is lacking

FUNDING

No funding was received for this systematic review with meta-analysis.

DATA AVAILABILITY STATEMENT

Data are available in a public, open access repository upon publication, including our RevMan file and the CSV-file used for meta-analysis. Please cite this article if our data synthesis, data analysis, or data interpretation is used as part of your publication.

REFERENCES

1. Avci S, Şayli U. Comparison of the results of short-term rigid and semi-rigid cast immobilization for the treatment of grade 3 inversion injuries of the ankle. *Injury*. 1998;29(8):581-584.
2. Baltes TPA, Arnáiz J, Geertsema L, et al. Diagnostic value of ultrasonography in acute lateral and syndesmotic ligamentous ankle injuries. *Eur Radiol*. Published online October 7, 2020.
3. Becker HP, Komischke A, Danz B, Bensel R, Claes L. Stress diagnostics of the sprained ankle: Evaluation of the anterior drawer test with and without anesthesia. *Foot Ankle*. 1993;14(8):459-464. PMID:8253439
4. Bleakley C, Wagemans J, Netterström-Wedin F. Understanding chronic ankle instability: model rich, data poor. *Br J Sports Med*. Published online 2020:bjsports-2020-103311.
5. Bridgman SA, Clement D, Downing A, Walley G, Phair I, Maffulli N. Population based epidemiology of ankle sprains attending accident and emergency units in the West Midlands of England, and a survey of UK practice for severe ankle sprains. *Emerg Med J*. 2003;20(6):508-510. PMID:14623833
6. Broström L. Sprained ankles. VI. Surgical treatment of "chronic" ligament ruptures. *Acta Chir Scand*. 1966;132(5):551-565. PMID:5339635
7. Cao S, Wang C, Ma X, Wang X, Huang J, Zhang C. Imaging diagnosis for chronic lateral ankle ligament injury: A systemic review with meta-analysis. *J Orthop Surg Res*. 2018;13(1).
8. Cho JH, Lee DH, Song HK, Bang JY, Lee KT, Park YU. Value of stress ultrasound for the diagnosis of chronic ankle instability compared to manual anterior drawer test, stress radiography, magnetic resonance imaging, and arthroscopy. *Knee Surgery, Sport Traumatol Arthrosc*. 2016;24(4):1022-1028.
9. Cochrane. Review Manager (RevMan) [Computer program]. Version 5.4. Copenhagen: The Nordic Cochrane Centre. *Cochrane Collab*. Published online 2020.
10. Cooke MW, Lamb SE, Marsh J, Dale J. A survey of current consultant practice of treatment of severe ankle sprains in emergency departments in the United Kingdom. *Emerg Med J*. 2003;20(6):505-507. PMID:14623832
11. Croy T, Koppenhaver S, Saliba S, Hertel J. Anterior talocrural joint laxity: Diagnostic accuracy of the anterior drawer test of the ankle. *J Orthop Sports Phys Ther*. 2013;43(12):911-919. PMID:24175608
12. Davidson M. The interpretation of diagnostic tests: A primer for physiotherapists. *Aust J Physiother*. 2002;48(3):227-232. PMID:12217073
13. Delahunt E, Bleakley CM, Bossard DS, et al. Clinical assessment of acute lateral ankle sprain injuries (ROAST): 2019 consensus statement and recommendations of the International Ankle Consortium. *Br J Sports Med*. 2018;52(20):1304-1310. PMID:29886432
14. Van Dijk CN, Lim LSL, Bossuyt PMM, Marti RK. Physical examination is sufficient for the diagnosis of sprained ankles. *J Bone Jt Surg - Ser B*. 1996;78(6):958-962. PMID:8951015
15. Doherty C, Bleakley C, Hertel J, Caulfield B, Ryan J, Delahunt E. Clinical Tests Have Limited Predictive Value for Chronic Ankle Instability When Conducted in the Acute Phase of a First-Time Lateral Ankle Sprain Injury. *Arch Phys Med Rehabil*. 2018;99(4):720-725.e1.
16. van der Ent FW. Lateral Ankle Ligament Injury: An Experimental and Clinical Study.

Johnsen & Johnsen. Published online 1984.
17. Fagan TJ. Nomogram for Bayes's Theorem. *N Engl J Med*. 1975;293(5):257.
18. Freeman SC, Kerby CR, Patel A, Cooper NJ, Quinn T, Sutton AJ. Development of an interactive web-based tool to conduct and interrogate meta-analysis of diagnostic test accuracy studies: MetaDTA. *BMC Med Res Methodol*. 2019;19(1):81. PMID:30999861
19. Frey C, Bell J, Teresi L, Kerr R, Feder K. A comparison of MRI and clinical examination of acute lateral ankle sprains. *Foot Ankle Int*. 1996;17(9):533-537.
20. Funder V, Jørgensen JP, Andersen A, et al. Ruptures of the lateral ligaments of the ankle: Clinical diagnosis. *Acta Orthop*. 1982;53(6):997-1000. PMID:7180412
21. Gaebler C, Kukla C, Breitenseher MJ, et al. Diagnosis of lateral ankle ligament injuries. Comparison between talar tilt, MRI and operative findings in 112 athletes. *Acta Orthop Scand*. 1997;68(3):286-290.
22. George J, Jaafar Z, Hairi IR, Hussein KH. The correlation between clinical and ultrasound evaluation of anterior talofibular ligament and calcaneofibular ligament tears in athletes. *J Sports Med Phys Fitness*. 2020;60(5):47-76.
23. Golanó P, Vega J, de Leeuw PAJ, et al. Anatomy of the ankle ligaments: a pictorial essay. *Knee Surgery, Sport Traumatol Arthrosc*. 2016;24(4):944-956. PMID:27052302
24. Gomes JLE, Soares AF, Bastiani CE, de Castro JV. Anterolateral talar palpation: A complementary test for ankle instability. *Foot Ankle Surg*. 2018;24(6):486-489.
25. Gondim Teixeira PA, Formery AS, Balazuc G, et al. Comparison between subtalar joint quantitative kinematic 4-D CT parameters in healthy volunteers and patients with joint stiffness or chronic ankle instability: A preliminary study. *Eur J Radiol*. 2019;114:76-84. PMID:31005181
26. Graphpad Software. GraphPad QuickCalcs.
27. Gremeaux V, Coudreuse JM, Collado H, et al. Comparative study of clinical and ultrasonographic evaluation of lateral collateral ligament sprains of the ankle. *J Sports Med Phys Fitness*. 2009;49(3):285-291. PMID:19861934
28. Gribble PA, Bleakley CM, Caulfield BM, et al. Evidence review for the 2016 International Ankle Consortium consensus statement on the prevalence, impact and long-term consequences of lateral ankle sprains. *Br J Sports Med*. 2016;50(24):1496-1505. PMID:27259753
29. Hertel J, Denegar CR, Monroe MM, Stokes WL. Talocrural and subtalar joint instability after lateral ankle sprain. *Med Sci Sports Exerc*. 1999;31(11):1501-1508.
30. Higgins J, Green S. Cochrane handbook for systematic reviews of interventions. Version 5.1.0. Oxford: The Cochrane Collaboration. https://handbook-5-1.cochrane.org/chapter_9/9_5_2_identifying_and_measuring_heterogeneity.htm
31. van den Hoogenband CR, van Moppes FI, Stapert JWJL, Greep JM. Clinical diagnosis, arthrography, stress examination and surgical findings after inversion trauma of the ankle. *Arch Orthop Trauma Surg*. 1984;103(2). PMID:6477070
32. Hubbard TJ, Cordova M. Mechanical Instability After an Acute Lateral Ankle Sprain. *Arch Phys Med Rehabil*. 2009;90(7):1142-1146.
33. Johannsen A. Radiological diagnosis of lateral ligament lesion of the ankle: A comparison between talar tilt and anterior drawer sign. *Acta Orthop*. 1978;49(3):295-301. PMID:685675
34. Kato T. The diagnosis and treatment of instability of the subtalar joint. *J Bone Jt Surg - Ser B*. 1995;77(3):400-406. PMID:7744924
35. Kikumoto T, Akatsuka K, Nakamura E, Ito W, Hirabayashi R, Edama M. Quantitative

evaluation method for clarifying ankle plantar flexion angles using anterior drawer and inversion stress tests: A cross-sectional study. *J Foot Ankle Res*. 2019;12(1). PMID:31073333

36. Korkala O, Lauttamus L, Tanskanen P. Lateral ligament injuries of the ankle. Results of primary surgical treatment. *Ann Chir Gynaecol*. 1982;71(3):161-163. PMID:7114782

37. Kovaleski JE, Norrell PM, Heitman RJ, Hollis JM, Pearsall IV AW. Knee and ankle position, anterior drawer laxity, and stiffness of the ankle complex. *J Athl Train*. 2008;43(3):242-248.

38. Lahde S, Putkonen M, Puranen J, Raatikainen T. Examination of the sprained ankle: Anterior drawer test or arthrography? *Eur J Radiol*. 1988;8(4):255-257. PMID:3148464

39. Lee KT, Park YU, Jegal H, Park JW, Choi JP, Kim JS. New method of diagnosis for chronic ankle instability: Comparison of manual anterior drawer test, stress radiography and stress ultrasound. *Knee Surgery, Sport Traumatol Arthrosc*. 2014;22(7):1701-1707.

40. Li Q, Tu Y, Chen J, et al. Reverse anterolateral drawer test is more sensitive and accurate for diagnosing chronic anterior talofibular ligament injury. *Knee Surgery, Sport Traumatol Arthrosc*. 2020;28(1):55-62. PMID:31559464

41. Lin CF, Gross MT, Weinhold P. Ankle syndesmosis injuries: Anatomy, biomechanics, mechanism of injury, and clinical guidelines for diagnosis and intervention. *J Orthop Sports Phys Ther*. 2006;36(6):372-384. PMID:16776487

42. Lindstrand A. Lateral lesions in sprained ankles. A clinical and roentgenological study with special reference to anterior instability of the talus. Published online 1976.

43. Lindstrand A, Mortensson W. Anterior instability in the ankle joint following acute lateral sprain. *Acta radiol*. 1977;18(5):529-539.

44. Lohrer H, Nauck T, Gehring D, Wissler S, Braag B, Gollhofer A. Differences between mechanically stable and unstable chronic ankle instability subgroups when examined by arthrometer and FAAM-G. *J Orthop Surg Res*. 2015;10(1). PMID:25890204

45. Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. Chapter 10: Analysing and Presenting Results. In: *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0*. ; 2010:1-61. http://methods.cochrane.org/sdt/handbook-dta-reviews

46. McHugh ML. Interrater reliability: The kappa statistic. *Biochem Medica*. 2012;22(3):276-282. PMID:23092060

47. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*. 2009;6(7):e1000097. PMID:19621072

48. Nyska M, Amir H, Porath A, Dekel S. Radiological assessment of a modified anterior drawer test of the ankle. *Foot Ankle*. 1992;13(7):400-403. PMID:1427531

49. Pasapula C, Memarzadeh A, Devany A, et al. The Modified Anterior Drawer Test (MADT): A New Clinical Test for Assessing Subtalar Instability a Cadaveric and Clinical Assessment. *Clin Res Foot Ankle*. 2018;06(02).

50. Patel A, Cooper N, Freeman S, Sutton A. Graphical enhancements to summary receiver operating characteristic plots to facilitate the analysis and reporting of meta-analysis of diagnostic test accuracy data. *Res Synth Methods*. Published online August 12, 2020:jrsm.1439.

51. Prins JG. Diagnosis and treatment of injury to the lateral ligament of the ankle. A comparative clinical study. *Acta Chir Scand Suppl*. 1978;486:3-149. PMID:282741
52. Raatikainen T, Putkonen M, Puranen J. Arthrography, clinical examination, and stress radiograph in the diagnosis of acute injury to the lateral ligaments of the ankle. *Am J Sports Med*. 1992;20(1):2-6.
53. Rosen AB, Ko J, Brown CN. Diagnostic accuracy of instrumented and manual talar tilt tests in chronic ankle instability populations. *Scand J Med Sci Sport*. 2015;25(2):e214-e221.
54. Ross SE, Guskiewicz KM, Gross MT, Yu B. Assessment tools for identifying functional limitations associated with functional ankle instability. *J Athl Train*. 2008;43(1):44-50. PMID:18335012
55. Schmidt RL, Factor RE. Understanding Sources of Bias in Diagnostic Accuracy Studies. *Arch Pathol Lab Med*. 2013;137(4):558-565.
56. Schneiders A, Karas S. The accuracy of clinical tests in diagnosing ankle ligament injury. *Eur J Physiother*. 2016;18(4):245-253.
57. Schwieterman B, Haas D, Columber K, Knupp D, Cook C. Diagnostic accuracy of physical examination tests of the ankle/foot complex: a systematic review. *Int J Sports Phys Ther*. 2013;8(4):416-426. PMID:24175128
58. Shinkins B, Thompson M, Mallett S, Perera R. Diagnostic accuracy studies: how to report and analyse inconclusive test results. *BMJ*. 2013;346. PMID:23682043
59. Simel DL, Feussner JR, Delong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. *Med Decis Mak*. 1987;7(2):107-114. PMID:3574020
60. De Simoni C, Wetz HH, Zanetti M, Hodler J, Jacob H, Zollinger H. Clinical examination and magnetic resonance imaging in the assessment of ankle sprains treated with an orthosis. *Foot Ankle Int*. 1996;17(3):177-182.
61. Singh K, Lal Thukral C, Gupta K, Singh A. Comparison of high resolution ultrasonography with clinical findings in patients with ankle pain. *J Ultrason*. 2018;18(75):316-324.
62. Sman AD, Hiller CE, Refshauge KM. Diagnostic accuracy of clinical tests for diagnosis of ankle syndesmosis injury: A systematic review. *Br J Sports Med*. 2013;47(10):620-628.
63. Staples OS. Result study of ruptures of lateral ligaments of the ankle. *Clin Orthop Relat Res*. 1972;85:50-58. PMID:4624809
64. Trikalinos TA, Balion CM, Coleman CI, et al. Chapter 8: Meta-analysis of test performance when there is a “gold standard.” *J Gen Intern Med*. 2012;27(SUPPL.1). PMID:22648676
65. University of Bristol. QUADAS-2 : Background Document. School of Social and Community Medicine. Published 2014. http://www.bristol.ac.uk/media-library/sites/quadas/migrated/documents/background-doc.pdf
66. Wallace BC, Dahabreh IJ, Trikalinos TA, Lau J, Trow P, Schmid CH. Closing the gap between methodologists and end-users: R as a computational back-end. *J Stat Softw*. 2012;49:5.
67. Weindel S, Schmidt R, Rammelt S, Claes L, Campe A V., Rein S. Subtalar instability: A biomechanical cadaver study. *Arch Orthop Trauma Surg*. 2010;130(3):313-319.
68. Whiting PF, Rutjes AWS, Westwood ME, et al. Quadas-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-536. PMID:22007046
69. Wiebking U, Pacha TO, Jagodzinski M. An accuracy evaluation of clinical,

arthrometric, and stress-sonographic acute ankle instability examinations. *Foot Ankle Surg*. 2015;21(1):42-48.

FIGURE LEGENDS

Figure 1

(number of hits)

Figure 2

Both authors independently examined each record for study inclusion eligibility and suitability for the subsequent meta-analysis.

Figure 3

TP = True Positive
FP = False Positive
TN = True Negative
FN = False Negative
*77 patients examined by two different examiners

Figure 4

The distance between the study points and the summary curve, as well as the width of the prediction ellipse, hints towards differences in positivity threshold (i.e the amount of laxity necessary for the clinician to classify the patient as injured) for the included studies.

Figure 5

The median disease prevalence of studies undergoing meta-analysis was used as the pretest probability of injury (any lateral ligamentous injury). A positive anterior drawer test is associated with a much greater shift in post-test probability of ligamentous damage in comparison to a negative test result.