



## Diagnostic Accuracy of Clinical Tests Assessing Ligamentous Injury of the Talocrural and Subtalar Joints: A Systematic Review With Meta-Analysis

Netterström-Wedin F, F., Matthews, M., & Bleakley, C. M. (2022). Diagnostic Accuracy of Clinical Tests Assessing Ligamentous Injury of the Talocrural and Subtalar Joints: A Systematic Review With Meta-Analysis. *Sports Health*, 14(3), 336-347. <https://doi.org/10.1177/19417381211029953>

[Link to publication record in Ulster University Research Portal](#)

**Published in:**  
Sports Health

**Publication Status:**  
Published (in print/issue): 01/05/2022

**DOI:**  
[10.1177/19417381211029953](https://doi.org/10.1177/19417381211029953)

**Document Version**  
Peer reviewed version

### General rights

Copyright for the publications made accessible via Ulster University's Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

### Take down policy

The Research Portal is Ulster University's institutional repository that provides access to Ulster's research outputs. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact [pure-support@ulster.ac.uk](mailto:pure-support@ulster.ac.uk).

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

4  
5 **Word count: 4265 words**

6

## 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

62 **KEYWORDS**

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

## INTRODUCTION

64

65 Each year, over 300,000 people present to UK Emergency Departments with ankle sprain  
66 (~800 per day).<sup>5</sup> Many occur during sporting or recreational activity, due to excessive  
67 inversion and internal rotation of the ankle at high velocity.<sup>28</sup> Ankle sprains are often  
68 regarded as innocuous injuries, but up to 70% of patients develop chronic ankle  
69 instability; characterised by mechanical laxity, subjective feelings of giving way, persistent  
70 pain and reinjury.<sup>28</sup> In the UK, the total average cost associated with a lateral ankle sprain  
71 is estimated at £940.<sup>10</sup> The high incidence of chronic symptoms, risk of recurrence, and  
72 long term risk of developing post-traumatic osteoarthritis, further contribute to the  
73 significant socioeconomic burden of lateral ankle sprains.<sup>28</sup>

74

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

87

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

97

99 **Protocol and registration**

100 We used the Preferred Reporting Items for Systematic Reviews and Meta-Analysis of  
 101 Diagnostic Test Accuracy Studies (PRISMA-DTA)<sup>47</sup> for our review.

102

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

105

106 **Eligibility criteria**

107 We assessed original research for eligibility using the criteria presented in Table 1, with  
 108 no restrictions on the language of the article nor the publication year. Most criteria were  
 109 decided on a priori, as part of the PROSPERO protocol. However, arthroscopy as an  
 110 inclusion criterion was extended to include other surgical techniques as well, and avulsion  
 111 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	

112 **Search**

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

119

120

[FIGURE 1]

121 **Study selection**

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

132

### 133 Risk of bias in individual studies

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

154

### 155 Data items

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

164

### 165 Synthesis of results

166

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

173

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

177

178 If ordinal level data were reported, these were extracted and analysed to see if clinical  
179 tests can accurately grade the degree of injury. We calculated the inter-rater agreement  
180 between index test and reference test with weighted Cohen's kappa (linear weighting),  
181 using an online calculator.<sup>26</sup> According to McHugh,<sup>46</sup> kappa values for agreement are to  
182 be interpreted as: 0–20 = none; 21–39 = minimal; 40–59 = weak; 60–79 = moderate;  
183 80–90 = strong; > 90 = almost perfect

184

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

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

### 191 Meta-analysis

192 We performed HSROC and bivariate meta-analyses with MetaDTA 2.0 software.<sup>18,50</sup> We  
193 calculated pooled summary estimates of test sensitivity, specificity, and positive and  
194 negative likelihood ratios (LR), each with 95% CI. Likelihood ratios are considered a  
195 useful diagnostic metric and represent the prevalence of positive tests in those with injury  
196 versus those without (LR+) and the prevalence of negative tests in those that are healthy  
197 versus those that are not (LR-).<sup>12</sup> We plotted the pooled LRs in Fagan's nomogram,<sup>17</sup> to  
198 examine the change in pre to post-test probability after positive and negative tests. We  
199 estimated the pretest probability through the median disease prevalence of studies  
200 eligible for meta-analysis. To determine heterogeneity, we used the Cochran Q test ( $p <$   
201  $0.05$  indicating presence of heterogeneity) and the *I*-squared statistic. *I*-squared values of  
202 0-40%, 30-60%, 50-90%, and 75-100% were considered non-important, moderate,  
203 substantial, and significant levels of heterogeneity, respectively.<sup>30</sup> This univariate analysis  
204 of heterogeneity was done with OpenMetaAnalyst software.<sup>66</sup> We also considered the  
205 correlation between sensitivity and specificity during bivariate modelling, the distance  
206 between each study and the HSROC curve, and the width of the prediction ellipse. Since  
207 some amount of heterogeneity is to be expected in studies on diagnostic test accuracy, we  
208 used random-effects modelling for all analyses.<sup>45</sup>

209

### 210 Additional analyses

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

215

### 216 Counting inconclusive findings

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

237

### 238 Patient and public involvement

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



242

243

## RESULTS

### 244 Study selection

245 Our search yielded 4786 records. After the initial title and abstract screening, we assessed  
246 38 full-text articles for final eligibility. We excluded 24 articles due to: insufficient  
247 data<sup>19,36,61</sup> ( $n = 2$ ); not a diagnostic test accuracy study<sup>1,32,39,48</sup> ( $n = 4$ ); no clinical  
248 test<sup>2,3,21,25,33-35,38,43,54</sup> ( $n = 10$ ); no or inaccurate reference test<sup>15,29,44,49,53</sup> ( $n = 5$ ); case  
249 series<sup>6,63</sup> ( $n = 2$ ); testing delivered under anaesthesia<sup>69</sup> ( $n = 1$ ). We contacted three  
250 authors to help clarify details related to their data,<sup>24,60,61</sup> with none responding. In total,  
251 14 articles met the inclusion criteria of our systematic review, with six of them  
252 contributing to meta-analysis. Figure 2 contains a flow chart of the study selection  
253 process.

254

255

[FIGURE 2]

256

### 257 Study characteristics and results

258 Supplementary 2 provides detailed information on study characteristics. Two studies  
259 were retrospective reviews,<sup>8,27</sup> the rest being diagnostic case-control,<sup>24</sup> clinical trials,<sup>31</sup> or  
260 prospective cohort studies ( $n = 10$ ).<sup>11,14,16,20,22,40,42,51,52,60</sup> Studies included an aggregate of  
261 2391 participants. The proportion of females within each study ranged from 23 to 51%.  
262 Seven studies were conducted in emergency departments<sup>14,16,20,31,42,51,52</sup> and seven in  
263 outpatient clinics.<sup>8,11,22,24,27,40,60</sup> Eleven out of 14 studies included sporting  
264 populations.<sup>11,16,20,22,24,27,31,40,42,51,52</sup> Only Gremeaux et al<sup>27</sup> and van der Ent<sup>16</sup> specified the  
265 level of play; the majority of which were recreational practitioners (85%) and amateur  
266 competitors (46%), respectively. Most studies included participants with recent ( $\leq 7$  days)  
267 ankle injuries,<sup>14,16,20,27,31,42,51,52,60</sup> with the remainder enrolling participants with either  
268 chronic ankle instability,<sup>8,24,40</sup> or a mixture of both.<sup>11</sup> In addition to the binary  
269 classification of injury status, two out of the 14 studies also assessed the level of  
270 agreement for ordinal injury grading between index and reference testing.<sup>8,22</sup>

271

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

288

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

293

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

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

### 312 Risk of bias within studies

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

318  
 319 There was an unclear risk of bias for test interpretation in nine of the included studies.  
 320 Prins<sup>51</sup> performed reference testing before index testing, and Gremeaux et al<sup>27</sup> provided  
 321 insufficient details to determine test order. Van Dijk et al<sup>14</sup> mentioned that a positive  
 322 anterior drawer test was sometimes unwittingly interpreted based on pain response  
 323 instead of increased laxity. Still, it is unclear how many patients were deemed injured  
 324 based on the unintended pain criteria. In a further seven studies, it was unclear if the  
 325 reference test was interpreted without knowledge of the results of the previous index  
 326 tests.<sup>8,16,20,24,27,42,52</sup>

327  
 328 For study flow and timing, four studies carried a high risk of bias.<sup>16,20,24,60</sup> De Simoni et  
 329 al<sup>60</sup> employed an inappropriate time interval between index testing and reference testing  
 330 (mean delay 9.4 days). As the included patients were examined acutely (0–19 days  
 331 following injury), each day of delay represents a relatively larger proportional discrepancy  
 332 in study flow and timing, when compared to more prolonged periods of injury. Both  
 333 Funder et al<sup>20</sup> and van der Ent<sup>16</sup> limited their reference standard examination to patients  
 334 with high clinical suspicion and positive index tests, resulting in verification bias. Van der  
 335 Ent's<sup>16</sup> cohort was further stratified based on the arthrographic findings for the  
 336 subsequent treatment intervention. However, in the strata serving as the control group,  
 337 insufficient information regarding the affected structures made it impossible to discern  
 338 the diagnostic accuracy of the different palpation tests for this subset of patients. The  
 339 control group in Gomes et al.<sup>24</sup> did not receive the reference standard, and it is unclear  
 340 whether or not their data was used to calculate the sensitivity and specificity values of the  
 341 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	😊	😊	😊	😊	😊	😊	😊

De Simoni et al 1996	😊	😊	😞	😞	😊	😊	😊
Funder et al 1982	😊	😊	?	😞	😊	😊	😊
George et al 2020	😊	😊	😊	😊	😊	😊	😊
Gomes et al 2017	😞	😊	?	😞	😊	😊	😊
Gremeaux et al 2009	?	?	?	😊	😊	😊	😊
Li et al 2020	😊	😊	😊	😊	😊	😊	😊
Lindstrand 1976	😊	😊	?	😊	😊	😊	😊
Prins 1978	😊	?	😊	😊	😊	😊	😊
Raatikainen et al 1992	?	😊	?	😊	😊	😊	😊
van den Hoogenband et al 1984	😞	😊	😊	😊	😞	😊	😊
van der Ent 1984	😊	😊	?	😞	😊	😊	😊
van Dijk et al 1996	😊	?	😊	😊	😊	😊	😊

Legend:

😊 Low Risk      😞 High Risk      ? Unclear Risk

343

344

[FIGURE 3]

345

### 346 Results of individual studies

347 Figure 3 presents the diagnostic accuracy of each test from the individual studies. In total,  
348 6302 observations from 14 studies spread over nine clinical tests contributed to the  
349 narrative synthesis.

### 350 Manual stress tests

351 The drawer test has higher specificity than sensitivity for diagnosing injury to the  
352 ATFL<sup>8,22,24,31,40,42,51,52</sup>, any lateral ligamentous injury,<sup>14,20</sup> or excessive joint instability.<sup>11</sup>  
353 This was typically observed, regardless of the technique employed: anterior drawer  
354 test<sup>8,11,14,20,22,24,31,40,51,52</sup> (sensitivity range 12–80%, specificity range 67–100%);  
355 anterolateral drawer test<sup>40</sup> (47% sensitivity and 99% specificity); reverse anterolateral  
356 drawer test<sup>40,42</sup> (sensitivity range 83–89%, specificity range 70–90%). The talar tilt  
357 test<sup>20,22,51</sup> and the heel adduction test<sup>20</sup> were also more specific than sensitive for  
358 diagnosing any lateral ligamentous injury<sup>20,31</sup> or injury to the CFL<sup>22,51</sup> displaying 17–66%  
359 sensitivity with 82–100% specificity, and 35% sensitivity with 77% specificity,  
360 respectively. Conversely, the supination test<sup>20,42</sup> proved more sensitive (73–98%) than  
361 specific (4–23%) for diagnosing ATFL injury<sup>42</sup> or any lateral ligamentous injury.<sup>20</sup>  
362

### 363 Palpation

364 Palpation is more sensitive than specific. Anterolateral talar palpation<sup>24</sup> displayed a  
365 perfect sensitivity (100%) and 80% specificity for diagnosing injury to the ATFL. Direct  
366 palpation of the ATFL<sup>14,20,27,42,60</sup> consistently showed high sensitivity (95–100%) across six  
367 studies but low (0–32%) specificity when diagnosing ATFL rupture<sup>16,27,42,60</sup> or any affected  
368 lateral collateral ligament.<sup>14,20</sup> Palpation of the CFL<sup>14,16,20,27,60</sup> had worse sensitivity,  
369 ranging between 49–100%, whilst specificity ranged between 26–79% for diagnosing  
370 partial to total tearing of the CFL<sup>16,27,42,60</sup> or any lateral ligamentous tear.<sup>14,20</sup>  
371

371

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

374

375 **Meta-analysis**

376 Six studies (885 observations) examining the anterior drawer test were included in our  
377 meta-analysis.<sup>11,14,22,40,51,52</sup> Using a bivariate model, the pooled metrics for the anterior  
378 drawer test were: sensitivity 54% (95% CI 35 to 71%); specificity 87% (95% CI 63 to 96%);  
379 LR+ 3.97 (95% CI 1.50 to 10.47); and LR- 0.54 (95% CI 0.39 to 0.75) ( $n = 6$ ). Sensitivity  
380 and specificity were negatively correlated (-0.73). When modelled independently,  
381 sensitivity displayed significant heterogeneity ( $I$ -square = 94.17%, Cochran's Q p-value <  
382 0.001) and specificity displayed substantial heterogeneity ( $I$ -square = 62.083%,  
383 Cochran's Q p-value = 0.022). It is plausible that a threshold effect in test interpretation  
384 (i.e. the amount of laxity required during translation for the clinician to say that the  
385 patient is injured) explains some of the between-study variations in sensitivity and  
386 specificity.<sup>64</sup> A threshold effect is further supported by the distance of the studies from the  
387 summary curve and the prediction ellipse (Figure 4).<sup>45</sup>

388

389

[FIGURE 4]

390

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

396

397

[FIGURE 5]

398

399 **Assessing the degree of ligamentous injury**

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

407

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

415

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

424

425

## DISCUSSION

426

### 427 **Principal findings**

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

439

### 440 **Explanations and implications for clinicians**

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

454

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

473

### 474 **Strength and limitations**

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



480 resonance imaging, and arthroscopy). Although only two studies incorporated the current  
481 gold standard reference (arthroscopy or surgery), a previous meta-analysis show that high  
482 diagnostic accuracy is possible using MRI, ultrasound or stress radiography (81–99%  
483 sensitivity and 79–91% specificity).<sup>7</sup> Still, as these reference standards are not perfect  
484 (and showcase variability), the diagnostic accuracy of the clinical tests of many of our  
485 included studies should be interpreted accordingly. Only three out of the 14 studies that  
486 we included had a low risk of bias across all QUADAS-2 domains. Verification bias was  
487 the most frequent, either due to improper time frames between the index and reference  
488 test or selective criteria. The generalisation of our findings is also affected by poor  
489 reporting of test interpretation, being commonly ambiguous and presenting with an  
490 unclear risk of bias. Only one study made direct comparisons between modified  
491 techniques for routine stress tests,<sup>40</sup> and just two studies incorporated an ordinal scale to  
492 grade injury severity.<sup>8,22</sup> As their results were contradictory, it is unclear if clinical tests of  
493 the talocrural joint can grade ligament damage beyond the binary. This review focuses on  
494 lateral ligament injuries, but we acknowledge that ankle sprains can also involve the ankle  
495 syndesmosis. Injuries to the syndesmosis will often have a different injuring  
496 mechanism,<sup>41</sup> and are assessed through alternative clinical tests featured in previous  
497 diagnostic reviews.<sup>62</sup> Although our meta-analysis excluded studies at a high risk of bias,  
498 the generalisability of our reported pooled diagnostic estimates to any specific setting  
499 might still be limited by reported differences in test technique, time since injury,  
500 reference standard used, and potential differences in remittance time. Lastly, our  
501 proposed diagnostic algorithm of performing palpation and anterior drawer testing of the  
502 ATFL for accurate diagnosis has not yet been validated with patient paired data.  
503

#### **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

504

## FUNDING

505

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

507

## DATA AVAILABILITY STATEMENT

508

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

## REFERENCES

- 512
- 513 1. Avci S, Şayli U. Comparison of the results of short-term rigid and semi-rigid cast  
514 immobilization for the treatment of grade 3 inversion injuries of the ankle. *Injury*.  
515 1998;29(8):581-584.
- 516 2. Baltes TPA, Arnáiz J, Geertsema L, et al. Diagnostic value of ultrasonography in  
517 acute lateral and syndesmotic ligamentous ankle injuries. *Eur Radiol*. Published  
518 online October 7, 2020.
- 519 3. Becker HP, Komischke A, Danz B, Bense R, Claes L. Stress diagnostics of the  
520 sprained ankle: Evaluation of the anterior drawer test with and without  
521 anesthesia. *Foot Ankle*. 1993;14(8):459-464. PMID:8253439
- 522 4. Bleakley C, Wagemans J, Netterström-Wedin F. Understanding chronic ankle  
523 instability: model rich, data poor. *Br J Sports Med*. Published online 2020:bjsports-  
524 2020-103311.
- 525 5. Bridgman SA, Clement D, Downing A, Walley G, Phair I, Maffulli N. Population  
526 based epidemiology of ankle sprains attending accident and emergency units in  
527 the West Midlands of England, and a survey of UK practice for severe ankle  
528 sprains. *Emerg Med J*. 2003;20(6):508-510. PMID:14623833
- 529 6. Broström L. Sprained ankles. VI. Surgical treatment of "chronic" ligament ruptures.  
530 *Acta Chir Scand*. 1966;132(5):551-565. PMID:5339635
- 531 7. Cao S, Wang C, Ma X, Wang X, Huang J, Zhang C. Imaging diagnosis for chronic  
532 lateral ankle ligament injury: A systemic review with meta-analysis. *J Orthop Surg  
533 Res*. 2018;13(1).
- 534 8. Cho JH, Lee DH, Song HK, Bang JY, Lee KT, Park YU. Value of stress ultrasound for  
535 the diagnosis of chronic ankle instability compared to manual anterior drawer test,  
536 stress radiography, magnetic resonance imaging, and arthroscopy. *Knee Surgery,  
537 Sport Traumatol Arthrosc*. 2016;24(4):1022-1028.
- 538 9. Cochrane. Review Manager (RevMan) [Computer program]. Version 5.4.  
539 Copenhagen: The Nordic Cochrane Centre. *Cochrane Collab*. Published online  
540 2020.
- 541 10. Cooke MW, Lamb SE, Marsh J, Dale J. A survey of current consultant practice of  
542 treatment of severe ankle sprains in emergency departments in the United  
543 Kingdom. *Emerg Med J*. 2003;20(6):505-507. PMID:14623832
- 544 11. Croy T, Koppenhaver S, Saliba S, Hertel J. Anterior talocrural joint laxity: Diagnostic  
545 accuracy of the anterior drawer test of the ankle. *J Orthop Sports Phys Ther*.  
546 2013;43(12):911-919. PMID:24175608
- 547 12. Davidson M. The interpretation of diagnostic tests: A primer for physiotherapists.  
548 *Aust J Physiother*. 2002;48(3):227-232. PMID:12217073
- 549 13. Delahunt E, Bleakley CM, Bossard DS, et al. Clinical assessment of acute lateral  
550 ankle sprain injuries (ROAST): 2019 consensus statement and recommendations of  
551 the International Ankle Consortium. *Br J Sports Med*. 2018;52(20):1304-1310.  
552 PMID:29886432
- 553 14. Van Dijk CN, Lim LSL, Bossuyt PMM, Marti RK. Physical examination is sufficient for  
554 the diagnosis of sprained ankles. *J Bone Jt Surg - Ser B*. 1996;78(6):958-962.  
555 PMID:8951015
- 556 15. Doherty C, Bleakley C, Hertel J, Caulfield B, Ryan J, Delahunt E. Clinical Tests Have  
557 Limited Predictive Value for Chronic Ankle Instability When Conducted in the  
558 Acute Phase of a First-Time Lateral Ankle Sprain Injury. *Arch Phys Med Rehabil*.  
559 2018;99(4):720-725.e1.
- 560 16. van der Ent FW. Lateral Ankle Ligament Injury: An Experimental and Clinical Study.

- 561         *Johnsen & Johnsen*. Published online 1984.
- 562 17. Fagan TJ. Nomogram for Bayes's Theorem. *N Engl J Med*. 1975;293(5):257.
- 563 18. Freeman SC, Kerby CR, Patel A, Cooper NJ, Quinn T, Sutton AJ. Development of an  
564 interactive web-based tool to conduct and interrogate meta-analysis of diagnostic  
565 test accuracy studies: MetaDTA. *BMC Med Res Methodol*. 2019;19(1):81.  
566 PMID:30999861
- 567 19. Frey C, Bell J, Teresi L, Kerr R, Feder K. A comparison of MRI and clinical  
568 examination of acute lateral ankle sprains. *Foot Ankle Int*. 1996;17(9):533-537.
- 569 20. Funder V, Jørgensen JP, Andersen A, et al. Ruptures of the lateral ligaments of the  
570 ankle: Clinical diagnosis. *Acta Orthop*. 1982;53(6):997-1000. PMID:7180412
- 571 21. Gaebler C, Kukla C, Breitenheher MJ, et al. Diagnosis of lateral ankle ligament  
572 injuries. Comparison between talar tilt, MRI and operative findings in 112 athletes.  
573 *Acta Orthop Scand*. 1997;68(3):286-290.
- 574 22. George J, Jaafar Z, Hairi IR, Hussein KH. The correlation between clinical and  
575 ultrasound evaluation of anterior talofibular ligament and calcaneofibular  
576 ligament tears in athletes. *J Sports Med Phys Fitness*. 2020;60(5):47-76.
- 577 23. Golanó P, Vega J, de Leeuw PAJ, et al. Anatomy of the ankle ligaments: a pictorial  
578 essay. *Knee Surgery, Sport Traumatol Arthrosc*. 2016;24(4):944-956.  
579 PMID:27052302
- 580 24. Gomes JLE, Soares AF, Bastiani CE, de Castro JV. Anterolateral talar palpation: A  
581 complementary test for ankle instability. *Foot Ankle Surg*. 2018;24(6):486-489.
- 582 25. Gondim Teixeira PA, Formery AS, Balazuc G, et al. Comparison between subtalar  
583 joint quantitative kinematic 4-D CT parameters in healthy volunteers and patients  
584 with joint stiffness or chronic ankle instability: A preliminary study. *Eur J Radiol*.  
585 2019;114:76-84. PMID:31005181
- 586 26. Graphpad Software. GraphPad QuickCalcs.
- 587 27. Gremeaux V, Coudreuse JM, Collado H, et al. Comparative study of clinical and  
588 ultrasonographic evaluation of lateral collateral ligament sprains of the ankle. *J*  
589 *Sports Med Phys Fitness*. 2009;49(3):285-291. PMID:19861934
- 590 28. Gribble PA, Bleakley CM, Caulfield BM, et al. Evidence review for the 2016  
591 International Ankle Consortium consensus statement on the prevalence, impact  
592 and long-term consequences of lateral ankle sprains. *Br J Sports Med*.  
593 2016;50(24):1496-1505. PMID:27259753
- 594 29. Hertel J, Denegar CR, Monroe MM, Stokes WL. Talocrural and subtalar joint  
595 instability after lateral ankle sprain. *Med Sci Sports Exerc*. 1999;31(11):1501-1508.
- 596 30. Higgins J, Green S. Cochrane handbook for systematic reviews of interventions.  
597 Version 5.1.0. Oxford: The Cochrane Collaboration. [https://handbook-5-  
598 1.cochrane.org/chapter\\_9/9\\_5\\_2\\_identifying\\_and\\_measuring\\_heterogeneity.htm](https://handbook-5-1.cochrane.org/chapter_9/9_5_2_identifying_and_measuring_heterogeneity.htm)
- 599 31. van den Hoogenband CR, van Moppes FI, Stapert JWJL, Greep JM. Clinical  
600 diagnosis, arthrography, stress examination and surgical findings after inversion  
601 trauma of the ankle. *Arch Orthop Trauma Surg*. 1984;103(2). PMID:6477070
- 602 32. Hubbard TJ, Cordova M. Mechanical Instability After an Acute Lateral Ankle Sprain.  
603 *Arch Phys Med Rehabil*. 2009;90(7):1142-1146.
- 604 33. Johannsen A. Radiological diagnosis of lateral ligament lesion of the ankle: A  
605 comparison between talar tilt and anterior drawer sign. *Acta Orthop*.  
606 1978;49(3):295-301. PMID:685675
- 607 34. Kato T. The diagnosis and treatment of instability of the subtalar joint. *J Bone Jt*  
608 *Surg - Ser B*. 1995;77(3):400-406. PMID:7744924
- 609 35. Kikumoto T, Akatsuka K, Nakamura E, Ito W, Hirabayashi R, Edama M. Quantitative



- 610 evaluation method for clarifying ankle plantar flexion angles using anterior drawer  
611 and inversion stress tests: A cross-sectional study. *J Foot Ankle Res.* 2019;12(1).  
612 PMID:31073333
- 613 36. Korkala O, Lauttamus L, Tanskanen P. Lateral ligament injuries of the ankle. Results  
614 of primary surgical treatment. *Ann Chir Gynaecol.* 1982;71(3):161-163.  
615 PMID:7114782
- 616 37. Kovaleski JE, Norrell PM, Heitman RJ, Hollis JM, Pearsall IV AW. Knee and ankle  
617 position, anterior drawer laxity, and stiffness of the ankle complex. *J Athl Train.*  
618 2008;43(3):242-248.
- 619 38. Lahde S, Putkonen M, Puranen J, Raatikainen T. Examination of the sprained ankle:  
620 Anterior drawer test or arthrography? *Eur J Radiol.* 1988;8(4):255-257.  
621 PMID:3148464
- 622 39. Lee KT, Park YU, Jegal H, Park JW, Choi JP, Kim JS. New method of diagnosis for  
623 chronic ankle instability: Comparison of manual anterior drawer test, stress  
624 radiography and stress ultrasound. *Knee Surgery, Sport Traumatol Arthrosc.*  
625 2014;22(7):1701-1707.
- 626 40. Li Q, Tu Y, Chen J, et al. Reverse anterolateral drawer test is more sensitive and  
627 accurate for diagnosing chronic anterior talofibular ligament injury. *Knee Surgery,  
628 Sport Traumatol Arthrosc.* 2020;28(1):55-62. PMID:31559464
- 629 41. Lin CF, Gross MT, Weinhold P. Ankle syndesmosis injuries: Anatomy,  
630 biomechanics, mechanism of injury, and clinical guidelines for diagnosis and  
631 intervention. *J Orthop Sports Phys Ther.* 2006;36(6):372-384. PMID:16776487
- 632 42. Lindstrand A. Lateral lesions in sprained ankles. A clinical and roentgenological  
633 study with special reference to anterior instability of the talus. Published online  
634 1976.
- 635 43. Lindstrand A, Mortensson W. Anterior instability in the ankle joint following acute  
636 lateral sprain. *Acta radiol.* 1977;18(5):529-539.
- 637 44. Lohrer H, Nauck T, Gehring D, Wissler S, Braag B, Gollhofer A. Differences between  
638 mechanically stable and unstable chronic ankle instability subgroups when  
639 examined by arthrometer and FAAM-G. *J Orthop Surg Res.* 2015;10(1).  
640 PMID:25890204
- 641 45. Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. Chapter 10: Analysing  
642 and Presenting Results. In: *Cochrane Handbook for Systematic Reviews of  
643 Diagnostic Test Accuracy Version 1.0.* ; 2010:1-61.  
644 <http://methods.cochrane.org/sdt/handbook-dta-reviews>
- 645 46. McHugh ML. Interrater reliability: The kappa statistic. *Biochem Medica.*  
646 2012;22(3):276-282. PMID:23092060
- 647 47. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic  
648 reviews and meta-analyses: The PRISMA statement. *PLoS Med.*  
649 2009;6(7):e1000097. PMID:19621072
- 650 48. Nyska M, Amir H, Porath A, Dekel S. Radiological assessment of a modified  
651 anterior drawer test of the ankle. *Foot Ankle.* 1992;13(7):400-403. PMID:1427531
- 652 49. Pasapula C, Memarzadeh A, Devany A, et al. The Modified Anterior Drawer Test  
653 (MADT): A New Clinical Test for Assessing Subtalar Instability a Cadaveric and  
654 Clinical Assessment. *Clin Res Foot Ankle.* 2018;06(02).
- 655 50. Patel A, Cooper N, Freeman S, Sutton A. Graphical enhancements to summary  
656 receiver operating characteristic plots to facilitate the analysis and reporting of  
657 meta-analysis of diagnostic test accuracy data. *Res Synth Methods.* Published  
658 online August 12, 2020:jrsm.1439.

- 659 51. Prins JG. Diagnosis and treatment of injury to the lateral ligament of the ankle. A  
660 comparative clinical study. *Acta Chir Scand Suppl.* 1978;486:3-149. PMID:282741
- 661 52. Raatikainen T, Putkonen M, Puranen J. Arthrography, clinical examination, and  
662 stress radiograph in the diagnosis of acute injury to the lateral ligaments of the  
663 ankle. *Am J Sports Med.* 1992;20(1):2-6.
- 664 53. Rosen AB, Ko J, Brown CN. Diagnostic accuracy of instrumented and manual talar  
665 tilt tests in chronic ankle instability populations. *Scand J Med Sci Sport.*  
666 2015;25(2):e214-e221.
- 667 54. Ross SE, Guskiewicz KM, Gross MT, Yu B. Assessment tools for identifying  
668 functional limitations associated with functional ankle instability. *J Athl Train.*  
669 2008;43(1):44-50. PMID:18335012
- 670 55. Schmidt RL, Factor RE. Understanding Sources of Bias in Diagnostic Accuracy  
671 Studies. *Arch Pathol Lab Med.* 2013;137(4):558-565.
- 672 56. Schneiders A, Karas S. The accuracy of clinical tests in diagnosing ankle ligament  
673 injury. *Eur J Physiother.* 2016;18(4):245-253.
- 674 57. Schwieterman B, Haas D, Columber K, Knupp D, Cook C. Diagnostic accuracy of  
675 physical examination tests of the ankle/foot complex: a systematic review. *Int J*  
676 *Sports Phys Ther.* 2013;8(4):416-426. PMID:24175128
- 677 58. Shinkins B, Thompson M, Mallett S, Perera R. Diagnostic accuracy studies: how to  
678 report and analyse inconclusive test results. *BMJ.* 2013;346. PMID:23682043
- 679 59. Simel DL, Feussner JR, DeLong ER, Matchar DB. Intermediate, indeterminate, and  
680 uninterpretable diagnostic test results. *Med Decis Mak.* 1987;7(2):107-114.  
681 PMID:3574020
- 682 60. De Simoni C, Wetz HH, Zanetti M, Hodler J, Jacob H, Zollinger H. Clinical  
683 examination and magnetic resonance imaging in the assessment of ankle sprains  
684 treated with an orthosis. *Foot Ankle Int.* 1996;17(3):177-182.
- 685 61. Singh K, Lal Thukral C, Gupta K, Singh A. Comparison of high resolution  
686 ultrasonography with clinical findings in patients with ankle pain. *J Ultrason.*  
687 2018;18(75):316-324.
- 688 62. Sman AD, Hiller CE, Refshauge KM. Diagnostic accuracy of clinical tests for  
689 diagnosis of ankle syndesmosis injury: A systematic review. *Br J Sports Med.*  
690 2013;47(10):620-628.
- 691 63. Staples OS. Result study of ruptures of lateral ligaments of the ankle. *Clin Orthop*  
692 *Relat Res.* 1972;85:50-58. PMID:4624809
- 693 64. Trikalinos TA, Balion CM, Coleman CI, et al. Chapter 8: Meta-analysis of test  
694 performance when there is a "gold standard." *J Gen Intern Med.*  
695 2012;27(SUPPL.1). PMID:22648676
- 696 65. University of Bristol. QUADAS-2 : Background Document. School of Social and  
697 Community Medicine. Published 2014. [http://www.bristol.ac.uk/media-](http://www.bristol.ac.uk/media-library/sites/quadas/migrated/documents/background-doc.pdf)  
698 [library/sites/quadas/migrated/documents/background-doc.pdf](http://www.bristol.ac.uk/media-library/sites/quadas/migrated/documents/background-doc.pdf)
- 699 66. Wallace BC, Dahabreh IJ, Trikalinos TA, Lau J, Trow P, Schmid CH. Closing the gap  
700 between methodologists and end-users: R as a computational back-end. *J Stat*  
701 *Softw.* 2012;49:5.
- 702 67. Weindel S, Schmidt R, Rammelt S, Claes L, Campe A V., Rein S. Subtalar instability:  
703 A biomechanical cadaver study. *Arch Orthop Trauma Surg.* 2010;130(3):313-319.
- 704 68. Whiting PF, Rutjes AWS, Westwood ME, et al. Quadas-2: A revised tool for the  
705 quality assessment of diagnostic accuracy studies. *Ann Intern Med.*  
706 2011;155(8):529-536. PMID:22007046
- 707 69. Wiebking U, Pacha TO, Jagodzinski M. An accuracy evaluation of clinical,

708 arthrometric, and stress-sonographic acute ankle instability examinations. *Foot*  
709 *Ankle Surg.* 2015;21(1):42-48.  
710  
711

## FIGURE LEGENDS

712

713 [Figure 1](#)

714 (number of hits)

715

716 [Figure 2](#)

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

719

720 [Figure 3](#)

721 TP = True Positive

722 FP = False Positive

723 TN = True Negative

724 FN = False Negative

725 \*77 patients examined by two different examiners

726

727 [Figure 4](#)

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

731

732 [Figure 5](#)

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