Manual testing for ankle instability

Emily Jane Wilkin, Adrienne Hunt, Elizabeth Jean Nightingale*, Joanne Munn, Sharon Lynne Kilbreath, Kathryn Margaret Refshauge

Discipline of Physiotherapy, Faculty of Health Sciences, University of Sydney, Cumberland Campus, C42, 75 East St., Lidcombe, NSW 2141, Australia

ARTICLE INFO

Article history:
Received 16 August 2011
Received in revised form 2 March 2012
Accepted 7 March 2012

Keywords:
Ankle injury
Joint instability
Physical therapy modalities
Manual therapy

ABSTRACT

Aim: To assess inter-rater reliability of ankle manual tests. We also correlated the manual tests with the Cumberland Ankle Instability Tool (CAIT).

Method: One ankle from each of 60 participants was assessed using four different manual tests (anterior drawer in supine and crook lying, talar tilt, inversion tilt). Three different raters, varying in experience, tested each participant. The CAIT questionnaire was also administered. The study received ethics approval from the University of Sydney Human Research Ethics Committee. Intraclass correlation coefficients (ICC), standard error of the mean (SEM) and percent close agreement (PCA) were used to determine reliability of the four tests. Pearson’s correlation coefficients were used to determine relationships between the manual tests and CAIT scores.

Results: Inter-rater reliability for the four manual tests was poor regardless of therapist experience (ICC[1,1] −0.12 to 0.33; SEM 0.93–1.69). Correlations between the CAIT and manual tests were also low varying between r = −0.12 and −0.42.

Conclusions: Inter-rater reliability was poor for manual tests of ankle stability. Reliability may be improved by using a grading scale with fewer intervals. The CAIT scores and manual tests correlated poorly, potentially reflecting the variety of conditions leading to ankle instability.

Crown Copyright © 2012 Published by Elsevier Ltd. All rights reserved.

1. Introduction

Manual tests are a common clinical method for assessing ankle ligament integrity. The inter-rater reliability of these tests has not been evaluated. However, as instrumented approaches such as the Telos device, stress radiography and dynamic and quasi static ankle testers are costly, invasive and not readily available, clinicians rely on manual tests. Manual tests require a clinician to apply a manual load to the ankle, evaluate the range of motion and “end feel” and compare the findings with their experience of a normal joint.

Two commonly used tests to ascertain ankle stability are the anterior drawer and the talar tilt tests. These tests have been used to differentiate ankles with and without injury (Hertel et al., 1999; Denegar et al., 2002). Intra-rater reliability of both the anterior drawer and the talar tilt tests has been shown to be reliable when using a 3 to 5 point scale (Baumhauer et al., 1995; Hertel et al., 1999; Denegar et al., 2002). The only studies to investigate inter-rater reliability used cadaveric specimens, finding poor reliability between examiners (Fujii et al., 2000). Inter-rater reliability has not been established in vivo.

Issues that can affect reliability of manual tests are the force applied to a joint, position of the patient and the range of scale over which the stability is assessed. The force applied to the ankle for the execution of these tests has not been standardised. Since the reported variability of forces used between experienced therapists in spinal mobilisation is high (Harms and Bader, 1997) one would expect this to be a source of error and may contribute to poor inter-rater reliability. Positioning for manual tests potentially affects reliability by altering tension in the soft tissues of the tested joint and distorting the direction and magnitude of the applied force. The rating scale against which the movement is graded also varies among studies. Grading scales with positive and negative anchors and up to 5 point scales have been used (Baumhauer et al., 1995; Hertel et al., 1999; Denegar et al., 2002).

Ankle instability is a global term used to describe a range of symptoms and impairments, including “mechanical” instability, “perceived” instability and recurrent sprain. Two questionnaires which seek to assess the patient’s perception of instability are the Cumberland Ankle Instability Tool (CAIT) (Hiller et al., 2006) and the Ankle Instability Instrument (AII) (Docherty et al., 2006). Both instruments have excellent test retest reliability, (ICC 0.96 and 0.95 respectively). The CAIT has sensitivity of 82.9% and specificity of 74.7%. The Cronbach alpha coefficient is 0.89 for the AII.
The present study investigated the inter-rater reliability of ankle manual tests, including the contribution of clinical experience, due to the subjective nature of the test. A secondary aim was to determine the correlation between the manual tests and the CAIT to explore further the relationship between mechanical and perceived instability.

2. Methods

2.1. Participants

Sixty participants (51 females, 9 males) aged between 17 and 50 years volunteered to participate. There were no exclusion criteria. Ethical approval for the study was obtained from The University of Sydney Human Research Ethics Committee and written consent was gained.

Thirty-eight (63%) participants were sprainers, twenty-two (37%) were non-sprainers, and three (5%) participants reported additional conditions: ligament reconstructive surgery, previous malleolar fracture and congenital metatarsus varus. Twenty-one participants reported modifications due to their ankle symptoms with modifications including ankle bracing/strapping (n = 14), changing technique (n = 1), decreased activity level (n = 14), and treatments (icing, orthotics, analgesia, anti-inflammatories or topical creams, each n = 1). Some participants used multiple strategies.

2.2. Raters

Five raters from The University of Sydney participated: four raters were experienced physiotherapists and the fifth rater was an undergraduate student with limited clinical experience. Raters were supplied with a manual outlining the performance of each manual test and underwent a practise session to familiarise themselves with the tests and protocol.

2.3. Measures

2.3.1. Manual tests

An 8-point Likert scale was adopted to enable analysis as continuous data with anchors –2 (very stiff) to 5 (hypermobility) with 0 representing normal stability. This scale was chosen based on our pilot of visual analogue methods because it best represented physiotherapists' perceived ease of use.

i) Anterior Drawer in Supine (Brukner and Khan, 2002): Participants were supine with the limb extended. A posterolateral force was applied to the talus to displace the talus anteriorly with respect to the tibia.

ii) Anterior Drawer in Crook Lying (Magee, 2002): Participants were in crook lying with the knee flexed to approximately 90°. An anteroposterior force was applied to the distal shank to displace the tibia posteriorly on the talus.

iii) Talar Tilt (Brukner and Khan, 2002): Participants were prone, with the knee flexed to 90° and the ankle in neutral. Raters grasped either side of the ankle just distal to the malleoli, while the thumbs rested on the lateral aspect of the talus as an inversion force was applied.

iv) Inversion tilt (Magee, 2002): Using the same position as iii), raters cupped the rear foot and mid foot, applying an inversion force to move the ankle into inversion.

2.3.2. CAIT

Participants completed the CAIT questionnaire (Hiller et al., 2006) during the test session. The CAIT has a maximum score of 30 with a score of ≤27 indicative of perceived ankle instability.

2.4. Protocol

During a single assessment session, participants provided demographic information and completed the CAIT questionnaire to determine the test ankle. The test ankle was the one with the greatest perceived instability (i.e. the lowest CAIT score). For participants without instability, the test ankle was randomised. There were a total of four experienced raters and one novice rater. At each test session, the novice and two of the experienced raters, made judgements on ankle stability based on findings from using each of the four manual tests. Individual recording sheets were used to blind the raters to the CAIT scores and to other raters' judgements.

The order of raters and tests was randomised. Each rater made judgements on all manual tests prior to the next rater’s assessment. Between raters, participants were encouraged to walk for 4 min to minimise pre-conditioning or viscous deformation.

2.5. Statistical analysis

The Intraclass Correlation Coefficient (ICC[1,1]) with 95% confidence intervals was determined for each manual test between the experienced raters and among all three raters (i.e. 2 experienced raters and the novice rater). ICCs were interpreted according to Fleiss (1986), whereby coefficients <0.4 were considered poor, between 0.4 and 0.75 were fair/good, and coefficients >0.75 were excellent.

The standard error of the measurement (SEM) was also calculated for each manual test, using: \[ SEM = SD \times \sqrt{(1 - r_{xx})} \] in which SD is the standard deviation of the scores and \( r_{xx} \) is the ICC[1,1] for the tests (Portney and Watkins, 2000). Lastly, percent close agreement, which measures the frequency of agreement between raters on individual participant scores, was determined (Portney and Watkins, 2000).

The relationship between scores from the manual tests and the CAIT scores were determined using Pearson’s \( r \) correlations, using scores from the novice examiner who tested all participants (n = 60), and the experienced examiner who had tested the most participants (n = 53). Descriptors of the strength of the correlation were interpreted using Munro’s (1997) guidelines (0–0.25 no correlation, 0.26–0.49 low, 0.5–0.69 moderate, 0.7–0.89 high, and 0.9–1.0 very high correlation).

3. Results

The mean score for the CAIT amongst the sprainers was 22.3 (range 9–30), indicative of an ankle perceived as unstable, on average, and for the non-sprainers was 27 (range 16–30). Based on the CAIT score, 7 (18%) of the 38 sprainers perceived their ankle as stable despite their sprain history and 14 (64%) of the non-sprainers considered their ankle stable.

3.1. Reliability

Inter-rater reliability of all four manual tests was poor (Table 1). The ICC[1,1] ranged between 0.06 and 0.33 among the three raters (1 novice and 2 experienced) and the SEM ranged from 0.93 to 1.39. The ICC[1,1] was highest for the talar tilt (ICC 0.33, 95% CI 0.17–0.50) and lowest for the anterior drawer test in crook lying (ICC 0.06, 95% CI –0.08–0.23). The reliability of these tests remained poor even when the novice rater was removed from the comparisons (ICC[1,1] –0.12 to 0.26). Similarly, the SEM ranged from 1.04 to 1.69.

The absolute agreement among the three raters ranged from 5% for the anterior drawer in crook lying to 32% for the talar tilt test (Fig. 1). The 80% percent close agreement among the three raters for
the anterior drawer test in supine, talar tilt and inversion tilt tests was within 2 units on the Likert scale. Measurement of the tilt tests were within 1 unit 78% of the time.

3.2. Correlation between manual testing and CAIT scores

Correlations were poor between the CAIT score and both the novice’s ($r = -0.11$ to $-0.09$) judgements and the experienced raters’ ($r = 0.12$ to $-0.42$) judgements of ankle instability (Table 2). The higher correlations were achieved by the experienced rater between the CAIT score and the talar tilt and inversion tilt tests.

4. Discussion

Manual tests are specifically aimed at assessing ligamentous integrity, or “mechanical instability” at the ankle. Our findings demonstrated that, using an 8-point Likert scale, inter-rater reliability was poor, regardless of therapist experience. A 3-point scale is the smallest scale worth evaluating, and is highly likely to achieve excellent reliability because the choice is so limited. We chose to use an 8-point Likert scale to force clinicians to make more finely graded decisions about ankle stability. In this way, the data could also be used as continuous data for analysis. However, the small, but clinically meaningful, changes in joint arthrokinematic motion resulting from ankle injury may be too small in magnitude for clinicians to perceive manually. Furthermore, the judgements on an 8-point Likert scale were too difficult to differentiate, with 80% agreement on most tests being at least 2 points apart. These scoring differences would result in judgements of quite different levels of instability. We therefore do not recommend use of such a finely graded scale.

The difficulty in testing mechanical stability has been the focus of much debate (Frost and Amendola, 1999; Spahn, 2004; Hubbard and Hicks-Little, 2008). Intra-rater reliability of manual tests has been demonstrated (Baumhauer et al., 1995; Denegar et al., 2002; Spahn, 2004). These tests are also more sensitive in ruling out ligament injury, with poorer sensitivity in distinguishing injury severity (Hertel et al., 1999; Denegar et al., 2002; Spahn, 2004). Inter-rater reliability has not previously been fully investigated (Spahn, 2004).

Poor inter-rater reliability may be improved by the use of instrumented measures, such as the Telos stress device or Hollis

---

### Table 1

<table>
<thead>
<tr>
<th>Test</th>
<th>ICC (95% CI) between the 3 raters</th>
<th>SEM</th>
<th>ICC (95% CI) between the experienced raters</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anterior drawer in supine</td>
<td>0.16 (0.10 to 0.33)</td>
<td>1.11</td>
<td>0.23 (−0.02 to 0.46)</td>
<td>1.19</td>
</tr>
<tr>
<td>Anterior drawer in crook lying</td>
<td>0.06 (−0.08 to 0.23)</td>
<td>1.39</td>
<td>−0.12 (−0.36 to 0.14)</td>
<td>1.69</td>
</tr>
<tr>
<td>Talar tilt</td>
<td>0.33 (0.17 to 0.50)</td>
<td>0.93</td>
<td>0.22 (−0.02 to 0.45)</td>
<td>1.06</td>
</tr>
<tr>
<td>Inversion tilt</td>
<td>0.29 (0.13 to 0.46)</td>
<td>0.98</td>
<td>0.26 (0.00 to 0.48)</td>
<td>1.04</td>
</tr>
</tbody>
</table>

---

### Table 2

<table>
<thead>
<tr>
<th>Test</th>
<th>Novice rater</th>
<th>Experienced rater</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anterior drawer in supine</td>
<td>0.016</td>
<td>−0.124</td>
</tr>
<tr>
<td>Anterior drawer in crook lying</td>
<td>0.088</td>
<td>−0.141</td>
</tr>
<tr>
<td>Talar tilt</td>
<td>−0.113</td>
<td>−0.348</td>
</tr>
<tr>
<td>Inversion tilt</td>
<td>−0.049</td>
<td>−0.420</td>
</tr>
</tbody>
</table>

---

**Fig. 1.** Percent close agreement (cumulative frequency) versus absolute difference for all three raters for each of the four tests.
ankle arthrometer to standarise the force applied. However, a review of stress radiography utilising instrumented force production found sufficient variability that acceptable normal values could not be determined (Frost and Amendola, 1999) compounding the difficulty of subjectively rating laxity. Taken together, these findings confirm the opinion of Hertel et al. (1999), that manual tests be used to rule out ligament injury and little else.

Manual tests of instability have high intra-rater reliability and the CAIT has excellent psychometric properties. We therefore correlated the CAIT score, measuring perceived instability, with one rater’s manual test scores, of mechanical stability, to explore the links between different aspects of ankle instability. We found the correlation to be poor. However, ankle instability may well be a combination of perceived giving way, neuromuscular changes, proprioceptive impairments, recurrent injury and joint laxity. Since the normal function of the anterior talofibular ligament is to act as a restraint to end range of motion (Milner and Soames, 1997), instability in mid-range is unlikely due to ligament insufficiency. The lack of correlation was therefore unsurprising, and reflect that manual stress tests measure a single construct of ankle instability.

Original research and systematic reviews have shown that the manual tests are unreliable, both for experienced and inexperienced clinicians. Manual tests rely on identification of changes in talocrural and/or subtalar joint laxity that are small, but clinically meaningful, and that statistically significant changes in joint laxity are difficult to identify using manual methods and Likert scales. There continues to be a clear need for a readily available, reliable, safe and quantifiable method for measuring talocrural and subtalar joint laxity so that real differences between groups can be detected.

5. Conclusions

Manual testing of ankle ligament laxity has poor inter-rater reliability. The variability in the tests themselves, the applied force and subjective judgement of the degree of movement are the most likely sources of variability. Reliability is highly likely to improve if raters merely rate the test as normal or abnormal. In their current form, manual tests are unreliable for identifying degrees of mechanical instability.

There was only a poor correlation between manual test and CAIT scores. This finding has support from current studies investigating the lack of relationship between perceived and mechanical instability and suggests that perceived instability status cannot be inferred from tests of mechanical instability.

References