A study to explore the reliability and precision of intra and inter-rater measures of ULNT1 on an asymptomatic population

George S. Oliver*, Alison Rushton

*Department of Physiotherapy, Cardiff and Vale UHB, Cardiff, UK

1 School of Health and Population Sciences, University of Birmingham, Birmingham, UK

ARTICLE INFO

Article info

Received 15 August 2009
Received in revised form 15 May 2010
Accepted 19 May 2010

Keywords:
Neurodynamics
Upper Limb Neurodynamic Tests (ULNT)
Reliability
Precision

ABSTRACT

Upper Limb Neurodynamic Test 1 (ULNT1) is commonly used within clinical practice. However, the existing evidence regarding its reliability is conflicting and raises methodological questions. Therefore, the aim of this study was to investigate how reliable and precise physiotherapists are at recording both intra and inter-rater measurements of ULNT1 on an asymptomatic population. Forty asymptomatic subjects, 29 females and 11 males (18–42 years, mean 23.35), were recruited into this intra (stability) and inter-rater (equivalence) reliability and precision study. ULNT1 was recorded twice using an electrogoniometer by two experienced physiotherapists using a standardised operational description in conditions replicating clinical practice. Reliability was analysed using the Intraclass Correlation Coefficient (ICC), and precision using Standard Error of Measurement (SEM) and Smallest Detectable Difference (SDD). The findings demonstrated excellent intra-rater (ICC2,1 0.98 Rater 1; ICC2,1 0.96 Rater 2) and good inter-rater (ICC2,1 0.80) reliability. Precision was acceptable for both intra-rater (SEM 2.59° Rater 1; SEM 0.97° Rater 2; SDD 7.16° Rater 1; SDD 2.68° Rater 2), and inter-rater (SEM 3.83°; SDD 10.58°) measurements. These findings demonstrate that physiotherapists can use ULNT1 reliably and with precision for intra and inter-rater measurements of asymptomatic subjects in conditions that replicate clinical practice. The reproduction of this study on a population of symptomatic subjects is now warranted.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

Physiotherapists commonly use the Upper Limb Neurodynamic Test 1 (ULNT1) both as a diagnostic tool and outcome measure (Butler and Gifford, 1989) in patients with neurogenic upper limb symptomology. Reproduction of symptoms and available range of motion (ROM) are the important criteria in its assessment (Butler and Gifford, 1989), with ROM for elbow extension deficits during ULNT1 of between 17 and 53° demonstrated (Pullos, 1986; Hines et al., 1993; Block et al., 1998). These criteria are assessed pre and post intervention, enabling evaluation of outcome (Sweeney and Harms, 1996). This requires the same, or in some cases different physiotherapists, to record two measures of a patient’s ULNT1; and to have confidence in its use reliability and precision of intra and inter-rater measurement is essential.

The ULNT1 aims to assess the mechanical and neurophysiological integrity of the median nerve (Shacklock, 2005), and has support within the literature regarding its validity (Kleinrensink et al., 1995; Wright et al., 1996; Lewis et al., 1998; Coppieters et al., 2001). In contrast, the literature relating to its reliability appears inconclusive. Studies investigating intra-rater reliability have claimed good (ICC2,1 0.83) (Selvaratnam et al., 1994) to excellent reliability (ICC2,1 0.98–0.99) (Coppieters et al., 2002) but the methodological quality of these studies has been questioned with limited sized and heterogenous samples. The literature relating to inter-rater reliability is also problematic owing to inappropriate data analysis and possible practice effects, and conflicting findings with reports of low (Hines et al., 1993) and good to excellent (Coppieters et al., 2002) reliability.

Precision of the ULNT1 (degree of mutual agreement among a series of individual measurements) (Sim and Wright, 2000), is important in addition to an Intraclass Correlation Coefficient (ICC) evaluating reliability, as the ICC does not convey the expected magnitude of measurement error; thus a reliable measurement tool could be producing a large magnitude of undisclosed measurement error (Haas, 1991). Varying levels of intra-rater precision have been demonstrated (Pullos, 1986; Hines et al., 1993; Block et al., 1998).
demonstrated, ranging from unacceptable (Selvaratnam et al., 1994) (Standard Error of Measurement (SEM) 16.9°, ±33° 95% Confidence Interval (CI)) to acceptable (Coppieeters et al., 2002) (SEM 2.2°–2.9°, 3.0°–6.4° 95% CI). Evidence for inter-rater precision is limited with one study reporting SEM 2.2°–3.7° to date (Coppieeters et al., 2002).

The aim of this study was to investigate the reliability and precision of intra and inter-rater ULNT1 measurements in a population of asymptomatic subjects in conditions that replicate those used in the clinical setting.

2. Methods

2.1. Design and raters

An intra (stability) and inter-rater (equivalence) reliability and precision study was designed, and ethical approval gained. Two raters (6 and 10 years post qualification experience) were randomly selected (drawing names from a hat) from consenting students studying the MSc Advancing Practice (Specialist Manipulative Physiotherapy) programme, at the University of Birmingham.

2.2. Subjects

Forty subjects (29 females/11 males) aged 18–42 (mean 23.35 years) were recruited using posters to ensure freedom required for ICC(2,1) (Chinn, 1991). Application of inclusion and exclusion criteria ensured a relatively homogenous sample, limiting the between subject variation and preventing over inflation of the ICC (Rankin and Stokes, 1998).

Inclusion criteria:

- Undergraduate physiotherapy students

Exclusion criteria:

- Ongoing cervical or upper limb symptomology (Selvaratnam et al., 1994)
- History of, or treatment for, cervical or upper limb neurogenic condition (Coppieeters et al., 2001)
- Upper limb joint restriction that prevents measures required for ULNT1 being taken (Coppieeters et al., 1999)
- Subjects unable to assume test position comfortably for the duration of testing (Coppieeters et al., 2002)

2.3. Equipment

A Biometrics K100 Electrogoniometer was used, which has demonstrated acceptable inter-rater reliability, with no significant differences in elbow ROM when inter-rater measurements were recorded (Goodwin et al., 1992), and acceptable levels of precision with measurement errors up to 3° (Lantz et al., 2003).

2.4. Procedure

Subjects were provided with a participant information sheet and provided written informed consent. The raters were provided with information regarding their role, but were not made aware of the exact study aims. The raters were provided with a written outline of the testing procedure, and were able to practice the procedure on a model without feedback, to control for possible training effects, and ensure familiarity.

Subjects were positioned in supine on a plinth, without a pillow and with their cervical spine maintained in neutral. Previous authors found no difference in elbow extension between left or right arm for ULNT1 (Kennelly, 1985; Rubenach, 1985; Bell, 1987; van der Heide et al., 2001), and consequently the right arm was used arbitrarily, with the subjects’ left arms positioned across their abdomen. A standardised procedure calibrated and applied the electrogoniometer for each subject using a handheld goniometer at 0°, 90° and 135° elbow flexion (Coppieeters et al., 1999).

The sequence for the ULNT1 followed the operational description formulated by Butler (2000), that is widely accepted (Coppieeters et al., 2002). The first onset of pain or tingling described by the subject (P1) was noted, and the position when pain or tingling increased and the subject wanted the test to be ceased (P2) was defined as the end point of the test (van der Heide et al., 2001). For reporting situations where P2 was not reached, the end point occurred when resistance prevented further elbow extension (R2).

The researcher recorded the electrogoniometer measurements, enabling blinding of the raters from their own and each other’s measurements. In an attempt to control for the effect of serial measures, the first three measures on each subject were discarded (informing by a pilot study (n = 10) which demonstrated smaller interquartile ranges for measures 4–7 compared to 1–3). Following these, discarded measures, each rater recorded two measurements of each subjects’ ULNT1. Rater 1 performed the discarded measures prior to recording measures 4–5 on the first subject, with rater 2 recording measures 6–7. To enable counterbalancing, the raters alternated performing measures 1–5 and 6–7, to control for possible order effects.

2.5. Data analysis

2.5.1. Descriptive analysis

Sums, means, Differences (Diff) and Standard Deviations (SD) were calculated using SPSS 14.0 statistical package for the intra and inter-rater measurements of elbow extension ROM deficit. Intra-rater Diff was calculated by subtracting measure 2 from measure 1, and inter-rater Diff by subtracting measure 1 rater 1 from measure 1 rater 2. The individual means and differences scores were used to produce intra and inter-rater means-vs-differences plots (Fig. 1).

2.5.2. Inferential analysis

The ICC(2,1) was used to evaluate intra and inter-rater reliability (Sim and Wright, 2000). Intra-rater reliability was calculated using both the first and second measurements that the raters recorded from each subject. Each rater’s first measure was used in the calculation of inter-rater reliability, to control for possible practice effect that may have occurred with use of their second measurement. When interpreting the study’s reliability findings the following criteria were used:

- < 0.75 Poor to moderate
- 0.75–0.90 Good
- 0.91–1 Adequate reliability for clinical measurement (Portney and Watkins, 1993)

Previous literature (Coppieeters et al., 2002) has described ‘adequate reliability for clinical measurement’ as ‘excellent reliability’. This term will therefore be used to aid comparison with existing literature.

The SEM, Smallest Detectable Difference (SDD) and 95% range enabled comparison to existing literature. Peat et al. (2002) describe SEM as a measurement of the within subject re-test variation (SDpooled × 1–ICC(2,1)) (Safrit and Wood, 1989). The SDD provides evidence regarding the smallest degree of change...
3. Results

3.1. Descriptive analysis

The descriptive data demonstrates combined means and SD for the intra and inter-rater measures (Table 1). The SD demonstrates that the variation between measurements 1 and 2 for rater 1 (1.45) and rater 2 (2.44) was low. In contrast, the SD for rater 1 measure 1 compared to rater 2 measure 1 was 5.91°. In Fig. 1, the narrow scatter either side of the line of no difference demonstrates the measures to be repeatable, the error random, and that a systematic bias is not suggested.

3.2. Inferential analysis

Excellent intra-rater (ICC2,1 Rater 1 = 0.98, Rater 2 = 0.96) and good inter-rater (ICC2,1 0.80) reliability were demonstrated (Table 2). When the mean of three measures was used, the inter-rater ICC2,1 improved from 0.80 to 0.89. Both intra-rater (SEM Rater 1 = 2.59°, Rater 2 = 0.97°) and inter-rater precision (SEM 3.83°) were acceptable. The SDD was 7.16° and 2.68° for raters 1 and 2 respectively, and 10.58° for inter-rater measures, providing an indication of differences in ULNT1 measures that could be considered significant within the clinical setting.

4. Discussion

The excellent intra-rater reliability (ICC2,1 0.98 Rater 1; 0.96 Rater 2) of ULNT1 improves upon previous claims of good reliability (ICC2,1 0.83) (Selvaratnam et al., 1994), and support excellent reliability findings (ICC2,1 0.98–0.99) (Coppieters et al., 2002), with key methodological limitations of the previous studies addressed. This study’s results support the inter-rater reliability for single measures to be good (ICC2,1 0.80) (Coppieters et al., 2002). However, if the mean of three measures were used, the ICC2,1 improved to 0.89, bordering excellent reliability. This opens the clinical debate regarding use of mean rather than single inter-rater measures to provide greater reliability, but at the expense of increased examination time and possible changes in asymptomatic population from repeated testing.

In line with Coppieters et al. (2002), acceptable levels of intra and inter-rater precision were demonstrated with all SEMs below the 5° limit accepted by the American Medical Association (Table 2) for the evaluation of movement in the clinical context (Nitschke et al., 1999). Interpretation of these findings means that a change in measurement of >7.16° for intra-rater or 10.58° for inter-rater (SD, Table 2) can be considered the result of a clinical change or an applied intervention and not the result of measurement error.

4.1. Limitations

Although the use of a probability sample would have strengthened the methodology, Armitage and Berry (1994) argue that poster recruitment provides a sample subject to random, unsystematic variation, and therefore comparable to data from a probability sample. The ICC using n = 40, was above the minimum
recommended by Chinn (1991), but with greater resources \( n = 60–70 \) would have allowed a more precise estimation of the variance from the ICC (Peat et al., 2002). Previous studies have described the normal end elbow ROM for ULNT1 as 17–53° (Pullos, 1986; Hines et al., 1993; Block et al., 1998), in contrast to the current recommendation by Chinn (1991), but with greater resources.

Table 2

<table>
<thead>
<tr>
<th>Intra-rater (R1 M1&amp;2)</th>
<th>SEM</th>
<th>ICC(2,1)</th>
<th>95% Range</th>
<th>SDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra-rater (R2 M1&amp;2)</td>
<td>0.98</td>
<td>0.99</td>
<td>2.58°</td>
<td>±5.08°</td>
</tr>
<tr>
<td>Intra-rater (R1 M1&amp;2)</td>
<td>0.96</td>
<td>0.98</td>
<td>0.97°</td>
<td>±1.90°</td>
</tr>
<tr>
<td>Inter-rater (R1 M1&amp; R2 M1)</td>
<td>0.80</td>
<td>0.89</td>
<td>3.83°</td>
<td>±7.51°</td>
</tr>
</tbody>
</table>

R – rater 1; R2 – rater 2; M1 – measurement 1; M2 – measurement 2; SEM – Standard Error of Measurement; ICC – Intraclass Correlation Coefficient; SDD – Smallest Detectable Difference.

5. Conclusion

Physiotherapists are able to take intra and inter-rater measures of ULNT1 reliably on asymptomatic subjects in conditions that replicate daily practice. In addition, the precision of these measures was acceptable for its use as a measurement tool. With this demonstrated reliability and precision it is possible to support the use of ULNT1 as an outcome measure on asymptomatic subjects. It is vital that this methodologically strong evidence is developed further, with the investigation of symptomatic subjects, in order to fully evaluate the clinical usefulness of ULNT1 as an outcome measure within clinical practice.

References


