Inter-rater reliability for measurement of passive physiological movements in lower extremity joints is generally low: a systematic review

Emiel van Trijffel[1], Rachel J van de Pol[2], Rob AB Oostendorp[3] and Cees Lucas[1]

[1]University of Amsterdam, Academic Medical Centre, [2]Private Practice for Physiotherapy, The Hague, [3]Radboud University Nijmegen Medical Centre The Netherlands

Question: What is the inter-rater reliability for measurements of passive physiological or accessory movements in lower extremity joints? **Design:** Systematic review of studies of inter-rater reliability. **Participants:** Individuals with and without lower extremity disorders. **Outcome measures:** Range of motion and end-feel using methods feasible in daily practice. **Results:** 17 studies were included of which 5 demonstrated acceptable inter-rater reliability. Reliability of measurements of physiological range of motion ranged from Kappa −0.02 for measuring knee extension using a goniometer to ICC 0.97 for measuring knee flexion using vision. Measuring range of knee flexion consistently yielded acceptable reliability using either vision or instruments. Measurements of end-feel were unreliable for all hip and knee movements. Two studies satisfied all criteria for internal validity while reporting acceptable reliability for measuring physiological range of knee flexion and extension. Overall, however, methodological quality of included studies was poor. **Conclusion:** Inter-rater reliability of measurement of passive movements in lower extremity joints is generally low. We provide specific recommendations for the conduct and reporting of future research. Awaiting new evidence, clinicians should be cautious when relying on results from measurements of passive movements in joints for making decisions about patients with lower extremity disorders. **[van Trijffel E, van de Pol RJ, Oostendorp RAB, Lucas C (2010) Inter-rater reliability for measurement of passive physiological movements in lower extremity joints is generally low: a systematic review. *Journal of Physiotherapy* 56: 223–235]**

Key words: Lower extremity, Reliability, Reproducibility of results, Physical examination, Range of motion, Passive motion, Physiotherapy

Introduction

Physiotherapists commonly assess and treat patients with lower extremity joint disorders. Despite varying levels of evidence, a growing number of studies have shown that manual joint mobilisations or manipulations are effective in certain disorders such as hip and knee osteoarthritis, patellofemoral pain syndrome, ankle inversion sprain, plantar fasciitis, metatarsalgia, and hallux limitus/rigidus (Brantingham et al 2009). Measurement of passive movement is indicated in order to assess joint restrictions and to help diagnose these disorders. Passive movement, either physiological or accessory, can be reported as range of motion, end-feel, or pain and is an indication of the integrity of joint structures (Cyriax 1982, Hengeveld and Banks 2005, Kaltenborn 2002). Passive physiological range of motion may be measured using vision or instruments such as goniometers or inclinometers.

An essential requirement of clinical measures is that they are valid and reliable so that they can be used to discriminate between individuals (Streiner and Norman 2008). Inter-rater reliability is a component of reproducibility along with agreement and refers to the relative measurement error, ie, the variation between patients as measured by different raters in relation to the total variance of the measurements (De Vet et al 2006, Streiner and Norman 2008). High inter-rater reliability for measurements of lower extremity joints is a prerequisite for valid and uniform clinical decisions about joint restrictions and related disorders (Bartko and Carpenter 1976).

Several reviews have systematically summarised and appraised the evidence with respect to the inter-rater reliability of passive movements of human joints. Seven systematic reviews have been published on passive spinal and pelvic movement including segmental intervertebral motion assessment (Haneline et al 2008, Hestbæk and Leboeuf-Yde 2000, May et al 2006, Seffinger et al 2004, Stochkendahl et al 2006, Van Trijffel et al 2005, Van der Wurff et al 2000). In general, inter-rater reliability was found to be poor and studies were of low methodological quality. A recent systematic review showed better inter-rater reliability for measurements of passive physiological range of motion in upper extremity joints using instruments compared to measurements using vision and compared to measurements of end-feel or accessory range of motion (Van de Pol et al 2010). To date, no systematic appraisal of studies on inter-rater reliability of measurement of passive movements in lower extremity joints has been conducted. Therefore, the research question for this systematic review was:

> What is the inter-rater reliability for measurements of passive physiological or accessory movements in lower extremity joints?

Method

Identification and selection of studies

MEDLINE, EMBASE, and CINAHL were searched for studies published up to 1 March 2010. Search terms included all lower extremity joints and all synonyms for *reliability* and *rater* (see Appendix 1 on the eAddenda for the detailed search strategy for MEDLINE). The titles and abstracts were screened for eligibility by two reviewers (EvT, RJvdP) independently. When necessary, full text articles were retrieved. Reference lists of all

retrieved papers were hand searched for relevant studies. A supplemental hand search of 13 journals relevant to the field of physiotherapy from 1 January 2005 to 1 March 2010 (see Appendix 2 on the eAddenda for journals) was performed by one reviewer (EvT). Finally, four experts in lower extremity musculoskeletal research were approached to ask if they could provide any additional published studies. Additionally retrieved papers were checked for eligibility by a second reviewer (RJvdP).

Studies were included if they met all inclusion criteria (Box 1). No restrictions were imposed on language or date of publication. Studies were excluded if they were abstracts and documents that were anecdotal, speculative, or editorial in nature. Studies were also excluded if they investigated: active movement or restriction in passive movement due to pain or ligament instability; people with neurological conditions in which abnormal muscle tone may interfere with joint movement; people after arthroplasty; animals or cadavers. Study selection was performed by two reviewers (EvT, RJvdP) independently. Disagreements on eligibility were first resolved by discussion between the two reviewers and decided by a third reviewer (CL) if disagreement persisted.

Box 1. Inclusion criteria.

Design
- Repeated measures between raters

Participants
- Symptomatic and asymptomatic adults

Measurement procedure
- Performed passive (ie, manual) physiological or accessory movements in any of the joints of the hip, knee, or ankle–foot–toes
- Reported range of motion or end-feel
- Used methods feasible in daily practice (considering instruments, costs, amount of training required)

Outcomes
- Estimates of inter-rater reliability

Assessment of characteristics of the studies

Description: We extracted data on participants (number, age, clinical characteristics), raters (number, profession, training), measurements (joints and movement direction, participant position, movement performed, method of measurement, outcomes reported), and inter-rater reliability (point estimates, estimates of precision). Two reviewers (EvT, RJvdP) extracted data independently and were not blind to journal, authors, or results. When disagreement between the two reviewers could not be resolved by discussion, a third reviewer (CL) made the final decision.

Quality: No validated instrument was available for assessing methodological quality of inter-rater reliability studies. Therefore, a list of criteria for quality was compiled derived from the QUADAS tool, the STARD statement, and criteria used for assessing studies on reliability of measuring passive spinal movement (Bossuyt et al 2003a, Bossuyt et al 2003b, Van Trijffel et al 2005, Whiting et al 2003). Criteria 1 to 4 assess external validity, Criteria 5 to 9 assess internal validity, and Criterion 10 assesses statistical methods (Box 2). Criteria were rated as 'yes', 'no', or 'unclear' where insufficient information was provided. External validity was considered sufficient if Criteria 1 to 4 were rated 'yes'. With respect to internal validity, Criteria 5, 6, and 7 were assumed to be decisive in determining risk of bias. A study was considered to have a low risk of bias if Criteria 5, 6, and 7 were all rated 'yes', a moderate risk if two of these criteria were rated 'yes', and a high risk if none or only one of these criteria were rated 'yes'. After training, two reviewers (EvT, RJvdP) independently assessed methodological quality of all included studies and were not blind to journal, authors, and results. If discrepancy between reviewers persisted, a decisive judgement was passed by a third reviewer (CL).

Box 2. Criteria for assessing methodological quality.

1. Was a representative sample of participants used?
2. Was a representative sample of raters used?
3. Is replication of the assessment procedure possible?
4. Was clinical information from participants available to raters and comparable to daily practice?
5. Were participants' characteristics under study stable during research?
6. Were raters' characteristics under study stable during research?
7. Were raters blinded to each other's results?
8. Can non-random loss to follow-up be ruled out?
9. Was an estimate of intra-rater reliability validly determined and was it above 0.80?
10. Were appropriate measures (Kappa, ICC) used for calculating reliability?

Data analysis

Data were analysed by examining ICC and Kappa (95% CI). If at least 75% of a study's ICC or Kappa values were above 0.75, the study was considered to have shown acceptable reliability (Burdock et al 1963, cited by Kramer and Feinstein 1981). Corresponding Kappa levels were used as assigned by Landis and Koch (1977) where < 0.00 = poor, 0.00–0.20 = slight, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial, and 0.81–1.00 = almost perfect reliability. In addition, reliability was analysed relating it to characteristics of the studies (participants' clinical characteristics, raters' profession and training, movement performed, method of measurement) and methodological quality. Reliability from studies not fulfilling Criteria 5 or 6 could have been underestimated, while reliability from studies not fulfilling Criterion 7 could have been overestimated. Negative scores on combinations of Criteria 5–7 could have led to bias in an unknown direction. Where one or more of these three criteria were rated 'unknown' because insufficient information was provided, no statement was made regarding the presence or direction of potential bias. Finally, clinical and methodological characteristics of included studies were examined for homogeneity in order to judge the possibility of statistically summarising results by calculating pooled estimates of reliability.

Results

Flow of studies through the review

Searching MEDLINE yielded 199 citations, of which 29 papers were retrieved in full text. After removing double citations, EMBASE (196 citations) provided another three potentially relevant studies. CINAHL (98 citations) then yielded no additional relevant articles. Hand searching of reference lists identified another 14 potentially eligible studies. Of these 46, 31 studies were excluded (see Appendix 3 on the eAddenda for excluded studies). Hand searching