Musculoskeletal Science and Practice 27 (2017) 78–84

Contents lists available at ScienceDirect

Musculoskeletal Science and Practice

journal homepage: www.journals.elsevier.com/aquaculture-and-fisheries/

Review article

Inter-examiner classification reliability of Mechanical Diagnosis and Therapy for extremity problems – Systematic review

Hiroshi Takasaki [a, *], Kousuke Okuyama [a], Richard Rosedale [b]

[a] Division of Physical Therapy, Saitama Prefectural University, Sannomiya 820, Koshigaya, Saitama, 343-8540, Japan
[b] Occupational Health and Safety, London Health Science Centre, London, N6A 5A5, Canada

ARTICLE INFO

Article history:
Received 23 September 2016
Received in revised form
26 December 2016
Accepted 29 December 2016

Keywords:
Classification
McKenzie
Musculoskeletal pain
Reliability

ABSTRACT

Background: Mechanical Diagnosis and Therapy (MDT) is used in the treatment of extremity problems. Classifying clinical problems is one method of providing effective treatment to a target population. Classification reliability is a key factor to determine the precise clinical problem and to direct an appropriate intervention.
Objectives: To explore inter-examiner reliability of the MDT classification for extremity problems in three reliability designs: 1) vignette reliability using surveys with patient vignettes, 2) concurrent reliability, where multiple assessors decide a classification by observing someone's assessment, 3) successive reliability, where multiple assessors independently assess the same patient at different times.
Design: Systematic review with data synthesis in a quantitative format.
Method: Agreement of MDT subgroups was examined using the Kappa value, with the operational definition of acceptable reliability set at $\geq$ 0.6. The level of evidence was determined considering the methodological quality of the studies.
Results/findings: Six studies were included and all studies met the criteria for high quality. Kappa values for the vignette reliability design (five studies) were $\geq$ 0.7. There was data from two cohorts in one study for the concurrent reliability design and the Kappa values ranged from 0.45 to 1.0. Kappa values for the successive reliability design (data from three cohorts in one study) were < 0.6.
Conclusion: The current review found strong evidence of acceptable inter-examiner reliability of MDT classification for extremity problems in the vignette reliability design, limited evidence of acceptable reliability in the concurrent reliability design and unacceptable reliability in the successive reliability design.

1. Introduction

The McKenzie System of Mechanical Diagnosis and Therapy (MDT) (McKenzie and May 2000; McKenzie and May 2003; McKenzie and May 2006) is a conservative approach to musculoskeletal disorders, where mechanical loading is used to identify a classification in order to guide the patient management strategy. Although the system is most widely practiced for the management of patients with spinal symptoms, there is an increasing body of research regarding the application of MDT to extremity problems (May and Rosedale, 2012; Kaneko et al., 2009; Littlewood and May 2007; Aina and May 2005; Rosedale et al., 2014; Aytona and Dudley, 2013; Kidd, 2013; Lynch and May 2013). In MDT, a classification is identified through a detailed history taking and physical evaluations including mechanical loading strategies. There are six primary classifications for extremity problems in MDT: 1) Derangement; 2) Articular Dysfunction; 3) Contractile Dysfunction; 4) Postural; 5) Spinal; and 6) OTHER, which can be further catabolized into 10 subgroups. The features of each classification are detailed elsewhere (Takasaki, 2016; Takasaki and May 2014; May and Ross, 2009). A provisional classification is made at the initial session and a definite classification is made through careful observations of symptomatic and mechanical changes at follow-up evaluations to undertake a stratified model of care.

Stratified models of care are used to provide a targeted management strategy for specific subgroups (Foster et al., 2013). These models have had an increasing prominence in the research as

* Corresponding author.
E-mail addresses: physical.therapy.takasaki@gmail.com (H. Takasaki), richard.rosedale@lhsc.on.ca (R. Rosedale).

http://dx.doi.org/10.1016/j.msksp.2016.12.016
2468-7812/

studies have demonstrated only modest treatment effects when patients are managed as a heterogeneous group (Foster et al., 2009, 2013). Well-designed randomized control trials (RCTs) are ultimately required to establish the best treatment for a subgroup in stratified models of care, but this requires considerable cost and effort. The 3-step Assessment-Diagnosis-Treatment-Outcome (A-D-T-O) model is a useful guide to design a high quality RCT. Its features are detailed elsewhere (Huijbregts, 2007; Spratt, 2002). The A-D-T-O model requires the establishment of the inter-examiner reliability of classification as the initial step, before undertaking subsequently higher study designs; cohort studies (second step) and a well-designed RCT (third step).

There has been a systematic review of the assessment process establishing inter-examiner reliability of MDT classification for patients with low back pain (May et al., 2006) but no systematic review has been conducted for extremity problems. In exploring the first step of the A-D-T-O model, some studies demonstrated strong inter-examiner reliability of MDT classification for extremity problems (May and Ross, 2009; Takasaki et al., 2015) but one study (Takasaki, 2016) did not. It is important to ascertain a comprehensive understanding of the inter-examiner reliability of MDT classification for extremity problems to facilitate future clinical trials using the MDT approach.

MDT classification of clinical presentations uses three methodologies; vignettes reliability, concurrent reliability and successive reliability. Vignette reliability design uses phantom patients while concurrent and successive reliability designs use actual patients. The vignette reliability design examines if assessors have reliability for identifying a classification from the reporting of the history and physical examination on a standardized assessment form. The concurrent reliability design examines if assessors are reliable in detecting and interpreting findings from the live assessment as well as reliable in identifying a classification. In the successive reliability design, multiple assessors assess the same patient at different times and decide on a classification. The successive reliability design examines if assessors have reliability for undertaking the assessment as well as detecting and interpreting findings from the live assessment and identifying a classification.

The purpose of this study was to explore inter-examiner reliability of the MDT classification for extremity problems in the vignette reliability design, the concurrent reliability design and the successive reliability design.

2. Materials and methods

2.1. Identification and selection of studies

The current systematic review was undertaken based on a guideline for systematic reviews on musculoskeletal disorders (Ghogomu et al., 2014). A systematic search (Appendix 1) was performed in MEDLINE, EMBASE, CINAHL, PubMed, Scopus and Cochrane library from inception to April 2016 without language limitations. Cross-referencing was undertaken through communications with three experts in this field (one expert was an external researcher and the two experts were authors; HT and RR). Relevant references in the ClinicalTrials.gov and the World Health Organization International Clinical Trial Registry Platform portal were hand-searched. Further, relevant references cited in studies included in full text inspection were also hand-searched.

Eligibility criteria were reliability studies that examined MDT for extremity problems. Two assessors (KO: novice to MDT, and HT: MDT credential therapist) independently searched for studies to be included and undertook screening by inspecting the title and abstract. Studies that either assessor retained through the screening were subsequently inspected with full text by the two assessors to examine the eligibility of the study. The article title and authors were not blinded to the assessors. However, any decision of either assessor regarding screening, quality appraisal and data extraction was blinded until consensus evaluation was undertaken.

2.2. Assessment of characteristics of studies

In this study, acceptable reliability was operationally defined as 0.6 $\geq$ Kappa (McHugh, 2012). The lower limit of 95% confidence interval (95% CI) was used for reliability value (Moran et al., 2016). A corresponding author was asked to provide additional information, if possible, for the current systematic review when the kappa and/or 95% CI values were insufficient. A point estimate was used for reliability value when the 95% CIs were not available (Moran et al., 2016). Two assessors (KO and HT) independently extracted data.

The primary measure for methodological quality was the Quality Appraisal of Diagnostic Reliability (QAREL) (Lucas et al., 2010, 2013). The Guidelines for Reporting Reliability and Agreement Studies (GRRAS) (Kottner et al., 2011) was used as a secondary measure as recommended (Lange et al., 2015). Clarity of the QAREL (Appendix 2) was completed as recommended (Lucas et al., 2013) through pilot assessments of studies included in a previous study (Ann Flavell et al., 2014). Two assessors (KO and HT) independently examined the methodological quality. Any disagreement was resolved by the third assessor (RR: MDT Diploma therapist and MDT instructor), when criteria with the disagreement and those full texts were presented to the third assessor in order to determine satisfaction or dissatisfaction of the criteria. Agreement of the methodological quality between examiners was examined with Cohen's Kappa and % agreement, where the Kappa value were: < 0.4 = poor agreement, 0.41–0.60 = moderate agreement, 0.61–0.80 = good agreement, 0.81–1.0 = very good agreement (Altman, 1991).

In this review, sensitivity analyses were undertaken by changing the cut-point for defining high quality study at $\geq$ 50%, $\geq$ 60%, and $\geq$ 70% of applicable QAREL scored as 'Yes' (Moran et al., 2016; Gorgos et al., 2014). The level of evidence was determined considering the number and quality of studies (strong evidence: consistent findings from multiple high quality studies [n $\geq$ 3], moderate evidence: consistent findings from $\geq$ one high quality and $\geq$ one low quality study, limited evidence: consistent findings in $\geq$ one low quality study or only one study available, conflicting evidence: inconsistent evidence in multiple studies irrespective of study quality, and no evidence: no study) (Moran et al., 2016). The level of evidence was determined in the vignette reliability design, the concurrent reliability design and the successive reliability design, respectively. When multiple reliability coefficients were available in a study, each reliability coefficient was examined in the current review to enhance understanding but did not affect evaluation of the level of evidence.

3. Results

3.1. Flow of studies through the review

Fig. 1 presents the flow of study selection. Three studies (May and Ross, 2009; Takasaki et al., 2015; Heidar Abady et al., 2014) were identified through the systematic search. Manual searching and cross-referencing identified a Master's thesis (Willis, 2015) and a study in press at the time of screening (Willis et al., 2016). The data in both were duplicate and therefore the Master's thesis (Willis, 2015) was excluded from the review. Further, manual searching and cross-referencing identified two studies (Takasaki, 2016; Kelly et al., 2008). There was no disagreement for examination of eligibility through inspecting full texts between the two