Inter-rater reliability of the McKenzie System of Mechanical Diagnosis and Therapy in the examination of the knee

Sean Willis^a , Richard Rosedale^a, Ravi Rastogi^a and Shawn M. Robbins^b

^aDepartment of Physiotherapy, London Health Sciences Centre, University Hospital, London, Canada; ^bSchool of Physical and Occupational Therapy, Centre for Interdisciplinary Research in Rehabilitation, Constance Lethbridge Rehabilitation Centre, McGill University, Montreal, Canada

ABSTRACT

Objective: The McKenzie System of Mechanical Diagnosis and Therapy (MDT) is a widely used method of classification and management of musculoskeletal problems. Although MDT has been investigated for its reliability and efficacy in the management of spinal pain, few studies have evaluated the system when applying it to musculoskeletal problems in the extremities, in particular the knee. The purpose of this study was to investigate the inter-rater reliability of MDT when classifying clinical vignettes describing patients with musculoskeletal knee pain.

Methods: This study was divided into two phases. First, 10 clinicians experienced in the use of MDT were recruited to write a total of 60 clinical vignettes based upon the initial assessment of their past patients with knee pain. Second, six different MDT raters were recruited to rate 53 selected vignettes and reliability was determined using Fleiss Kappa.

Results: There was 'substantial agreement' among six MDT raters classifying the clinical vignettes into one of four categories ($\kappa = 0.72$). There was no statistically significant difference between therapists with different levels of training.

Discussion: MDT demonstrated acceptable reliability among trained raters to classify clinical vignettes describing patients with musculoskeletal knee pain. To generalize the use of the system to more users, future research should continue to investigate the reliability of MDT using raters with lower levels of training and experience and assess reliability in real patients.

Level of Evidence: 5.

KEYWORDS

McKenzie System; knee; Mechanical Diagnosis and Therapy; reliability; vignette

Introduction

In the US, the prevalence of knee pain has increased by 65% over the last 20 years.[1] Over 4 billion dollars are spent annually on arthroscopic knee surgery alone [2] despite evidence suggesting arthroscopic surgery does not result in superior patient outcomes.[3–5] With an emphasis on fiscal responsibility in health care, clinicians should possess the skills and use methods of assessment that have the clinical utility to identify the most appropriate, cost-effective intervention from which the patient will benefit. Therefore, it is essential that an orthopedic evaluation of the knee is reliable and valid, is guided by clear diagnostic criteria, and provides the clinician with prognostic value.

Common practice in the clinical examination of the knee is the use of orthopedic special tests (OSTs). While an established body of evidence has highlighted the limited diagnostic validity of OSTs,[6–8] the primary purpose of establishing a diagnosis should be to offer insight into the management of a condition. In the example of meniscal pathology, the diagnosis itself does not direct a specific

management pathway as the link between the diagnosis and the most effective conservative treatment (i.e. Exercise) has not been established. Confounding this is the research around the use of medical imaging that has drawn attention to the presence of asymptomatic pathology and abnormalities.[9–12] As a result, it has been suggested that specifically defined criteria used by non-pathoanatomical classification systems may offer the clinician a reliable diagnosis that will direct an appropriate path of management and could be considered an alternative to the current orthopedic assessment model.[13]

One system that has not been tested for use in the assessment of musculoskeletal knee pain is the McKenzie System of Mechanical Diagnosis and Therapy (MDT). MDT uses a non-pathoanatomical approach to classify patients based on their response to repeated end range loading strategies. MDT involves a detailed history and an examination in which baseline symptoms, both with functional activity and at rest, are established and then re-evaluated following the patient performing repeated end range loading movements to the affected area. The assessment screens out potential serious pathology such as fractures, neurological

Table 1. Summary of MDT classifications and clinical presentation.[29]

MDT classification	Clinical presentation
Derangement	Varied in its clinical presentation; associated with mechanical obstruction of an affected joint. Key characteristic is the presence of a directional preference with loading strategies. Directional preference is defined as the rapid improvement of a patient's symptoms with positioning or movement in one specific direction
Dysfunction	Intermittent pain that is consistently reproduced at the end-range of a restricted movement, but will not persist once mechanical loading strategies have ceased. In the extremities, the Dysfunction syndrome can be disseminated further as Articular and Contractile Dysfunction
Articular Dysfunction	Distinguished from contractile through the loss of active and passive range of motion with pain being produced at the end of available range and absent during resisted testing
Contractile Dysfunction	Characterized by pain brought on by active and resisted movements, where passive range of motion is generally preserved
OTHER	Category is made up of 10 diagnostic subgroups which together complete the full MDT classification system where each subgroup has its own definition and diagnostic criteria (Refer to Appendix A)

and vascular issues to determine if a patient's symptoms are mechanical in nature. A key characteristic of the system that has shown potential as a prognostic indicator of musculoskeletal pain is directional preference.[14] Directional preference is defined as the rapid and lasting improvement of a patient's symptoms with positioning or movement in one specific direction.[13] Based on the patients' symptomatic response to the loading strategies, the clinician is able to formulate a provisional classification and provide directed treatment. MDT has previously demonstrated moderate to substantial inter-rater reliability in the musculoskeletal assessment of the spine ($\kappa = 0.6\text{--}0.84$),[15–17] shoulder ($\kappa = 0.9$),[18] and extremities ($\kappa = 0.7\text{--}0.83$) [19,20] among MDT trained clinicians.

MDT is a widely used method of classification for the management of musculoskeletal problems. Although MDT has been investigated for its reliability and efficacy in the management of spinal pain, few studies have evaluated the system when applying it to musculoskeletal problems in the extremities, in particular the knee. The objectives of this study were: (1) To develop clinical vignettes of patients presenting with musculoskeletal knee pain which are representative of four MDT clinical classifications (Table 1), (2) to test the inter-rater reliability of MDT-trained experienced clinicians when classifying clinical vignettes describing patients with musculoskeletal knee pain into one of four MDT classifications, and (3) to investigate the influence of the level of MDT education on the reliability of classifying clinical vignettes of patients presenting with musculoskeletal knee pain.

Methods

Clinical vignettes were utilized in the current study to examine MDT classification reliability. Vignettes are often

used in the evaluation of decision-making and clinical judgment of health professionals.[21] Although often criticized for not reflecting actual practice, well-designed vignette studies can be practical, offer flexibility, avoid ethical and observational issues, and be generalizable to real world settings.[21–23] Clinical vignettes are an inexpensive option to control multiple variables, collect information simultaneously from multiple sources, and isolate clinical decision-making.[22,24,25] Thus they provide an initial step in the investigation of the reliability of MDT for knee conditions.

Study design

To achieve our objectives, a two-phase study was conducted. The first phase consisted of the recruitment of 10 MDT clinicians to develop 60 clinical vignettes representative of the prevalence of musculoskeletal knee pain classified through MDT, of which 53 vignettes would be selected for phase two. To meet objectives two and three, the second phase required the recruitment of an additional six MDT raters to classify the clinical vignettes describing patients with musculoskeletal knee pain and measure the reliability and level of agreement among the MDT raters. Ethical approval for the study was obtained from the Health Science Research Ethics Board at Western University.

Sample size

Rotondi and Donner [26] proposed a method of calculating the sample size for studies measuring inter-rater agreement for multiple outcomes and raters. To arrive at an estimated sample size, κ was set at 0.8 (0.7 lower limit, 0.9 upper limit) based on levels of agreement with previous work evaluating the reliability of MDT in the extremities.[18–20] With an alpha of 0.05 for six raters (phase two) and using the prevalence rate of 0.4 (Derangement), 0.4 (Other), 0.1 (Contractile Dysfunction), 0.1 (Articular Dysfunction) of the four common MDT syndromes (Table 1) as outlined by May and Rosedale,[27] a value of 53 was determined for the number of clinical vignettes required for phase two. The sample size was estimated using a program developed for the R Project for Statistical Computing (Vienna, Austria).[28]

Phase 1

Participants

MDT education has two levels of clinical competence, Credentialed and Diploma. Credentialed clinicians have completed 108 h of post-graduate course work covering MDT principles and application of the system in the spine and extremities. Credentialed clinicians have also successfully passed a standardized written and practical examination. Having attained Credentialed status, clinicians can then go on to acquire Diploma status

which requires the completion of two components: one University semester theoretical component and 360 h of clinical practice mentorship clinical component. Once completed, the clinician must then pass an oral examination to be awarded Diploma status (<http://www.mckenzieinstitute.org/>).

For the first phase of the study, 10 clinicians experienced in the use of MDT in the extremity were recruited based on previous willingness to participate in research. These clinicians were asked to generate 60 clinical vignettes, six vignettes per clinician, classified into one of four classifications. The sample size of 10 was chosen to minimize the burden of creating the vignettes on the consenting clinician. To be included, clinicians were Credentialed or Diplomats with the McKenzie Institute with more than three years of experience applying MDT to musculoskeletal disorders of the extremity and were registered on the publicly available list of MDT practitioners practicing in the United States or Canada. Clinicians were excluded if they were unable to understand written and spoken English, unable to provide informed consent, or unable to follow the instructions for generating the clinical vignettes. Correspondence was conducted and informed consent was obtained from each clinician through electronic mail. In total, 20 clinicians were approached for recruitment to which 10 consented for participation in this study.

Vignette development

May and Rosedale [27] identified four MDT classifications that were most prominent in patients presenting with musculoskeletal knee pain: Derangement, Contractile Dysfunction, Articular Dysfunction, and OTHER. The Derangement syndrome is the most common of the three mechanical syndromes and is varied in its clinical presentation; however, the key characteristic is the presence of a directional preference with loading strategies. [29] The clinical presentation for Dysfunction syndrome is intermittent pain that is consistently reproduced at the end-range of a restricted movement but will not persist once mechanical loading strategies have ceased. [29] The Dysfunction syndrome can be further disseminated into Contractile and Articular Dysfunction. Contractile Dysfunction is characterized by pain brought on by active and resisted movements, where passive range of motion is generally painfree and preserved. [29] Articular Dysfunction is distinguished from contractile through the loss of active and passive range of motion with pain being produced at the end of available range and absent during resisted testing. [29] The OTHER category is made up of 10 diagnostic subgroups which together complete the full MDT classification system where each subgroup has its own definition and diagnostic criteria (Appendix A). Appendix A has been modified from the original publication by May and Rosedale [27] to reflect revisions

made to the definitions and diagnostic criteria by the International Education Committee of the McKenzie Institute International.

For clinical vignettes to be clinically relevant, they should be based on or relating to a case study or clinical experience and follow a similar structure and natural flow of a typical assessment. [24,25] Also, it is preferential to generate more vignettes than will be used so they can be reviewed by an expert panel to determine appropriateness and to select the best sample for testing. [25,30] For this study, clinicians were asked to generate vignettes based on their past patient assessment files. Clinicians were instructed that each clinical vignette be de-identified to only include gender, age range (e.g. 35–40 years old), and a category of occupation. The written vignettes would be characteristic of one of four MDT classifications: Derangement, Articular Dysfunction, Contractile Dysfunction, and OTHER. A summary of these classifications that are currently used in the MDT education manuals is provided in Table 1. Clinicians were asked to write the clinical vignettes on a blank McKenzie extremity assessment form. The blank McKenzie extremity assessment form used was revised from the standard form to exclude entry areas for patient names and other identifying information (Appendix B). To correspond with established prevalence, each clinician was asked to submit six clinical vignettes consisting of two Derangements, one Articular Dysfunction, one Contractile Dysfunction, and two OTHER.

Once received, all vignettes were reviewed by two authors. The first author is an MDT Credentialed physiotherapist and has 17 years of clinical practice working with patients with musculoskeletal knee pain. The second author is a MDT Diploma physiotherapist with 23 years of clinical experience and is Senior Faculty of the McKenzie Institute. The review of the cases was done to ensure that each vignette was complete, possessed characteristics of one of the four MDT classifications requested, and that a level of ambiguity existed that would be present in the situation of a clinical patient presentation. Any discrepancies were identified, flagged, and discussed with the clinician who developed the vignette to ensure and verify accuracy of the case. In some situations, clinicians were unable to submit one or more of the number of vignettes matching the requested classification(s) because of the lack of a past patient assessment(s) that represented that classification(s). In those instances, clinicians chose another past patient assessment with one of the other requested classification(s) to submit to fulfill their quota of six clinical vignettes. In total, 60 vignettes were received: 24 Derangement, 8 Articular Dysfunction, 8 Contractile Dysfunction, and 20 OTHER. After a review of all vignettes was completed, 53 cases were selected that were representative of the established prevalence for use in phase two of the study. Of the 53 vignettes, 22 were Derangement, 7 Articular

Dysfunction, 7 Contractile Dysfunction, and 17 OTHER. An example vignette for each category of classification can be found in Appendix C–F.

Phase 2

Participants

For phase two, six different raters were recruited based on previous willingness to participate in research from the publicly available list of MDT practitioners registered with McKenzie Institute International who practice in Canada or the United States. They were required to classify 53 selected clinical vignettes generated in phase one of the study. The sample size of six raters was chosen to provide equal division of groups by level of MDT training and within group comparison. To be included, the rater had to have Credentialed or Diploma MDT status and have applied MDT to the extremities for more than three years. Raters were excluded if they participated in the creation of the clinical vignettes, did not wish to participate, were unable to understand written and spoken English, unable to provide informed consent or were unable to follow the instructions for rating the clinical vignettes. Correspondence was conducted and informed consent was obtained from each rater through electronic mail. In total, all six raters recruited consented to participate in the study.

Data collection and procedures

Demographic information (e.g. gender, age) were collected along with other relevant characteristics like clinical practice setting, years of practice, length of time since obtaining Credentialed/Diplomat status, discipline (e.g. Physiotherapist vs. Doctor of Chiropractic), estimated proportion of extremity patients treated and proportion of knee patients treated with MDT. For each vignette, the raters were instructed to review the vignette and based on the history and clinical presentation, assign the vignette a classification of Derangement, Articular Dysfunction, Contractile Dysfunction, or OTHER. Each vignette was randomly assigned a number from 1 to 53 to facilitate tracking of responses and data collection. All raters were blinded to the provisional MDT classification originally assigned to the vignette by its creator in phase one.

Analysis

Descriptive statistics for the demographic and clinical information for the raters were determined. Interrater reliability, our primary objective, was determined through the calculation of Fleiss kappa statistic along with 95% confidence interval (CI) and standard error (SE) across all six raters for all categories.[31,32] Data were analyzed for Fleiss kappa using a program written in

Matlab version 7.14 (MathWorks, Natick, MA, USA).[33] Kappa values were interpreted using definitions outlined by Landis and Koch [34]: 0.01–0.20 slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and 0.81–1.00 almost perfect agreement.

Although overall kappa for three or more raters may lead to a better representation of reliability, overall kappa may mask extreme cases of agreement or disagreement for paired raters.[35] A solution is to report both overall and paired kappa data to provide the most informative summary.[35] Paired comparisons of the agreement of vignette classification among the six raters were thus analyzed and reported as percentage agreement and kappa statistic with standard error. Additionally, individual raters and their agreement with the vignette provisional classification assigned in phase one were analyzed and reported as percentage agreement and kappa statistic with standard error. Raters were grouped based on their level of education.

To examine if the level of education influenced the reliability, differences in Fleiss kappa values between Credentialed and Diploma therapists were compared. A bootstrap method with a 1000 samples was utilized and Fleiss kappa coefficients were calculated separately for the Credentialed and Diploma raters for each of these samples.[36] The differences between the Fleiss kappa coefficients were determined. The mean of these differences was determined along with the 95% CI represented by the 25 and 975 values. If the 95% CI included zero, then no significant difference existed between Credentialed and Diploma raters.

Results

The six raters recruited to rate the clinical vignettes were all physiotherapists and comprised of three Credentialed and three Diploma therapists. Four raters practiced fee for service and two worked in multiple settings. Four raters were male and two were female. Demographic information obtained from each of the raters is displayed in Table 2.

The overall kappa value among the six raters demonstrated substantial agreement with $\kappa = 0.72$ (SE = 0.02, 95% CI = 0.71–0.73). The highest level of reliability was for the Derangement category with $\kappa = 0.83$; the lowest level of reliability was for the OTHER category with $\kappa = 0.64$. Articular and Contractile Dysfunction had a κ of 0.67 and 0.69, respectively. There was 100% agreement in classification among all six raters in 31 of the 53 (58.5%) clinical vignettes.

Paired comparison of agreement in vignette classifications across the six raters are displayed in Table 3. The top right half shows percentage agreement and the bottom left half shows kappa scores (standard error) for all possible pairings of raters. The highest percentage

Table 2. Demographic information of participating phase 2 raters (n = 6).

Variables	Mean (SD)	Range
Age, years	51 (13.4)	35–67
Years of practice	25 (13.7)	10–44
Proportion of extremity patients treated in practice, percentage	37 (16.0)	20–65
Proportion of knee patients of peripheral joints in practice, percentage	79 (37.2)	5–100

Note: SD – standard deviation.

Table 3. Percentage agreement and kappa (standard error) for paired comparisons among the six raters.

Rater	Percentage agreement					
	1	2	3	4	5	6
1	–	83%	89%	77%	92%	77%
2	0.74 (0.09)	–	75%	75%	81%	72%
3	0.83 (0.09)	0.64 (0.09)	–	81%	91%	79%
4	0.67 (0.08)	0.65 (0.08)	0.73 (0.08)	–	83%	75%
5	0.89 (0.09)	0.72 (0.09)	0.86 (0.09)	0.76 (0.08)	–	81%
6	0.66 (0.09)	0.58 (0.09)	0.72 (0.08)	0.65 (0.08)	0.72 (0.09)	–
Kappa (SE)						

of agreement (92%) and kappa (0.89) were between rater 1 and 5. The lowest percentage of agreement (72%) and kappa (0.58) were between rater 2 and 6. Reliability between raters showed moderate to substantial agreement.

Individual rater responses were compared to the provisional classifications of the vignettes and reliability calculated with results displayed in Table 4. Raters are grouped by level of MDT training. The highest percentage of agreement with the provisional classification was 91% for rater 5 and the lowest agreement was 81% for rater 6. Kappa scores ranged from 0.73 for rater 6 and 0.86 for rater 5. All raters showed substantial reliability ($\kappa \geq 0.73$) with the provisional classification.

Grouped by level MDT training, Credentialed ($\kappa = 0.71$; SE = 0.05, 95% CI = 0.61–0.8) and Diploma ($\kappa = 0.74$; SE = 0.05, 95% CI = 0.64–0.84) raters demonstrated substantial agreement. The mean difference value between kappa values for Credentialed and Diploma therapists was -0.03 (95% CI = -0.15 – 0.11). Since the confidence interval includes 0, there is no significant difference between rater groups based on level of education.

Discussion

To our knowledge, this is the first study to investigate the reliability of MDT specifically for musculoskeletal knee pain alone. The primary findings of this study suggest that the inter-rater reliability of MDT Credentialed and Diploma clinicians demonstrate ‘substantial agreement’ when classifying clinical vignettes based on patients presenting with musculoskeletal knee pain ($\kappa = 0.72$). There was no significant difference between Credentialed and Diploma therapists. Thus, it appears that clinicians with Credentialed or Diploma status can use MDT to classify clinical vignettes describing patients with knee pain.

The results are consistent with other studies evaluating the use of MDT in the extremities,[19,20] shoulder [18], and spine. [15–17] The reliability among raters for this study ($\kappa = 0.72$) is comparable to previous studies in the extremities ($\kappa = 0.7$ – 0.83).[19,20] Earlier studies for musculoskeletal lumbar and cervical spinal pain found moderate to substantial reliability ($\kappa = 0.6$ – 0.84) for paired raters which is similar to the current results ($\kappa = 0.58$ – 0.89).[15–17] Against the provisional classification for the current study, individual raters demonstrated substantial reliability ($\kappa = 0.73$ – 0.86) while percentage of agreement ranged from 81 to 91%. Similarly, Abady et al. [18] reported substantial reliability ($\kappa = 0.89$) with 95% overall agreement across raters against the provisional classification for musculoskeletal shoulder disorders. In contrast, Werneke et al. [37] found only fair to moderate reliability ($\kappa = 0.37$ to 0.44) despite high observed agreement (86–91%) for lumbar classification among therapists with pre-Credentialed level of training. It has been suggested this paradox results from the sensitivity of the kappa statistic to skewed prevalence of ratings and that the interpretation of the kappa statistic alone may result in conclusions that may be misleading.[38,39] The high prevalence of Derangement (81–86%) and the relative low prevalence for the remaining classifications (0–4.6%) in this study likely accounted for the low kappa values despite the high observed agreement.[7,35,37] Regardless, MDT seems to have similar reliability when classifying clinical vignettes or real patients with musculoskeletal spinal, knee, and shoulder pain.

For the current study, the highest level of reliability was for the Derangement category ($\kappa = 0.83$) while the level of reliability for the three remaining categories were lower ($\kappa = 0.62$ – 0.69). The difference between these levels of reliability may be explained by the general presentation of each category. The Derangement syndrome

Table 4. Percentage agreement and kappa (standard error) of individual raters vs. the provisional classification grouped by MDT education.

Rater	Statistic	Diploma				Credentialed	
		1	2	3	4	5	6
Provisional classification	% Agreement	88	88	85	83	91	81
	Kappa (SE)	0.83 (0.09)	0.83 (0.09)	0.78 (0.09)	0.76 (0.08)	0.86 (0.09)	0.73 (0.08)

Note: SE – standard error.

is readily identifiable by a lasting reduction or elimination of patients' symptoms through repeated movement in a particular direction.[29] The relative lower level of agreement of the remaining categories may be partially attributable to the absence of a unique identifiable characteristic, such as directional preference, that is present with Derangement syndrome. This is particularly true for the OTHER category as multiple subgroups are included which makes determining a classification based exclusively on an initial assessment more challenging.

The level of MDT training did not affect the overall level of reliability of the system when evaluating musculoskeletal knee pain. Grouped by education, both Diploma ($\kappa = 0.78\text{--}0.83$) and Credentialed ($\kappa = 0.73\text{--}0.86$) raters had similar agreement with the provisional classification. Although Diploma clinicians undergo further education, each rater in our study had a great deal of experience using the extremity assessment form and in treating patients with musculoskeletal pain in the extremities. We did not include raters without Credentialed or Diploma competence and as a result the effect of lower levels of MDT training was not evaluated.

It has been suggested that an assessment to simply identify structures as the cause of pain does not elicit enough information to understand the problem or to justify a course of management.[40] Yet it has been observed that clinicians performing an orthopedic assessment have become over reliant on the results of OSTs and medical imaging despite the limited value the results may provide toward the management of some conditions.[41] Lynch and May [42] documented the case study of directional preference of the knee using MDT. Although presenting with a positive McMurray's test and pain with swimming, following prescribed exercise matching the directional preference, the patient reported 95% improvement in function and symptoms and a negative McMurray's test. It was concluded by the authors that the use of McMurray's test was not diagnostic and that the management strategy for the patient was determined not by the structural diagnosis, but rather by the classification established through the symptomatic and mechanical response to repeated loading strategies applied to the knee.

Further to this, Rosedale et al. [13] published a randomized control trial using MDT to determine a directional preference for patients with end stage knee OA. Patients matched with exercises consistent with a directional preference demonstrated significant decreases in pain and increases in self-reported function with small to large effect sizes after two weeks ($d = 0.98\text{--}1.44$) and three months ($d = 0.42\text{--}0.80$) compared to the control group who received no exercise intervention and patients without a directional preference. Although the results could not be conclusively attributed to the MDT classification, they concluded that the response to directional preference matched exercises should be explored

further. These and the findings of Lynch and May [42] suggest that classification systems like MDT may be an effective alternative as it relies on a patient's symptomatic response to mechanical loading strategies to guide treatment rather than the patho-anatomy.[13,27,42]

The primary limitation of this study is the raters have achieved a high level of MDT education and experience. This limits the generalizability of the results to those individuals with similar training and experience. The vignettes with the provisional classification of OTHER were not further disseminated into subgroups for the raters to identify. As such, the reliability of raters classifying clinical vignettes describing patients into those subgroups cannot be determined from this study. Additionally, the reviewers of each of the vignettes were not blinded to the creator of the vignette or the provisional classification assigned to each vignette. Another potential limitation is the use of clinical vignettes as an alternative to real patients. However, clinical vignettes were used as they offer the convenience of collecting information from multiple sources simultaneously while allowing for the flexibility of variable manipulation to ensure a heterogeneous sample.

Conclusion

This is the first study investigating the inter-rater reliability of the McKenzie System of MDT in the examination of musculoskeletal knee pain. The McKenzie System of MDT demonstrated substantial agreement, indicating acceptable inter-rater reliability for trained raters when classifying clinical vignettes describing patients presenting with musculoskeletal knee pain. The results of this study offer preliminary support for the use of MDT in the assessment of musculoskeletal knee pain and support for future studies. To generalize the use of the system to more users, future research should continue to investigate the reliability of MDT using raters with lower levels of training and replicate the findings in real patients.

Acknowledgements

We would like to thank members of the MDT community for their participation and contributions to this study. We would also like to thank Dr. Trevor Birmingham for his influence and direction for this study.

Disclosure statement

The authors declare no conflicts of interest for this study.

Notes on contributor

Sean Willis, BScPT, MSc, Cred MDT, is a physiotherapist at London Health Sciences Centre. His research interests are MDT, Orthopedics, and Musculoskeletal Care.

Richard Rosedale, PT, DipMDT, is a physiotherapist at London Health Sciences Centre; also an instructor and reference coordinator with the McKenzie Institute International. He provides educational courses on Mechanical Diagnosis and Therapy for which he receives a teaching fee. His research interests are McKenzie System, Orthopedics, and Musculoskeletal Care.

Ravi Rastogi, BScPT, MSc, Cred MDT, is a physiotherapist at London Health Sciences Centre. His research interests are MDT, Musculoskeletal Care, and Orthopedics.

Shawn Robbins, BScPT, PhD, is an assistant professor, McGill University, School of Physical and Occupational Therapy. Robbins' research utilizes biomechanical and clinical measures to assess orthopedic health conditions in both clinical and laboratory settings. He also examines interventions used to treat these conditions. Additionally, he examines the mechanics of ice hockey skating and shooting.

ORCID

Sean Willis <http://orcid.org/0000-0001-9595-2810>

Ravi Rastogi <http://orcid.org/0000-0001-8577-6956>

Shawn M. Robbins <http://orcid.org/0000-0002-8108-0561>

References

- [1] Nguyen UDT, Zhang Y, Yanyan Z, et al. Increasing prevalence of knee pain and symptomatic knee osteoarthritis: survey and cohort data. *Ann. Intern. Med.* 2011;155:725–732.
- [2] Gage BE, McIlvain NM, Collins CL, et al. Epidemiology of 6.6 million knee injuries presenting to united states emergency departments from 1999 through 2008. *Acad. Emerg. Med.* 2012;19:378–385.
- [3] Kirkley A, Birmingham TB, Litchfield RB, et al. A randomized trial of arthroscopic surgery for osteoarthritis of the knee. *N. Engl. J. Med.* 2008;359:1097–1107.
- [4] Sihvonen R, Paavola M, Malmivaara A, et al. Arthroscopic partial meniscectomy versus sham surgery for a degenerative meniscal tear. *New Engl. J. Med.* 2013;369:2515–2524.
- [5] Thorlund JB, Juhl CB, Roos EM, et al. Arthroscopic surgery for degenerative knee: systematic review and meta-analysis of benefits and harms. *BMJ.* 2015;350:1–9.
- [6] Hegedus EJ, Cook C, Hasselblad V, et al. Physical examination tests for assessing a torn meniscus in the knee: a systematic review with meta-analysis. *J. Orthop. Sports Phys. Ther.* 2007;37:541–550.
- [7] Lange T, Freiberg A, Dröge P, et al. The reliability of physical examination tests for the diagnosis of anterior cruciate ligament rupture – a systematic review. *Man. Ther.* 2015;20:402–411.
- [8] Leblanc MC, Kowalczyk M, Andruszkiewicz N, et al. Diagnostic accuracy of physical examination for anterior knee instability: a systematic review. *Knee Surg. Sports Traumatol. Arthrosc.* 2015;23:2805–2813.
- [9] Beattie KB, Boulous P, Pui M, et al. Abnormalities identified in the knees of asymptomatic volunteers using peripheral magnetic resonance imaging. *Osteoarthr. Cartil.* 2005;13:181–186.
- [10] Boks SS, Vroegindeweij D, Koes BW, et al. Magnetic resonance imaging abnormalities in symptomatic and contralateral knees: prevalence and associations with traumatic history in general practice. *Am. J. Sports Med.* 2006;34:1984–1991.
- [11] Kaplan LD, Schurhoff MR, Selesnick H, et al. Magnetic resonance imaging of the knee in asymptomatic professional basketball players. *Arthroscopy.* 2005;21:557–561.
- [12] LaPrade RF, Quinter BM, Veenstra MA, et al. The prevalence of abnormal magnetic resonance imaging findings in asymptomatic knees: with correlation of magnetic resonance imaging to arthroscopic findings in symptomatic knees. *Am. J. Sports Med.* 1994;22:739–745.
- [13] Rosedale R, Rastogi R, May S, et al. Efficacy of exercise intervention as determined by the McKenzie System of mechanical diagnosis and therapy for knee osteoarthritis: a randomized controlled trial. *J. Orthop. Sports Phys. Ther.* 2014;44:173–181.
- [14] May S, Aina A. Centralization and directional preference: a systematic review. *Man. Ther.* 2012;17:497–506.
- [15] Clare HA, Adams R, Maher CG. Reliability of McKenzie classification of patients with cervical or lumbar pain. *J. Manipulative Physiol. Ther.* 2005;28:122–127.
- [16] Kilpikoski S, Airaksinen O, Kankaanpää M, et al. Interexaminer reliability of low back pain assessment using the McKenzie method. *Spine.* 2002;27:E207–E214.
- [17] Razmjou H, Kramer E, Yamada R. Intertester reliability of the McKenzie evaluation in assessing patients with mechanical low-back pain. *J. Orthop. Sports Phys. Ther.* 2000;30:368–389.
- [18] Abady AH, Rosedale R, Overend TJ, et al. Inter-examiner reliability of diplomats in the mechanical diagnosis and therapy system in assessing patients with shoulder pain. *J. Man. Manipulative Ther.* 2014;22:199–205.
- [19] May S, Ross J. The McKenzie classification system in the extremities: a reliability study using McKenzie assessment forms and experienced clinicians. *J. Manipulative Physiol Ther.* 2009;32:556–563.
- [20] Kelly E, May S, Ross J. The reliability of the McKenzie classification system using extremity McKenzie assessment forms. *Int. J. Mech. Diag. Ther.* 2008;3:3–6.
- [21] Evans SC, Roberts MC, Keeley JW, et al. Vignette methodologies for studying clinicians' decision-making: validity, utility, and application in ICD-11 field studies. *Int. J. Clin. Health Psychol.* 2015;15:160–170.
- [22] Peabody JW, Luck J, Glassman P, et al. Measuring the quality of physician practice by using clinical vignettes: a prospective validation study. *Ann. Intern. Med.* 2004;141:771–780.
- [23] Rutten GMJ, Harting J, Rutten STJ, et al. Measuring physiotherapists' guideline adherence by means of clinical vignettes: a validation study. *J. Eval. Clin. Pract.* 2006;12:491–500.
- [24] Gould D. Using vignettes to collect data for nursing research studies: how valid are the findings? *J. Clin. Nurs.* 1996;5:207–212.
- [25] Veloski J, Tai S, Evans AS, et al. Clinical vignette-based surveys: a tool for assessing physician practice variation. *Am. J. Med. Qual.* 2005;20:151–157.
- [26] Rotondi MA, Donner A. A confidence interval approach to sample size estimation for interobserver agreement studies with multiple raters and outcomes. *J. Clin. Epidemiol.* 2012;65:778–784.
- [27] May SJ, Rosedale R. A survey of the McKenzie classification system in the extremities: prevalence of mechanical syndromes and preferred loading strategies. *Phys. Ther.* 2012;92:1175–1186.
- [28] Rotondi MA. Sample size estimation functions for studies of interobserver agreement [cited 2013 Mar 12]. Available December 16, 2013 from: <http://cran.r-project.org/web/packages/kappaSize/index.html>.
- [29] McKenzie R, May S. The human extremities: mechanical diagnosis and therapy. New Zealand: Spinal Publications; 2000.

- [30] Atzmüller C, Steiner PM. Experimental vignette studies in survey research. *Methodology*. 2010;6:128–138.
- [31] Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol. Bull.* 1971;76:378–382.
- [32] Fleiss JL, Nee JC, Landis JR. Large sample variance of kappa in the case of different sets of raters. *Psychol. Bull.* 1979;86:974–977.
- [33] Cardillo G. Fleiss'es kappa: compute the Fleiss'es kappa for multiple raters [cited 2009 Dec 23]. Available March 10, 2015 from: <http://www.mathworks.com/matlabcentral/fileexchange/15426>.
- [34] Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
- [35] O'Leary S, Lund M, Ytre-Hauge TJ, et al. Pitfalls in the use of kappa when interpreting agreement between multiple raters in reliability studies. *Physiotherapy*. 2014;100:27–35.
- [36] McKenzie DP, Mackinnon AJ, Péladeau N, et al. Comparing correlated kappas by resampling: Is one level of agreement significantly different from another? *J. Psychiatr. Res.* 1996;30:483–492.
- [37] Werneke MW, Deutscher D, Hart DL, et al. McKenzie lumbar classification: inter-rater agreement by physical therapists with different levels of formal McKenzie postgraduate training. *Spine*. 2014;39:E182–E190.
- [38] Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J. Clin. Epidemiol.* 1993;46:423–429.
- [39] de Vet HC, Mokkink LB, Terwee CB, et al. Clinicians are right not to like Cohen's κ . *BMJ*. 2013;346:1–7.
- [40] Jones MA, Rivett DA. Introduction to clinical reasoning. In: Jones MA, Rivett DA, editor. *Clinical Reasoning for Manual Therapists*. Edinburgh: Butterworth Heinemann; 2004. p. 16–17.
- [41] Cook C. The lost art of the clinical examination: an overemphasis on clinical special tests. *J. Man. Manipulative Ther.* 2010;18:3–4.
- [42] Lynch G, May S. Directional preference at the knee: a case report using mechanical diagnosis and therapy. *J. Man. Manipulative Ther.* 2013;21:60–66.