Clinician's ability to identify neck and low back interventions: an inter-rater chance-corrected agreement pilot study

Mark W Werneke¹, Dennis L Hart², Daniel Deutscher³, Paul W Stratford⁴

¹CentraState Medical Center, Freehold, NJ, USA, ²Focus On Therapeutic Outcomes, Inc., White Stone, VA, USA, ³Physical Therapy Department, Maccabi Healthcare Services – HMO, Tel-Aviv, Israel, ⁴School of Rehabilitation Science and Department of Clinical Epidemiology and Biostatistics, McMaster University, Ontario, Canada

Objective: To estimate inter-rater agreement of physical therapists trained in MDT approach and participating in practice-based evidence (PBE) research to identify 72 physical therapy interventions in video demonstrations on a single model and clinical vignettes. PBE is a well designed observational study and demonstrating clinician observational consistency is an important step in conducting PBE research design.

Methods: Two physical therapists volunteered to participate in pilot reliability testing and seven other physical therapists trained in McKenzie Mechanical Diagnosis and Therapy (MDT) methods volunteered for the inter-rater chance-corrected agreement study. All therapists identified interventions presented within 52 videos and 5 written clinical vignettes describing 20 more intervention techniques. Therapists independently identified all interventions. We assessed inter-rater chance-corrected agreement of therapists' ability to identify intervention techniques using Kappa coefficients with associated 95% confidence intervals and indices for bias and prevalence.

Results: Of the 147 kappa coefficients estimated, 7% were ≤ 0.6 , 10% were > 0.6 and ≤ 0.8 , and 83% were > 0.8 . Agreement was lowest for identifying cognitive behavioral techniques (median kappa=0.79). The minimum and maximum prevalence and bias indices were 0.33 and 0.85 and 0 and 0.33, respectively suggesting kappa coefficient estimates were strong. Generalized kappa coefficients ranged from 0.73 to 1.00.

Discussion: Results provide evidence that substantial to almost perfect inter-rater agreement could be expected when trained therapists identify physical therapy interventions used for patients with spinal impairments from staged videos and vignettes. This may be helpful to reassure clinicians of the quality of the reporting of intervention(s) performed when conducting multivariable analyses in future pragmatic PBE studies. Additional studies are needed to test whether these results can be validated using larger groups of therapists, trained and not trained in MDT methods, as well as examining different methods to examine inter-rater agreement for identifying diverse interventions commonly used for managing patients during routine practice.

Keywords: Inter-rater reliability, Intervention, Practice based evidence, Spinal impairments

Introduction

Low back pain is a common and costly condition.¹ The majority of adults seeking care for their low back pain troubles are managed in primary care and of those a substantial number of patients are referred to physical therapy.² Over 25 to 45% of patients referred to outpatient physical therapy clinics are treated for low back pain.^{3,4} Clinical practice guidelines for managing patients with low back pain are starting to recognize the importance of classifying patients based on clinical

signs and symptoms in order to identify optimal rehabilitation standards of care, reduce practice variation, and improve the effectiveness and efficiency of treatment.^{5,6} There are many different classification approaches⁷⁻¹¹ routinely used by physical therapists to manage patients with lumbar impairment yet there is little agreement among clinicians and researchers on which classification system works best for which patients.¹²

McKenzie or Mechanical Diagnosis and Therapy (MDT) is a popular classification approach frequently utilized by physical therapists both in the US and internationally.^{9,13,14} For instance, 63% of physical

Correspondence to: M W Werneke, CentraState Medical Center, 901 West Main Street, Freehold, NJ 07728, USA. Email: mwernerke@centrastate.com

therapists surveyed in the US reported McKenzie as an evaluation preference for patients with chronic low back pain³ and nearly 36–47% of therapist in Europe preferred the McKenzie approach as a treatment method.^{15,16} The MDT system has been described in detail elsewhere.⁹ Briefly, MDT method includes an evaluation and a matched intervention approach guided by a classification scheme based on the patient's subjective and mechanical responses to end range repeated trunk movement tests. Patients are classified into three treatment subgroups i.e. derangement, dysfunction, and posture.⁹ The key MDT intervention principle is the prescription of individual exercises based on directional preference or centralization observed during the patient's assessment.⁹ The majority of studies investigating centralization have reported an important association between this patient response criterion and outcomes when patients are managed by clinicians adequately trained in the MDT approach.^{17–19} However, despite the established prognostic value of centralization, the efficacy of the McKenzie system has not yet been established^{20,21} and most trials have found no clinically important differences in outcomes comparing MDT to other intervention types such as manual orthopedic or cognitive behavioral therapies.^{22–24}

One reason which may partially explain the lack of difference in outcomes between patients managed by MDT trained physical therapists versus clinicians trained in other treatment approaches may be the lack of clear reliable descriptions of all possible interventions and treatment processes used during studies investigating the MDT approach. For example, Machado *et al.* reported similar pain, disability, and functional outcomes between patients managed with a first-line care approach (i.e. advice, reassurance, and time contingent acetaminophen) by primary care physicians compared to interventions used by credentialed MDT physical therapists.²² The physical therapists participating in this study were instructed to exclusively adhere to the treatment principles described in McKenzie textbooks.²² The core components of treatment principles described in McKenzie textbooks are specific exercise, manual therapy, education, and postural training.⁹ Yet the authors provided no data to identify if the therapists were choosing the same interventions in a reliable manner or which intervention or combinations of the treatment components were used for patients classified into each of the MDT classification subgroups. In addition, the authors did not consider or discuss what type of interventions were applied for patients whose symptoms did not centralize or could not be classified into one of the three MDT classification subgroups.²² Although the patient population and outcomes were clearly described by Machado *et al.*,

treatment processes associated with better or worse outcomes remained under-specified.²²

Rigorous and reliable definition of treatment processes is an essential component of scientific research to improve the dissemination and implementation of evidence-based interventions during routine clinical practice. Comparative effectiveness research (CER) analyzes following Practice-based evidence (PBE) criteria is a research design developed to enhance clinical understanding of possible associations between interventions or combinations of interventions administered and patient outcomes while controlling for important patient characteristics such as classification category. PBE has been recommended as a complementary and alternative design to RCTs^{25,26} and has been recently cited by Freburger and Carey as a new frontier to accelerate the advancement of the physical therapy profession by examining intervention comparative effectiveness and efficiency during routine clinical practice.²⁵ PBE are well designed observational studies that rely on evidence from daily clinical practice without establishing a strict experimental environment.^{27–29} PBE maybe a promising research design to identify which interventions are associated with best outcomes for which patients when patients with low back pain are managed by clinicians interested in MDT methods.

Practice based evidence studies require comprehensive and complex databases including detailed patient characteristics, reliable treatment care processes, and valid outcome measures.^{28–30} Seven specific steps are required in order to facilitate PBE.²⁶ Horn and Gassaway identified the need for testing inter-rated reliability regarding intervention care process data as one of the important steps for developing the foundation necessary to conduct PBE studies.²⁶ In order to implement extensive data collection PBE process, Horn reported that internal reliability testing of treatment care processes is required for quality assurance.²⁶ Inter-rater reliability data facilitates the ultimate goal of PBE to identify which intervention(s) used by physical therapists during everyday routine care, are associated with better outcomes.^{25,26,31–34}

In order to progress with PBE design studies investigating the McKenzie system, the importance of recording interventions applied by MDT therapists in a reliable and valid manner needs to be demonstrated. Explaining variance in outcomes using interventions as predictors in CER models can not begin until we can confirm that clinicians can identify and document interventions in a reliable way. Inter-rater reliability of all measures must be assessed and determined to be adequate. Inter-rater agreement of physical therapists ability to identify treatments applied during routine practice has not been determined for patients with lumbar or cervical impairment. We are not aware of

any prior research which specifically examined the inter-rater agreement of identifying a wide variety of treatment procedures commonly used by physical therapists for this population during routine clinical practice. Because demonstrating clinician observational consistency is an important step in PBE research design, the purpose of this study was to begin estimates for inter-rater agreement between clinicians trained in MDT participating in PBE research to identify physical therapy interventions in video demonstrations and clinical vignettes. Our study is a precursor to conducting PBE research to begin the process of identifying the most important intervention(s), i.e. active ingredient, associated with best patient outcomes when patients with low back pain are managed by therapists trained MDT.

Methods

Design

We conducted a prospective, inter-rater chance-corrected agreement study investigating the ability of seven physical therapists to agree on treatment techniques used for patients with cervical or lumbar impairments. All seven clinicians were collecting clinical outcome data using Focus On Therapeutic Outcomes, Inc. (FOTO) (Knoxville, TN, USA), an international medical rehabilitation data management company.^{35,36} All clinicians signed an informed consent prior to participating in this study. The FOTO Institutional Review Board for the Protection of Human Subjects approved the project.

Procedures

Agreement testing

One adult female (age 23 years; Body Mass Index 22) signed an informed consent and volunteered to be videotaped as the subject for the study. She was pain free and reported no history of neck or back pain complaints. A series of 52 treatments were demonstrated on the volunteer by one physical therapist (MW) experienced with all treatment techniques. The treatments were videotaped in a hospital-based outpatient rehabilitation clinic simulating actual equipment used and treatments rendered for a patient with neck or back pain.

Clinicians

Seven physical therapists participating in a PBE research project consented to participate as raters in this study. The ultimate goal for our PBE research initiative was to investigate comparative effectiveness between treatments delivered by MDT versus non-MDT trained therapists. Therefore, all therapists met the following inclusion criteria: (1) working in an outpatient rehabilitation setting; (2) treating patients with musculoskeletal problems including neck and back pain; and (3) completing post-graduate training (credentialing) in MDT methods.^{9,37} Therapists (mean age 37 years, minimum 30, maximum 44, 100% males)

worked in diverse practice settings: two therapists worked in hospital-based orthopedic outpatient clinics, 4 therapists worked in private practices, and one therapist worked in a military orthopedic outpatient clinic setting. Therapists worked in separate clinics in different states except for two clinicians who were co-owners of one private practice. All therapists earned baccalaureate degrees in physical therapy; three therapists earned master's degree in physical therapy; one therapist earned a doctorate degree in science; and 1 therapist earned a doctorate of physical therapy. The average number of years of clinical experience was 12 years (minimum 8, maximum 17 years).

Agreement testing procedures

For this study we investigated 72 treatment interventions outlined in the Appendix (treatment definitions available upon request). The taxonomy included interventions specifically practiced by clinicians trained in MDT⁹ as well as other intervention approaches reported and commonly used by physical therapists during routine care for this population.^{4,15,16,38} PBE designs encourage and allow both standardized and non-standardized intervention protocols to be included, so data reflect real-world practice settings.²⁶ Seventy-two treatment interventions were examined for inter-rater agreement via video images and written vignettes. Fifty-two intervention techniques were videotaped and 20 treatment techniques were described in one of five clinical vignettes (case studies available upon request). We selected video and vignette demonstration techniques because of practical and logistical issues. For example, as compared to having each therapist view actual patient interventions, videos and vignettes are less expensive and do not expose patients to undo risks. In addition, we believe several techniques could best be demonstrated using video, like manipulation and mobilization. Videos and vignettes also permitted flexibility in product format, for example the clinic where the video was developed did not have a traction table, so the traction technique was described in a vignette. The proportion of the videos ($N=52$) specifically demonstrating a cervical or lumbar intervention was 25% ($N=13$) and 31% ($N=16$), respectively. The largest proportion of videos 44% ($N=23$) were independent of body part. In other words, videotaping interventions such as aerobic exercise, lifting and sitting posture correction can be demonstrated on video regardless of whether the patient has a cervical or lumbar impairment. The proportion of case studies describing a cervical patient was 20% (1 of 5 vignettes). Therapists received a copy of the operational definitions describing the interventions as a training guide prior to the start of the reliability testing. Each therapist independently watched the videotape, read the written vignettes, and recorded which intervention technique was identified on

the videotape and which intervention techniques were being described in each vignette. Therapists were blinded to each other's answers and mailed completed answer sheets to co-investigator DH for agreement analyses.

Videotape procedure

To develop the videotape, a cameraman recorded each intervention and edited the videotape by separating and ordering each intervention using a computer-generated table of random numbers to produce a new videotape, which was used for therapist viewing. In order to minimize biasing therapists with intervention recognition, the video did not include audio except where required. The following videotaping procedures required audio: (1) the cameraman reported the intervention number being viewed to assist in therapists' data documentation on the videotape answer sheet; (2) the examiner asked the subject while performing therapeutic exercises for directional preference (i.e. extension, flexion, and lateral) 'what effect does this movement have on your symptoms?' The subject was allowed to respond by choosing 1 of 3 responses, i.e. the movement decreases my leg pain, the movement abolishes my leg pain, or the movement has no effect on my pain; and (3) the examiner verbally trained the subject for all educational and cognitive behavioral treatment techniques simulating patient education received during actual clinical practice.

After each intervention technique was viewed, the therapist was instructed to stop the videotape and write one of the seven possible intervention groups defined in Appendix (i.e. exercise, manual, educational, functional, modality, cognitive behavioral or administrative by referral) and name of the intervention technique on a blank answer sheet numbered 1–52. For example, the 15th videotaped treatment technique demonstrated the examiner performing lumbar extension mobilization. After viewing the videotape displaying this technique, the therapist documented the treatment group 'manual' and the treatment technique 'lumbar extension mobilization' on the 15th entry on the blank videotape answer sheet.

Vignette procedure

After watching the videotape and recording the descriptions of the techniques, therapists read the five vignettes. The vignettes detailed four out of five patients (80%) between the ages of 34 and 51 with lumbar syndromes and one 60-year-old patient with cervical radiculopathy. Each vignette included descriptions of a patient's history, symptoms, relevant physical examination findings, and treatment processes used to manage the patient based on his or her initial evaluation. One or more scenarios describing a specific intervention technique followed each vignette.

The therapist identified which intervention technique was being described. Consistent with the video intervention technique documentation, therapists wrote the intervention type and technique on blank numbered vignette answer sheet.

Pilot testing

To test the agreement process before conducting the formal test, a pilot test was undertaken to ensure the relevance of the content and clarity of the interventions being videotaped and described in the vignettes. Two physical therapists, who did not participate in the subsequent agreement study, volunteered for the pilot project: one therapist was MDT credentialed; and one therapist completed an American Academy of Orthopedic Manual Physical Therapy (AAOMPT) residency program without post-graduate MDT training. Therapists were instructed to document the treatments being performed after independently viewing the videotape and reading the vignettes as outlined above. Based on the pilot test, only three minor text edits for the vignette answer sheet were made to clarify data documentation. The edits did not change the content of the case studies for subsequent testing.

Data analyses

To estimate the inter-rater agreement of clinicians' ability to identify specific interventions (Appendix), categorical data from all raters were organized into seven sets of data for analyses. The first set of analyses was for the major groupings of seven types of interventions, i.e. therapeutic exercise, manual techniques, educational techniques, functional training activities, modality procedures, cognitive behavioral techniques, and an administrative technique. The second set of analyses was for the specific interventions within each major grouping, including: (1) therapeutic exercise, 8 types; (2) manual, 7 types; (3) education, 7 types; (4) functional training, 3 types; (5) modality, 7 types; (6) cognitive behavioral, 6 types, and (7) administrative by referral, 1 type. An unweighted kappa statistic³⁹ was calculated [i.e. (observed agreement minus chance agreement)/(1 minus chance agreement)] to estimate the proportion of agreement greater than that expected by chance⁴⁰ for each pair of raters for each grouping of treatments. For the pilot data, there were seven pairs of observations, one for each major grouping, which produced seven kappa estimates. Because the pilot data were only used to test the process, only kappa estimates per grouping were generated. For the study, there were 21 possible pairs of ratings from seven raters and seven treatment groupings for a total of 147 kappa estimates. In addition, generalized kappa coefficients were estimated for each major grouping.⁴¹ For each kappa, a 95% confidence interval was

calculated to estimate a range of plausible values for the population kappa.

As recommended,⁴⁰ to assist in the interpretation of the magnitude of each kappa coefficient, we calculated a prevalence index and a bias index for the study data. The prevalence index was calculated as the absolute value of the difference between the true positives (number of times a given technique was correctly identified by both raters from a pair) and true negatives (number of times a given technique is correctly not reported when not demonstrated) divided by the number of paired ratings because kappa coefficients are influenced by the prevalence of the attribute assessed. A prevalence effect exists when the proportion of agreement for a specific treatment differs from that of the other treatments. If the prevalence index is high, chance agreement will be high, and kappa will be reduced.⁴⁰ The effect of prevalence on kappa is greater for large compared to small values of kappa.⁴² A bias index represents the extent to which the raters disagree on the proportion of positive or present treatments and is calculated as the absolute value of the difference between the false positives and false negatives divided by the number of paired ratings. When disagreement is close to symmetrical, i.e. proportions are close, bias will be low. When disagreements are asymmetrical, i.e. proportions are not close, bias will be high. When there is a large bias index, kappa is higher than when bias is low or absent.⁴⁰ Finally, Landis and Koch proposed the following qualifying terminology for interpreting kappa coefficients: ≤ 0 =poor agreement, 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.⁴³

Results

Pilot study kappa values ranged from 0.68 to 1.00. Numeric results and communications with the two participating pilot therapists suggested the videotaped images and vignettes were understandable and ready for testing. For the inter rater agreement for the seven study therapists, kappas ranged from 0.68 to 1.00. Of the 147 kappa coefficients estimated, the smallest kappa was 0.17, 7% of kappas were ≤ 0.60 (all 10 were related to assessing cognitive techniques), 10% were >0.60 and ≤ 0.80 (of the 15 kappas, five were related to assessing cognitive techniques, five were related to assessing exercise techniques, and five were related to assessing educational techniques), 83% were >0.80 (of the 122 kappas, six were related to assessing cognitive techniques, 16 were related to assessing educational techniques, 16 were related to assessing exercise techniques, 21 were related to assessing functional techniques, 21 were related to assessing manual techniques, 21

were related to assessing modalities, and 21 were related to assessing major groupings of techniques). Percentage of perfect agreement (kappa=1) was 58%, and the median kappa was 1.00. Minimum and maximum prevalence indices were 0.33 and 0.85 with a median of 0.50, which can be interpreted as neither high nor low. Minimum and maximum bias indices were 0 and 0.33 with a median of 0.13, which can be interpreted as bias being either absent or low. Listing of kappa coefficients, 95% CIs, prevalence and bias for therapeutic exercises as an example are displayed in Table 1 (a full listing of all kappas and associated statistics can be obtained from DLH). Examining kappa estimates and associated statistics by group of intervention showed agreement was lowest for identifying cognitive behavioral techniques (median kappa=0.79, median prevalence=0.50, median bias=0.17) compared to one of the highest median kappas for identifying modality procedures (median kappa=1.00, median prevalence=0.50, median bias=0.13). Summary agreement statistics for each of the seven groupings are displayed in Table 2. Generalized kappa coefficients ranged from 0.73 to 1.00 (Table 3).

Discussion

Results of this study provide support that substantial to almost perfect levels of agreement for identifying diverse intervention techniques for cervical or lumbar impairments displayed in videos or vignettes can be achieved by physical therapists trained in MDT methods participating in the study. A frequently reported shortcoming of research examining the efficacy or effectiveness of various interventions is a lack of clear and reliable descriptions of key interventions being investigated.^{31,44}

Table 1 Inter-therapist agreement for identifying the therapeutic exercises in video demonstrations and vignettes describing a case

Raters		Kappa	95% CI	Prevalence	Bias
Rater 1	Rater 2	1.00	1.00, 1.00	0.65	0.00
Rater 1	Rater 3	1.00	1.00, 1.00	0.65	0.00
Rater 1	Rater 4	1.00	1.00, 1.00	0.68	0.00
Rater 1	Rater 5	1.00	1.00, 1.00	0.65	0.00
Rater 1	Rater 6	0.80	0.58, 1.00	0.47	0.18
Rater 1	Rater 7	1.00	1.00, 1.00	0.65	0.00
Rater 2	Rater 3	1.00	1.00, 1.00	0.65	0.00
Rater 2	Rater 4	1.00	1.00, 1.00	0.65	0.00
Rater 2	Rater 5	1.00	1.00, 1.00	0.65	0.00
Rater 2	Rater 6	0.80	0.58, 1.00	0.47	0.18
Rater 2	Rater 7	1.00	1.00, 1.00	0.65	0.00
Rater 3	Rater 4	1.00	1.00, 1.00	0.65	0.00
Rater 3	Rater 5	1.00	1.00, 1.00	0.65	0.00
Rater 3	Rater 6	0.80	0.58, 1.00	0.47	0.18
Rater 3	Rater 7	1.00	1.00, 1.00	0.65	0.00
Rater 4	Rater 5	1.00	1.00, 1.00	0.65	0.00
Rater 4	Rater 6	0.80	0.58, 1.00	0.47	0.18
Rater 4	Rater 7	1.00	1.00, 1.00	0.65	0.00
Rater 5	Rater 6	0.85	0.66, 1.00	0.50	0.19
Rater 5	Rater 7	1.00	1.00, 1.00	0.65	0.00
Rater 6	Rater 7	0.80	0.58, 1.00	0.47	0.18

Note: CI=confidence interval. Prevalence=prevalence index. Bias=bias index.

Although it would appear clinically intuitive that physical therapists should be able to reliably identify and describe all common interventions practiced during everyday clinical practice, we are unaware of any prior research suggesting that this intuitive belief is supported by statistical evidence. Whyte and Hart described identifying and defining rehabilitation interventions as complex 'It's more than a black box; it's a Russian doll' and wanted researchers to objectively define interventions to facilitate effectiveness research.⁴⁴ Therefore, determining agreement of intervention identification is a crucial step progressing to CER analyses.²⁶ We believe our results provide sufficient evidence that the therapists participating in this study can progress with CER analyses to begin the process for identifying which rehabilitation interventions are associated with better outcomes for which patients with spinal impairments when managed by therapists trained in MDT.

To improve interpretation of kappa coefficient estimates, we estimated prevalence and bias.⁴⁰ For example, a high kappa estimate might be caused by a high bias, i.e. therapists disagreed on the proportion of positive or negative cases. In the current data, there were no high bias estimates. High prevalence indices tend to lower kappa estimates compared to low prevalence indices, the effect of which is greater for large values of kappa, as in these data particularly for therapeutic exercise (median kappa 1.00 and median prevalence 0.65) and educational techniques (median kappa 0.90 and median prevalence 0.77). However, in these data (Table 2), relatively high prevalence indices did not appear to reduce the strong kappa estimates. Therefore, in the current data, bias indices were low or absent, and only two (exercise and education) kappa estimates were associated with relatively high prevalence indices, which implies the high kappa estimates were not affected by prevalence or bias.

We tested a diverse list of interventions commonly used by physical therapists during routine

practice.^{4,13,16,38} including treatment principles described in the McKenzie textbooks.^{9,37} Although our study's aim was specific, i.e. to examine the inter-rater reliability of clinicians trained in MDT, we tested a diverse taxonomy of interventions such as strengthening and stabilization exercises, passive modalities, and cognitive behavioral therapies. There are no data to indicate that clinicians trained in MDT only apply pure McKenzie treatment principles for all patients. For example in a recent study 28% of patients could not be classified into one of the three major McKenzie syndromes.¹² It is unclear how patients who are not classified or whose symptoms do not centralize are managed by MDT trained therapists during everyday clinical care across different practice settings. Prior research investigating outcomes for patients managed by clinicians credentialed in MDT suggest that additional interventions such as strengthening exercises and cognitive behavioral therapy may be prescribed,^{12,45-47} therefore it appears reasonable to include and test a comprehensive taxonomy of interventions for this study. It is also important to note that we did not include many other published manual intervention types or variations⁴⁸⁻⁵⁰ in our study. For instance, we did not test a lumbar manipulation technique used in the development and validation of a clinical prediction rule for spinal manipulation.⁴⁹ Yet it would not be feasible to test for all possible clinical variations in manual interventions applied routinely in everyday clinical practice in one study. Therefore additional inter-rater reliability studies would be needed to examine reliability of recognizing variations of manual techniques currently used in clinical practice in order to progress with PBE studies investigating the association between outcomes and interventions used by therapists with different training and practice patterns.

Although we investigated the raters' ability to recognize specific intervention techniques in a reliable manner, the competency levels for the clinicians to

Table 2 Agreement estimates within each major intervention groupings

Group*	Kappa			Prevalence			Bias		
	Median	Min.	Max.	Median	Min.	Max.	Median	Min.	Max.
Major	0.97	0.91	1.00	0.48	0.44	0.51	0.13	0.10	0.17
Exercise	1.00	0.80	1.00	0.65	0.47	0.68	0.00	0.00	0.19
Manual	0.94	0.82	1.00	0.59	0.50	0.66	0.09	0.09	0.18
Education	0.90	0.79	1.00	0.77	0.69	0.85	0.15	0.15	0.23
Functional	1.00	1.00	1.00	0.33	0.33	0.33	0.00	0.00	0.00
Modalities	1.00	1.00	1.00	0.50	0.50	0.50	0.13	0.13	0.13
Cognitive	0.79	0.17	1.00	0.50	0.33	0.67	0.17	0.00	0.33

Note: *Each group consisted of data from video tape and vignette presentations. The presentation data for each group were as follows: Major: 52 video, 20 vignette. Exercise: 17 video, 2 vignette. Manual: all 22 video. Education: 3 video, 10 vignette. Functional: 2 video, 1 vignette. Modalities: 6 video, 2 vignette. Cognitive: 2 video, 4 vignette. Prevalence=prevalence index. Bias=bias index. Min.=minimum. Max.=maximum. Major=major groupings. Exercise=therapeutic exercises. Manual>manual techniques. Education=educational techniques. Functional=functional training techniques. Modalities=modality procedures. Cognitive=cognitive behavioral techniques.

deliver each intervention in the same manner was beyond the scope of this study. Future studies examining therapists' ability to consistently deliver each intervention technique in the same manner during routine care are warranted. For our study examiners had completed post-graduate training and achieved credentialing status in MDT assessment and treatment methods.^{9,37} The MDT credentialing examination process accesses the therapist's basic skill set to classify and treat biomechanical spinal impairments with manual techniques for managing patients with cervical or lumbar impairments. Therefore, we believe the clinicians were competent at least at a basic level for performing the manual intervention techniques examined. In addition MDT training encompasses several of the competencies required to implement a cognitive behavioral approach (CBA) in primary care as outlined by van der Windt *et al.*⁵¹ For example, patient empowerment is a key tenet for both CBA and MDT approaches, which leaves responsibility for rehabilitation with the patient and considers patients as active participants in his or her own care versus passive receptors of care. Similarity, in educational messages between MDT and CBA approaches may have resulted in the high level of agreement observed between raters for the cognitive behavioral techniques investigated. It is also possible that the mode of presentation, i.e. case studies, relying heavily on non-video media, enhanced agreement between therapists for identifying cognitive behavioral interventions. Future studies exploring similarities between MDT and CBA educational approaches using a variety of different methods to best examine reliability are needed.

MDT training and the credentialing examination process however does not test the therapist's competence for therapeutic exercise prescription except for exercise prescribed based on directional preference nor functional training techniques, modality procedures, and administrative techniques examined in our study. Therefore, the influence of MDT training may only partially explain the high inter-rater chance-corrected agreement coefficients obtained in our study. Substantial agreement between the therapists may have been the result of providing all raters with standard

operational definitions for judging each intervention. Precise and standardized operational definitions have been recommended as an important method for improving inter-rater reliability regardless of training and experience levels.⁵²⁻⁵⁴ For example, we observed substantial agreement during the pilot test between two experienced therapists with different post-graduate educational training, i.e., AAOMPT versus MDT.

Limitations

The similarity of training in MDT across examiners and the scripted sample of interventions presented in this study may diminish the generalizability of our results beyond clinicians credentialed in MDT techniques, and we did not study whether advanced MDT training affected agreement statistics. In addition, the small number of therapists participating in the reliability testing needs to be considered when interpreting our inter-rater agreement results. The specific aim of our study, however, was to investigate inter-therapist reliability among MDT trained clinicians for identifying diverse treatment techniques following PBE research criteria.^{26,31,33,34,44} Researchers need to perform additional studies to assess how generalizable these reliability estimates are for larger samples of therapists with diverse post-graduate training. Because we are collaborating with therapists in Israel in our PBE studies, we are initiating a larger inter-rater reliability study to assess recognition of treatments outlined in Appendix among 400 physical therapists in Israel, including 168 therapists at different levels of MDT training from which 20 therapists have recently achieved a credential MDT level.

We did not test all possible intervention techniques, variations or dosage such as intensity and frequency of exercise prescription, which affects generalizability of our results. The techniques tested in our study are commonly reported by physical therapists managing patients with musculoskeletal back pains,^{4,16,38,55} and the reliability estimates were substantial suggesting our intervention descriptions were sufficient at least for the therapists participating in the study.

Only one volunteer and one treating therapist were videotaped in the study. Using one volunteer and therapist is not representative of all populations or clinicians, which may have influenced our reliability estimates. However, the volunteer was solicited because of a normal BMI, which allowed clear demonstration of the therapist's hand placement for mobilization and manipulation techniques during the video. We believe the other intervention techniques tested such as education, cognitive behavioral, or exercise are independent of specific patient characteristics, so we did not use additional volunteers during the videotaping.

Table 3 Generalized kappa agreement estimates within each major intervention groupings

Group	Kappa (95% CI)
Major groupings	0.96 (0.93, 0.98)
Therapeutic exercise	0.82 (0.85, 0.93)
Manual therapy techniques	0.95 (0.90, 0.99)
Educational techniques	0.94 (0.88, 0.99)
Functional training techniques	1.00 (0.83, 1.00)
Modality procedures	1.00 (0.94, 1.00)
Cognitive behavioral techniques	0.73 (0.65, 0.81)

Note: CI=confidence interval.

In addition, it is possible the selective use of sound during the video may have inflated the reliability estimates by labeling more clearly educational and cognitive behavior techniques and the directional preference exercises. We are not aware of a better method to test these particular techniques while avoiding undo patient burden and all potential rater bias. The majority of educational and cognitive behavioral techniques were presented in the case studies to mitigate the potential bias for using sound during video demonstration. It is also possible that the presentation of the vignettes may have helped inflate reliability by clearly breaking down the information to one intervention. For example in 1 case study the intervention describes a patient being treated on a lumbar traction table. It is intuitively logical that scripting the vignette would be very similar to the operational definition describing lumbar traction intervention. Regardless of whether the rater viewed a video of patient on traction table, observed a patient on the traction table in the clinic, or identified the same intervention in a case study, considerable overlap between scripting and defining the same intervention would be necessary regardless of the mode of presentation. Therefore, similarities between scripting and defining interventions are inherent within any study examining the reliability for identifying interventions applied during clinical practice.

Finally, videotaping and paper-based case vignettes used in our study merit further discussion because they may affect our results. Videotaping has been previously used by many researchers for examining reliability of visual analysis of gait^{56,57} and scapular positioning⁵⁸ as well as determining classification sub groupings for patients with cervical or lumbar impairments.^{53,59–63} The advantages of videotapes and vignettes include improved feasibility for testing different therapists from different geographical regions, standardizing testing procedures to mitigate possible subject changes between a test and retest, and a reduction of potential stress on patients that alters clinical presentation.⁶⁰ Despite these advantages we recognize the limitations of using video and paper case vignettes to estimate inter-therapist reliability for identifying interventions. The biggest concern with these methods is the purification of the intervention being viewed, which may not reflect all of the dynamics of a real patient encounter occurring during clinical practice, thereby making it easier to identify the intervention provided on the video tape or case study compared to actual clinical situations, all of which may inflate the kappa coefficients. Weighing the advantages and limitations for video/vignette methods and considering the specific aim of our paper, we concluded that possible

disadvantages cited for using videotapes/paper vignettes would not be enough to overshadow the benefits for the current study. Future research investigating head to head comparisons between different reliability testing methods such as videotaping/vignettes versus observing care provided by real therapists and real patients are recommended.

Conclusions

The results of the study provided evidence that a substantial level of agreement for identifying intervention techniques can be achieved by physical therapists trained in MDT methods. Inter-rater agreement of clinician's ability to identify rehabilitation interventions is important to examine associations between treatment care processes and patient outcomes when conducting practice-based evidence research designed studies. Further studies are needed to investigate generalizability of our results in larger samples of therapists with diverse post-graduate clinical background and interventions delivered by both MDT and non-MDT trained clinicians.

Acknowledgements

We thank physical therapists Patricia Guttormsen PT, FAAOMPT and Adrian Reyes PT, Cert MDT (CentraState Medical Center, Freehold, NJ, USA) for volunteering as raters for the pilot testing. We would also like to thank the following physical therapists for volunteering as raters for the reliability assessment consisting of Guillermo Cutrone, PT, DSc, OCS, FAAOMPT, Cert MDT, (Research and Clinical Coordinator at St. Vincent's Medical Center Northeast, Fisher, IN, USA), Maj Troy McGill, MPT, Dip MDT (Director of Physical Therapy Flight, USAF, Joint Base Elmendorf, Fort Richardson, AK, USA), Jon Weinberg, PT, Dip MDT, (Director, Team Care Physical Therapy, Oxford, NC, USA), William Oswald DPT, Cert MDT (NYU Hospital for Joint Diseases, New York, NY, USA), David Grigsby MPT, Cert MDT and Jason Ward MPT, Cert MDT (Director, MidSouth Orthopaedic Rehabilitation, Germantown, TN, USA), and Dave Oliver PT, Dip. MDT (Director, Physical Therapy in Motion, Saline, MI, USA).

References

- 1 Druss BG, Marcus SC, Olfson M, Pincus HA. The most expensive medical conditions in America. *Health Aff (Millwood)* 2002;21:105–11.
- 2 Freburger JK, Carey TS, Holmes GM. Physician referrals to physical therapists for the treatment of spine disorders. *Spine J* 2005;5:530–41.
- 3 Battie MC, Cherkin DC, Dunn R, Ciol MA, Wheeler KJ. Managing low back pain: attitudes and treatment preferences of physical therapists. *Phys Ther* 1994;74:219–26.
- 4 Jette AM, Smith K, Haley SM, Davis KD. Physical therapy episodes of care for patients with low back pain. *Phys Ther* 1994;74:101–10; discussion 110–5.
- 5 Guide to physical therapist practice. 2nd ed. American Physical Therapy Association. *Phys Ther* 2001;81:9–746.

- 6 Koes BW, van Tulder MW, Thomas S. Diagnosis and treatment of low back pain. *BMJ* 2006;332:1430-4.
- 7 Delitto A, Erhard RE, Bowling RW. A treatment-based classification approach to low back syndrome: identifying and staging patients for conservative treatment. *Phys Ther* 1995;75:470-85; discussion 485-8.
- 8 Laslett M. Evidence-based diagnosis and treatment of the painful sacroiliac joint 9. *J Man Manip Ther* 2008;16:142-52.
- 9 McKenzie R, May S. The Lumbar spine: mechanical diagnosis and therapy. 2nd ed. Waikanae: Spinal Publication, Ltd; 2003.
- 10 Petersen T, Laslett M, Thorsen H, Manniche C, Ekdahl C, Jacobsen S. Diagnostic classification of non-specific low back pain. A new system integrating patho-anatomical and clinical categories. *Physiother Theory Pract* 2003;19:213-37.
- 11 Spratt KF, Lehmann TR, Weinstein JN, Sayre HA. A new approach to the low-back physical examination. *Behavioral assessment of mechanical signs. Spine* 1990;15:96-102.
- 12 Werneke MW, Hart DL, Oliver D, McGill T, Grigsby D, Ward J, et al. Prevalence of classification methods for patients with lumbar impairments using the McKenzie syndromes, pain pattern, manipulation, and stabilization clinical prediction rules. *J Man Manip Ther* 2010;18:197-210.
- 13 Hefford C. McKenzie classification of mechanical spinal pain: profile of syndromes and directions of preference. *Man Ther* 2008;13:75-81.
- 14 May S, Donelson R. Evidence-informed management of chronic low back pain with the McKenzie method. *Spine J* 2008;8:134-41.
- 15 Byrne K, Doody C, Hurley DA. Exercise therapy for low back pain: a small-scale exploratory survey of current physiotherapy practice in the Republic of Ireland acute hospital setting. *Man Ther* 2006;11:272-8.
- 16 Foster NE, Thompson KA, Baxter GD, Allen JM. Management of nonspecific low back pain by physiotherapists in Britain and Ireland. A descriptive questionnaire of current clinical practice. *Spine* 1999;24:1332-42.
- 17 Long A, Donelson R, Fung T. Does it matter which exercise? A randomized control trial of exercise for low back pain. *Spine* 2004;29:2593-602.
- 18 Werneke M, Hart DL. Centralization phenomenon as a prognostic factor for chronic low back pain and disability. *Spine* 2001;26:758-64.
- 19 Werneke MW, Hart DL. Categorizing patients with occupational low back pain by use of the Quebec Task Force Classification system versus pain pattern classification procedures: discriminant and predictive validity. *Phys Ther* 2004;84:243-54.
- 20 Clare HA, Adams R, Maher CG. A systematic review of efficacy of McKenzie therapy for spinal pain. *Aust J Physiother* 2004;50:209-16.
- 21 Machado LA, de Souza MS, Ferreira PH, Ferreira ML. The McKenzie method for low back pain: a systematic review of the literature with a meta-analysis approach. *Spine* 2006;31:E254-62.
- 22 Machado LA, Maher CG, Herbert RD, Clare H, McAuley JH. The effectiveness of the McKenzie method in addition to first-line care for acute low back pain: a randomized controlled trial. *BMC Med* 2010;8:10.
- 23 Moffett JK, Jackson DA, Gardiner ED, Torgerson DJ, Coulton S, Eaton S, Mooney, et al. Randomized trial of two physiotherapy interventions for primary care neck and back pain patients: 'McKenzie' vs brief physiotherapy pain management. *Rheumatology* 2006;45:1514-21.
- 24 Paatelma M, Kilpikoski S, Simonen R, Heinonen A, Alen M, Videman T. Orthopaedic manual therapy, McKenzie method or advice only for low back pain in working adults: a randomized controlled trial with one year follow-up. *J Rehabil Med* 2008;40:858-63.
- 25 Freburger JK, Carey TS. Comparative effectiveness research: opportunities and challenges for physical therapy. *Phys Ther* 2010;90:327-32.
- 26 Horn SD, Gassaway J. Practice-based evidence study design for comparative effectiveness research. *Med Care* 2007;45:S50-7.
- 27 Horn SD. Clinical practice improvement methodology: implementation and evaluation. New York: Faulkner & Gray; 1997.
- 28 Horn SD, Gassaway J. Practice based evidence: incorporating clinical heterogeneity and patient-reported outcomes for comparative effectiveness research. *Med Care* 2010;48:S17-22.
- 29 Horn SD, Gassaway J, Pentz L, James R. Practice-based evidence for clinical practice improvement: an alternative study design for evidence-based medicine. *Stud Health Technol Inform* 2010;151:446-60.
- 30 Gassaway J, Horn SD, DeJong G, Smout RJ, Clark C, James R. Applying the clinical practice improvement approach to stroke rehabilitation: methods used and baseline results. *Arch Phys Med Rehabil* 2005;86:S16-33.
- 31 Deutscher D, Horn S, Dickstein R, Hart DL, Smout RJ, Gutvirth M, et al. Associations between treatment processes, patient characteristics and outcomes in outpatient physical therapy practice. *Arch Phys Med Rehabil* 2009;90:1349-63.
- 32 Deutscher D, Hart DL, Dickstein R, Horn SD, Gutvirth M. Implementing an integrated electronic outcomes and electronic health record process to create a foundation for clinical practice improvement. *Phys Ther* 2008;88:270-85.
- 33 Horn SD. Performance measures and clinical outcomes. *JAMA* 2006;296:2731-2.
- 34 Horn SD, DeJong G, Ryser DK, Veazie PJ, Teraoka J. Another look at observational studies in rehabilitation research: going beyond the holy grail of the randomized controlled trial. *Arch Phys Med Rehabil* 2005;86:S8-15.
- 35 Swinkels IC, van den Ende CH, de Bakker D, Van der Wees PJ, Hart DL, Deutscher D, et al. Clinical databases in physical therapy. *Physiother Theory Pract* 2007;23:153-67.
- 36 Swinkels ICS, Hart DL, Deutscher D, van den Bosch WJH, Dekker J, de Bakker DH, et al. Comparing patient characteristics and treatment processes in patients receiving physical therapy in the United States, Israel and the Netherlands. Cross sectional analyses of data from three clinical databases. *BMC Health Serv Res* 2008;8:163.
- 37 McKenzie R, May S. The Cervical and thoracic spine: mechanical diagnosis and therapy. 2nd ed. Waikanae: Spinal Publications Ltd; 2006.
- 38 Gracey JH, McDonough SM, Baxter GD. Physiotherapy management of low back pain: a survey of current practice in Northern Ireland. *Spine* 2002;27:406-11.
- 39 Cohen JA. Coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37-46.
- 40 Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005;85:257-68.
- 41 Fleiss J. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;76:378-82.
- 42 Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol* 1993;46:423-9.
- 43 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
- 44 Whyte J, Hart T. It's more than a black box; it's a Russian doll: defining rehabilitation treatments. *Am J Phys Med Rehabil* 2003;82:639-52.
- 45 Werneke M, Hart DL, Cook D. A descriptive study of the centralization phenomenon. A prospective analysis. *Spine* 1999;24:676-83.
- 46 Werneke MW, Hart DL, George SZ, Stratford PW, Matheson JW, Reyes A. Clinical outcomes for patients classified by fear-avoidance beliefs and centralization phenomenon. *Arch Phys Med Rehabil* 2009;90:768-77.
- 47 Werneke MW, Hart DL. Centralization: association between repeated end-range pain responses and behavioral signs in patients with acute non-specific low back pain. *J Rehabil Med* 2005;37:286-90.
- 48 Cyriax J. Textbook of orthopaedic medicine. 10th ed. London: Bailliere Tindall; 1982.
- 49 Flynn T, Fritz J, Whitman J, Wainner R, Magel J, Rendeiro D, et al. A clinical prediction rule for classifying patients with low back pain who demonstrate short-term improvement with spinal manipulation. *Spine* 2002;27:2835-43.
- 50 Maitland GD. Vertebral manipulation. 5th ed. London: Butterworth; 1986.
- 51 van der Windt D, Hay E, Jellema P, Main C. Psychosocial interventions for low back pain in primary care: lessons learned from recent trials. *Spine* 2008;33:81-9.
- 52 Cleland JA, Childs JD, Fritz JM, Whitman JM. Interrater reliability of the history and physical examination in patients with mechanical neck pain. *Arch Phys Med Rehabil* 2006;87:1388-95.
- 53 Fritz JM, Delitto A, Vignovic M, Busse RG. Interrater reliability of judgments of the centralization phenomenon and status change during movement testing in patients with low back pain. *Arch Phys Med Rehabil* 2000;81:57-61.

- 54 Strender LE, Sjoblom A, Sundell K, Ludwig R, Taube A. Interexaminer reliability in physical examination of patients with low back pain. *Spine* 1997;22:814–20.
- 55 Holdom A. The use of the Mckenzie approach to treat back pain. *Br J Ther Rehabil* 1996;3:7–10.
- 56 Eastlack ME, Arvidson J, Snyder-Mackler L, Danoff JV, McGarvey CL. Interrater reliability of videotaped observational gait-analysis assessments. *Phys Ther* 1991;71:465–72.
- 57 Krebs DE, Edelstein JE, Fishman S. Reliability of observational kinematic gait analysis. *Phys Ther* 1985;65:1027–33.
- 58 Kibler WB, Uhl TL, Maddux JW, Brooks PV, Zeller B, McMullen J. Qualitative clinical evaluation of scapular dysfunction: a reliability study. *J Shoulder Elbow Surg* 2002;11:550–6.
- 59 Clare HA, Adams R, Maher CG. Reliability of McKenzie classification of patients with cervical or lumbar pain. *J Manipulative Physiol Ther* 2005;28:122–7.
- 60 Dankaerts W, O'Sullivan PB, Straker LM, Burnett AF, Skouen JS. The inter-examiner reliability of a classification method for non-specific chronic low back pain patients with motor control impairment. *Man Ther* 2006;11:28–39.
- 61 Delitto A, Shulman AD, Rose MJ, Strube MJ, Erhard RE, Bowling RW. Reliability of a clinical examination to classify patients with low back syndrome. *Physiother Theory Pract* 1992;1:1–9.
- 62 Dionne C, Bybee RF, Tomaka J. Correspondence of diagnosis to initial treatment for neck pain. *Physiotherapy* 2007;93:62–8.
- 63 Dionne CP, Bybee RF, Tomaka J. Inter-rater reliability of McKenzie assessment in patients with neck pain. *Physiotherapy* 2006;92:75–82.

Appendix: Intervention techniques

No.	Intervention techniques
	Therapeutic exercise
1	Directional preference – extension
2	Directional preference – flexion
3	Directional preference – lateral
4	Upper extremity stretch
5	Lower extremity stretch
6	Spinal stretch
7	Core spinal stabilization exercises
8	General spinal stabilization exercises
9	Pilates spinal stabilization exercises
10	Upper extremity stabilization – strengthening/endurance
11	Lower extremity stabilization – strengthening/endurance
12	Aerobic exercise
13	Circuit training
14	Group exercise therapy
15	Upper extremity neural tension exercise
16	Lower extremity neural tension exercise
17	Balance, coordination, and agility exercises
18	Gait exercises
19	Assistive exercise device(s)
	Manual techniques
20	Lumbar extension in lying with clinician overpressure
21	Lumbar extension mobilization
22	Lumbar rotation mobilization and extension
23	Lumbar sustained rotation/mobilization in flexion

Appendix: Continued

No.	Intervention techniques
24	Lumbar therapist correction of lateral shift
25	Cervical extension mobilization – prone
26	Cervical extension mobilization – supine
27	Cervical flexion mobilization – supine
28	Cervical lateral flexion mobilization – sitting
29	Cervical lateral flexion mobilization – supine
30	Cervical retraction with therapist overpressure
31	Cervical rotation mobilization – sitting
32	Cervical rotation mobilization – supine
33	Cervical traction – supine
34	Retraction and extension with traction and rotation – supine
35	Lumbar extension manipulation
36	Lumbar rotation manipulation in extension
37	Lumbar rotation manipulation in flexion
38	Cervical lateral flexion manipulation in sitting or supine
39	Cervical rotation manipulation in sitting or supine
40	Joint accessory mobilization/manipulation
41	Soft tissue mobilization/massage
	Educational techniques
42	Self exercise program
43	Bending or stooping posture position
44	Lumbar roll prescription
45	Lying posture position
46	Night roll prescription
47	Sitting posture position
48	Standing posture position
49	Patient-therapist educational treatment agreement/treaty
50	Pathoanatomical impairment model explanation
51	Back school
52	McKenzie booklet
53	Bed rest
54	Active rest in presence of healing
	Functional training techniques
55	Work and activities daily living
56	On site ergonomic modification
57	Lifting
	Modality procedures
58	Heat modality
59	Ultrasound
60	Ice modality
61	REPEX repeated end range passive exercise
62	Lumbar mechanical traction
63	Cervical mechanical traction
64	Electric stimulation modality
65	Electrotherapeutic delivery of medication
	Cognitive behavioral techniques
66	Operant graded activity program
67	Exposure <i>in vivo</i> program
68	Problem solving
69	Positive thinking
70	Relaxation breathing
71	Meditation
	Administrative technique
72	Refer to another medical clinic