Evaluation of a Treatment-Based Classification Algorithm for Low Back Pain: A Cross-Sectional Study

Tasha R. Stanton, Julie M. Fritz, Mark J. Hancock, Jane Latimer, Christopher G. Maher, Benedict M. Wand, Eric C. Parent

Background. Several studies have investigated criteria for classifying patients with low back pain (LBP) into treatment-based subgroups. A comprehensive algorithm was created to translate these criteria into a clinical decision-making guide.

Objective. This study investigated the translation of the individual subgroup criteria into a comprehensive algorithm by studying the prevalence of patients meeting the criteria for each treatment subgroup and the reliability of the classification.

Design. This was a cross-sectional, observational study.

Methods. Two hundred fifty patients with acute or subacute LBP were recruited from the United States and Australia to participate in the study. Trained physical therapists performed standardized assessments on all participants. The researchers used these findings to classify participants into subgroups. Thirty-one participants were reassessed to determine interrater reliability of the algorithm decision.

Results. Based on individual subgroup criteria, 25.2% (95% confidence interval [CI] = 19.8%–30.6%) of the participants did not meet the criteria for any subgroup, 49.6% (95% CI = 43.4%–55.8%) of the participants met the criteria for only one subgroup, and 25.2% (95% CI = 19.8%–30.6%) of the participants met the criteria for more than one subgroup. The most common combination of subgroups was manipulation/specific exercise (68.4% of the participants who met the criteria for 2 subgroups). Reliability of the algorithm decision was moderate (κ = 0.52, 95% CI = 0.27–0.77, percentage of agreement = 67%).

Limitations. Due to a relatively small patient sample, reliability estimates are somewhat imprecise.

Conclusions. These findings provide important clinical data to guide future research and revisions to the algorithm. The finding that 25% of the participants met the criteria for more than one subgroup has important implications for the sequencing of treatments in the algorithm. Likewise, the finding that 25% of the participants did not meet the criteria for any subgroup provides important information regarding potential revisions to the algorithm’s bottom table (which guides unclear classifications). Reliability of the algorithm is sufficient for clinical use.
Despite the ever-increasing numbers of randomized controlled trials (RCTs) evaluating interventions for low back pain (LBP), treatment outcomes remain less than optimal. Unfortunately, the majority of RCTs testing the effectiveness of LBP treatments demonstrate negligible to small effects. One hypothesis for these small effects is that by using broad inclusion criteria in trials, we are testing interventions in a heterogeneous population of patients with LBP (eg, those with nonspecific LBP) and such a diverse group of patients may not all respond to the one treatment. This hypothesis (that not all patients should be expected to respond to a single treatment) has led to a movement in research to identify subgroups of patients who may respond preferentially to certain treatments, in the hope that treatment effects will be larger in these subgroups.

Recent research has identified subgroups of patients who may respond best to 4 common physical therapy interventions for LBP: spinal manipulation, stabilization therapy (eg, motor control training), specific exercise (eg, direction-specific exercise), and traction. Criteria for inclusion in 2 of the treatment subgroups (manipulation and stabilization) were identified by creating clinical prediction rules, with the manipulation subgroup later tested in RCTs. The criteria for the other 2 treatment subgroups (specific exercise and traction) were based on prevalent theories of clinical decision making, subsequently supported by RCTs. These studies limited the inclusion criteria to those patients believed to respond to the treatment, then compared groups who received and did not receive the treatment of interest. For the remainder of this article, the criteria for the 4 subgroups identified by these studies will be termed individual subgroup criteria.

In order to translate the results of these individual studies into a clinical reasoning strategy applicable to more routine clinical circumstances, a decision-making flowchart (termed the “classification algorithm”) was developed (Figure). The classification algorithm attempted to clarify 2 important considerations required for clinical utility. First, to be a clinically useful decision-making aid, a classification algorithm should place each patient into only one treatment category. Second, it should be comprehensive, able to categorize all patients for whom it is intended. If either of these requirements is not met, it will be unclear what treatment a therapist should provide, a result that is contrary to the purpose of an algorithm.

To meet these requirements, modifications to the individual study subgroup criteria were necessary. Because an individual patient with LBP may exhibit subgroup criteria from more than one category, a hierarchical order of clinical findings had to be proposed in the algorithm to result in a single treatment category for each patient. Conversely, because an individual patient may not clearly fit the subgroup criteria for any one treatment category, the algorithm had to provide additional criteria to help a therapist choose the “best” treatment subgroup for these patients. For the remainder of the article, the revised subgroup criteria used by the classification algorithm will be termed comprehensive algorithm. Preliminary evidence exists to support the effectiveness of the comprehensive algorithm (and thus these modifications). An RCT showed that patients who received “matched” treatment according to their classification based on the algorithm (3-treatment subgroup version) had significantly lower disability levels at 4 weeks and 1 year than those who received “unmatched” treatment.

By studying the translation of individual studies into the broader clinical decision-making algorithm, it may be possible to improve or refine the comprehensive algorithm. In creation of the comprehensive algorithm, this translation included the decisions made regarding modification of the individual subgroup criteria and development of the hierarchical structuring. The majority of these decisions were based upon anecdotal evidence and assumptions of what findings should take precedence over others, and so need to be tested. For example, although mutual exclusivity among classification categories is forced by the structure of the comprehensive algorithm, the degree to which treatment subgroups identified by the comprehensive algorithm actually overlap based on individual subgroup criteria is currently unclear.

There are case reports in the literature of patients who have met the individual subgroup criteria of more than one treatment subgroup, but there has been no systematic attempt to quantify the nature or degree of overlap among categories in a clinically representative sample. Data are needed to determine the
extent of overlap among subgroups and establish in which subgroups overlap is most common using the individual subgroup criteria that informed the comprehensive algorithm. In addition, it is unclear how many patients do not actually meet the criteria for any of the 4 treatment subgroups based on the individual subgroup criteria and how often the additional criteria in the second part of the comprehensive algorithm are used to classify a patient. The extent of these issues would be best established by a prospective survey of consecutive cases of patients receiving physical therapy. Such a study would provide important information on the comprehensive algorithm and could assist in recommendations for its modification and improvement.

Therefore, this study had 3 purposes: (1) to determine the proportion of patients meeting the criteria for each treatment subgroup, more than one subgroup, or none of the subgroups and characterize the subgroups that overlap if applicable using the criteria from the individual studies that informed the classification algorithm (individual subgroup criteria); (2) to determine the prevalence of subgroups based on the comprehensive classification algorithm (comprehensive algorithm); and (3) to assess the interrater reliability of the classification decision when using the comprehensive algorithm in therapists with limited experience using the algorithm.

**Method**

**Study Design**

A cross-sectional, observational study was used to determine the prevalence of patients meeting the criteria for each treatment subgroup. A test-retest study design was used in a subset of patients to determine
Evaluation of a Treatment-Based Classification Algorithm for Low Back Pain

interrater reliability of the classification decision when using the comprehensive algorithm.

Setting
Study recruitment occurred from June 2008 to June 2010 in Sydney, New South Wales, Australia, and Salt Lake City, Utah, and from June 2009 to June 2010 in Perth and Bunbury, Western Australia, Australia. Patients were recruited from teaching hospitals (Sydney, Australia) and from private physical therapy clinics (Australia and United States). The interrater reliability study was performed only at the Sydney, Australia, sites.

Participants
A total of 250 patients with acute or subacute LBP were recruited into the study. In order to be included in this study, patients had to be between the ages of 18 and 65 years, have LBP in the area bounded superiorly by T12 and inferiorly by the buttock crease (with or without leg pain) that had lasted for more than 24 hours but less than 90 days, and have a modified Oswestry Low Back Pain Disability Questionnaire (ODQ) score of ≥20%. Patients were excluded if their most prominent symptoms were something other than LBP or leg pain (eg, pain in lumbar, thoracic, and cervical spine, where cervical pain is the worst), if they were currently pregnant, or if they had lumbar surgery in the preceding 6 months, previous spinal fusion or scoliosis rods or screws, any known or suspected serious pathology (eg, fracture, tumor), spinal steroid injections within the last month, or any previous sclerotic injections, botulinum toxin injections (in the low back, abdominal, or pelvic area), or denervation procedures.

Therapists
Registered physical therapists (from Australia and the United States) were recruited to act as assessors for this study. Therapists were instructed to enroll consecutive patients with LBP. In Australia, all therapists underwent standardized training in assessing the criteria needed to classify patients into subgroups and in applying the comprehensive algorithm. The training session lasted approximately 45 minutes and involved practice and discussion of assessment techniques and algorithm use. Continued communication between researchers and therapists occurred throughout the recruitment period to address any questions or problems. In the United States, patient data were obtained from 3 physical therapists who had been involved in clinical research using the comprehensive algorithm and had incorporated it into clinical care for at least 5 years. An additional 7 physical therapists in the United States with less experience using the algorithm received additional training conducted with similar methods as in Australia. Data regarding each physical therapist’s clinical experience (years at work), postgraduate training, and exposure to the algorithm and length of time using the algorithm (United States versus Australia) were recorded. Two of the authors (T.R.S., M.J.H.) also acted as assessors at the Sydney sites, particularly for the reliability assessments.

Examination Procedures and Classification Criteria
Prevalence study. Once consent was given to participate in the study, each eligible patient completed a numeric pain rating questionnaire to assess severity of back pain using an 11-point scale (0–10) and completed a pain drawing to show the location of the pain. The ODQ was used to assess LBP-specific disability. The Fear-Avoidance Beliefs Questionnaire (FABQ) was used to measure fear-avoidance beliefs about work and physical activity.

During the initial examination of the patient, a registered physical therapist collected data on all items required to classify patients using the comprehensive algorithm. Some items were assessed via questionnaires (as described above), whereas other items were assessed during the history taking or physical examination. The Appendix (available at ptjournal.apta.org) lists the items required to classify a patient and describes the protocol for assessing each of the individual items. A standardized assessment form was used for all patients.

Following a brief history taking (eg, collecting information regarding age, sex, prior history of LBP, best and worst positions for LBP, and duration of LBP), therapists began by completing a neurological examination on each patient. Muscle strength (force-generating capacity), sensation, and reflexes (knee and ankle jerk) were tested, with each category coded as either normal or abnormal. Straight-leg-raise range of motion then was assessed bilaterally using an inclinometer (also checking for a positive straight leg raise or crossed straight leg raise). Patients were observed during spinal movement for the presence of aberrant movements such as an instability “catch,” painful arc, reversal of lumbopelvic rhythm, or thigh climbing (present or absent). Following this examination, an assessment of active spinal movements was performed. Single and repeated trunk flexion and trunk extension in a standing position, single and repeated trunk flexion in a sitting position, sustained extension in a prone position (30 seconds), and single and repeated trunk extension in a prone position were performed. Repeated movements were capped at 10 repetitions. The effect of spinal movements on pain intensity (no change, increased, or decreased) and...
pain location (no change, centralization, or peripheralization) was recorded.

While the patient was in a prone position, passive range of hip medial (internal) rotation was assessed using an inclinometer. Spinal mobility then was assessed using a posterior-anterior pressure test on the L1–5 spinous processes, with spinal segments judged as being normal, hypermobile, or hypomobile. Whether the mobility assessment caused concordant pain also was recorded. If pain was present with mobility testing, the prone instability test was performed and its outcome recorded (positive or negative). Therapists were advised that they could perform any additional examination procedures on the patient that they felt necessary; however, they were instructed to perform these procedures following their assessment of the comprehensive algorithm criteria.

Interrater reliability study. Patients giving informed consent to participate in the reliability study were reassessed by a researcher (T.R.S., M.J.H.) who was blinded to the outcome of the initial assessment. The same testing procedure as in the initial assessment was used, and the reliability assessment was completed within 4 days of the initial assessment. Only patients who remained stable between examinations were included. Stable patients were defined as those having less than 6 points of change on the ODQ. Six points has been shown to be the minimal clinically important difference on the ODQ.

Outcome Measures
Classifying patients for the prevalence study. To determine the prevalence of patients meeting the criteria for each treatment subgroup, 2 methods of classification were used. First, patients were classified using the individual subgroup criteria that informed the algorithm (Table 1).

The criteria for specific exercise were updated to include patients who exhibited either centralization or a directional preference with spinal movement testing. Using the data on each criterion provided by the assessing therapist, the researchers rated each patient as either meeting or not meeting the criteria for each of the 4 treatment subgroups (manipulation, stabilization, specific exercise, and traction) for the individual studies.

Second, patients were classified using the comprehensive algorithm. The algorithm is structured in 2 parts (Figure). The first part is a set of hierarchical boxes that guide the therapist to give precedence to certain subgrouping findings over other findings. If a patient meets the criteria within one of these boxes, the classification category is determined for that patient and the classification criteria (no change, centralization, or peripheralization) was recorded.

While the patient was in a prone position, passive range of hip medial (internal) rotation was assessed using an inclinometer. Spinal mobility then was assessed using a posterior-anterior pressure test on the L1–5 spinous processes, with spinal segments judged as being normal, hypermobile, or hypomobile. Whether the mobility assessment caused concordant pain also was recorded. If pain was present with mobility testing, the prone instability test was performed and its outcome recorded (positive or negative). Therapists were advised that they could perform any additional examination procedures on the patient that they felt necessary; however, they were instructed to perform these procedures following their assessment of the comprehensive algorithm criteria.

| Treatment Subgroup | Individual Subgroup Criteria |
|--------------------|-----------------------------|
| Manipulation\(^{12}\) (must meet 4 or more criteria) | Duration of symptoms <16 days At least one hip with less than 35° of medial (internal) rotation Lumbar hypomobility No symptoms distal to the knee Fear-Avoidance Beliefs Questionnaire work subscale score <19 |
| Stabilization\(^{14}\) (must meet 3 or more criteria) | Age <40 y Average straight leg raise >91° Aberrant movement present Positive prone instability test |
| Specific exercise\(^{10,35}\) | Demonstrated centralization or a directional preference (an improvement in pain intensity) during repeated movement testing in any one position (standing, sitting, or lying) |
| Traction\(^{11}\) (must meet all criteria) | Signs and symptoms of nerve root compression (positive straight leg raise or reflex, sensory, or muscle strength deficit) and Pain or numbness extending distal to the buttock in the previous 24 hours and Peripheralization of pain with extension or positive crossed straight leg raise |

In this RCT, patients exhibiting a directional preference (decrease in pain intensity with spinal movement testing) who received treatment matched to their directional preference had better outcomes than those who received unmatched treatment. Therefore, the criteria for specific exercise were updated to include patients who exhibited either centralization or a directional preference with spinal movement testing. Using the data on each criterion provided by the assessing therapist, the researchers rated each patient as either meeting or not meeting the criteria for each of the 4 treatment subgroups (manipulation, stabilization, specific exercise, and traction) for the individual studies.
process is complete. If the patient does not meet the subgroup criteria in any of the hierarchical boxes in the first part of the algorithm, the therapist is directed to the second part of the algorithm that provides additional classification criteria in the bottom box. At this point, the therapist must select the one classification that best fits the patient based on the criteria provided. Using this 2-part process, the comprehensive algorithm ensures that each patient will be placed into one, and only one, classification category. The treatment subgroup classification made by the assessing therapist was used for the comprehensive algorithm prevalence rates.

Interrater reliability of the comprehensive algorithm classification decision. To determine the interrater reliability of the classification decision, the treatment subgroup chosen by each therapist using the comprehensive algorithm was compared. In order to evaluate potential sources of error in the comprehensive algorithm decision, the reliability of the specific classification criteria used in the algorithm was also assessed (eg, criteria in the hierarchical boxes and in the bottom table).

Sample size estimation. We designed the study to include 250 participants, as this sample size would provide sufficiently precise estimates of subgroup prevalence across the likely ranges of prevalence. For example, if the prevalence of a subgroup was 5.2%, the 95% confidence interval (CI) would span 2.5% to 9.0%; for a prevalence of 50%, the 95% CI would span 44% to 56%.

Data Analysis
Prevalence. Prevalence rates were calculated in 2 ways. First, using the individual subgroup criteria, the prevalence of patients meeting the criteria for each treatment subgroup was calculated as: number of patients meeting subgroup criteria/N, where N is the total number of patients (N=250). These prevalence rates include patients who meet the criteria for more than one subgroup or no subgroups; therefore, they may not sum to 100%. These prevalence rates represent the proportion of patients seeking care who would be expected to meet the criteria for each treatment subgroup based on the individual subgroup criteria evaluated in isolation from the other subgroups. In addition, the prevalence of patients meeting the criteria for no subgroup, only 1 subgroup, only 2 subgroups, only 3 subgroups, or only 4 subgroups was calculated by dividing the number of patients by N (calculated generally and for each subgroup or subgroup combination). These prevalence rate calculations allow a patient to meet the criteria for only one of these groups (eg, patients who meet the criteria for the manipulation + stabilization subgroup combination are only counted under patients meeting only 2 subgroups); therefore, the prevalence rates sum to 100%. These prevalence rates represent the specific breakdown of which treatment subgroups tend to overlap and how many patients do not meet the individual subgroup criteria for any treatment subgroups.

Second, using the comprehensive algorithm, the prevalence rate for each classification category was calculated as the number of patients placed into that category using the algorithm/N. Because the comprehensive algorithm requires each
Interrater reliability. Percentage of agreement and kappa coefficients with 95% CIs were calculated for the reliability of the classification decision between raters (when using the comprehensive algorithm) and for the reliability of the specific criteria comprising the comprehensive algorithm. If appropriate, a prevalence and bias adjusted kappa (PABAK) value also was given. When missing data values were present for a variable for one rater only, this patient was removed from the reliability analysis, as a kappa coefficient could not be calculated. A sensitivity analysis was performed to determine whether any differences in reliability occurred between patients with “clear classification” (ie, based on the hierarchical boxes) and those with “unclear classification” (ie, based on using the second part of the comprehensive algorithm, the criteria in the bottom box).

**Results**

**Prevalence Study**

A total of 545 patients were screened for inclusion to this study, with 295 excluded, reaching the goal of 250 patients. Reasons for exclusion were: LBP for more than 90 days (n=82), no current LBP (n=8), age greater than 65 years (n=62), age less than 18 years (n=6), ODQ score less than 20% (n=22), primary complaint not LBP (n=4), surgery within the previous 6 months (n=8), injections within the previous month.
(n=6), currently pregnant (n=2), did not provide consent (n=10), unable to complete assessment due to therapist or patient timing issues (n=78), and incomplete data (n=12). Of the 250 patients included, 125 were recruited from the United States and 125 were recruited from Australia (98 recruited from Sydney, Australia, and 27 recruited from Perth or Bunbury, Australia). Patient demographics are shown in Table 2. Of all patients recruited, 82.8% were recruited from private practice settings, 16% from hospital settings, and 1.2% from another source (eg, seeing a physical therapist but not recruited from that therapist’s clinic). Information regarding the physical therapists involved in the study is presented in Table 3.

The prevalence rates for each treatment subgroup for classification using the individual subgroup criteria are provided in Table 4. Based on the individual subgroup criteria, the most common treatment subgroup for which the criteria were met was specific exercise (prevalence of 44.8%), followed by manipulation (prevalence of 35.2%). Of the whole sample, 49.6% of the patients met the criteria for only one subgroup (n=124), and of these patients, 21.6% met the criteria for only the specific exercise group. The majority of patients met the criteria for at least one treatment subgroup, with 25.2% of the patients meeting the criteria for more than one subgroup. The most common combination of subgroups for which the criteria were met was manipulation + specific exercise. Of the 22.8% of patients meeting the criteria for only 2 subgroups, 68.4% met the criteria for the manipulation + specific exercise subgroup combination. Only a small number of patients met the criteria for 3 treatment subgroups (manipulation, stabilization, and specific exercise), calculated to be 2.4% of the total sample. The prevalence of patients who did not meet the individual subgroup criteria for any of the treatment subgroups was 25.2%.

Using the comprehensive algorithm, manipulation was the most commonly met treatment subgroup, with 42.0% of the patients assigned by the algorithm to this subgroup (Tab. 5). The prevalence of patients who met the criteria for the specific exercise subgroup was 30.8%. Of the whole

---

**Table 4.**

Prevalence Rates Using the Individual Subgroup Criteria

| Treatment Subgroup | No. of Patients | Prevalence, % (95% CI) |
|--------------------|----------------|-----------------------|
| Patient can meet criteria for multiple subgroups (prevalence rates do not sum to 100%) | | |
| Manipulation | 88 | 35.2 (29.3 to 41.1) |
| Stabilization | 32 | 12.8 (8.7 to 17.0) |
| Specific exercise | 112 | 44.8 (38.6 to 51.0) |
| Traction | 24 | 9.6 (6.0 to 13.3) |
| Patient can meet criteria for only one subgroup or subgroup combination (prevalence rates sum to 100%) | | |
| No subgroups | 63 | 25.2 (19.8 to 30.6) |
| One subgroup | 124 | 49.6 (43.4 to 55.8) |
| Manipulation | 38 | 15.2 (10.8 to 19.7) |
| Stabilization | 16 | 6.4 (3.4 to 9.4) |
| Specific exercise | 54 | 21.6 (16.3 to 26.7) |
| Traction | 16 | 6.4 (3.4 to 9.4) |
| Two subgroups | 57 | 22.8 (17.6 to 28) |
| Manipulation + stabilization | 4 | 1.6 (0.04 to 3.16) |
| Manipulation + specific exercise | 39 | 15.6 (11.1 to 20.1) |
| Manipulation + traction | 1 | 0.4 (~0.38 to 1.18) |
| Stabilization + specific exercise | 6 | 2.4 (0.5 to 4.3) |
| Specific exercise + traction | 7 | 2.8 (0.75 to 4.85) |
| Three subgroups | 6 | 2.4 (0.5 to 4.3) |
| Manipulation + stabilization + specific exercise | 6 | 2.4 (0.5 to 4.3) |

*Prevalence = number of patients/250. CI = confidence interval.
Evaluation of a Treatment-Based Classification Algorithm for Low Back Pain

Table 5.
Prevalence Rates for Treatment Subgroups Using the Comprehensive Algorithm, Further Separated Into Clear Classification and Unclear Classification

| Treatment Subgroup | Comprehensive Algorithm | Clear Classification | Unclear Classification |
|--------------------|-------------------------|---------------------|------------------------|
|                    | No. of Patients | Prevalence, % (95% CI) | No. of Patients | Prevalence, % (95% CI) | No. of Patients | Prevalence, % (95% CI) |
| Manipulation       | 105           | 42.0 (35.9 to 48.1)    | 78           | 31.2 (25.5 to 36.9)    | 27           | 10.8 (7.0 to 14.7)    |
| Stabilization      | 44            | 17.6 (12.9 to 22.3)    | 19           | 7.6 (4.3 to 10.9)      | 25           | 10.0 (6.3 to 13.7)    |
| Specific exercise  | 77            | 30.8 (25.1 to 36.5)    | 44           | 17.6 (12.9 to 22.3)    | 33           | 13.2 (9.0 to 17.4)    |
| Traction           | 24            | 9.6 (6.0 to 13.3)      | 24           | 9.6 (6.0 to 13.3)      | 0            | -              |
| Total              | 250           | 100                   | 165          | 66.0 (60.1 to 71.9)    | 85           | 34.0 (28.1 to 39.9)   |

* Prevalence-number of patients/250. CI = confidence interval.

sample, 66.0% had a clear classification using the comprehensive algorithm, and 34.0% had an unclear classification (ie, the second part of the algorithm consisting of the criteria in the bottom box was used). The unclear classifications were divided between the manipulation subgroup (10.8%), the stabilization subgroup (10.0%), and the specific exercise subgroup (13.2%); every patient assigned to the traction subgroup was classified using the hierarchical boxes in the top part of the algorithm.

The clinical tests required for classification of patients were consistently performed on all patients, with the exception of the prone instability test. Various reasons were given by the therapists for not performing this test, such as the patient having too much pain or the patient refusing or being unable to perform the test. The prone instability test was not performed by 19 patients, 6 of whom had 2 of 4 stabilization criteria. If the prone instability test had been completed and was positive, these patients would have met the stabilization subgroup criteria (eg, at least 3 of 4 criteria present). Therefore, in 6 patients, these missing test results may have affected their classification. If all 6 patients had been positive on the test, this would have increased the prevalence of the stabilization subgroup to 15.2% (95% CI=10.8% to 19.7%) for the individual subgroup criteria.

Reliability Study
A total of 32 patients (of the 98 patients recruited from Sydney for the prevalence study) agreed to participate in the reliability study, with only 1 patient not remaining stable between the initial and second assessments. Of these patients, 74.7% were recruited from hospital settings, 12.9% from private practice settings, and 9.7% from another source (eg, seeing a physical therapist but not recruited from that therapist’s clinic). Table 2 presents the patient demographics, and Table 3 presents the therapist demographics. Completion of the second reliability assessment was done within a time frame ranging from immediately following the first assessment to 4 days after the first assessment. The majority of assessments (90%) were completed directly following the first assessment.

Reliability of the classification decision using the comprehensive algorithm was moderate, with a kappa value of 0.52 (95% CI=0.27 to 0.77) and a percentage of agreement of 67%. In the first assessment, 38.7% of the patients had an unclear classification (ie, the additional classification criteria in the bottom table of the algorithm were used). For the second assessment, 61.3% of the patients had an unclear classification. When considering only patients with a clear classification (using the first rater’s data; n=19), the interrater reliability of the classification decision (using the comprehensive algorithm) was good, with a kappa value of 0.69 (95% CI=0.42 to 0.96) and a percentage of agreement of 78.9%. Conversely, when considering only patients with an unclear classification (n=12), the reliability of the classification decision was poor (kappa=0.23, 95% CI=−0.21 to 0.66, percentage of agreement=50%). Table 6 presents the reliability values of the specific criteria of the comprehensive algorithm, with kappa values ranging from −0.11 to 0.78.

Discussion
This study demonstrated that subgroups of patients with LBP identified by the individual subgroup criteria were mutually exclusive in approximately 50% of cases (eg, patients met the criteria for only one subgroup). In the remaining 50% of cases, patients either met the criteria for more than one subgroup (25.2%) or did not meet the criteria for any subgroups (25.2%). Important aspects of the content validity of classification systems are exhaustiveness and mutual exclusivity of their
classification categories. Therefore, these results confirm the necessity of developing a comprehensive algorithm for translation of the classification criteria into clinical practice. When classified based on the comprehensive algorithm that was constructed to overcome the concerns about exhaustiveness and exclusivity of the classification categories identified in individual research studies, 34.0% of the patients had an unclear classification. An unclear classification appeared to adversely affect the interrater reliability of the comprehensive algorithm. These findings provide important information for evaluating the translation of the individual subgroup criteria into the comprehensive algorithm. They also provide essential information for guiding additional research and refinement of the comprehensive algorithm in terms of optimal hierarchical ordering of treatments, potential addition of further treatments to the algorithm, and modifications to the criteria used in the algorithm. The reliability of the classification decision when using the comprehensive algorithm was moderate, suggesting it is appropriate for use in a clinical setting.

The prevalence rates for each subgroup based on the individual subgroup criteria were generally similar to those in the literature, with the exception of the stabilization subgroup. Direct comparison of prevalence rates between the current study and previous work are made with caution due to differences in outcome definitions and patient populations. For the manipulation subgroup, previous studies’ prevalence rates ranged from 23% to 59%. (present study: 35.2%, 95% CI=29.3% to 41.1%). For stabilization, prevalence rates from the literature ranged from 24% to 26% (present study: 12.8%, 95% CI=8.7% to 17.0%). The reason for

![Table 6](image-url)

Evaluation of a Treatment-Based Classification Algorithm for Low Back Pain

Table 6.
Reliability of the Specific Criteria of the Comprehensive Algorithm (n=31)

| Classification Category | Kappa (95% CI) | Percentage of Agreement |
|-------------------------|---------------|------------------------|
| Centralize with 2 or more movements in the same direction? | -0.11 (-0.19 to -0.02) 0.61 | 80.6 |
| Centralize with movement in one direction and peripheralize with movement in opposite direction? | -0.08 (-0.16 to 0.00) 0.67 | 83.9 |
| Directional preference for flexion or extension | 0.24 (-0.09 to 0.57) | 61.3 |
| Centralization with motion testing | 0.26 (-0.13 to 0.65) | 74.2 |
| Low back pain onlya | | |
| Status quo with all movements | 0.33 (0.01 to 0.66) | 67.7 |
| Duration of symptoms <16 days | 1.00 | 100 |
| No symptoms below the knee | 0.56 (0.24 to 0.87) | 80.6 |
| Hypomobility with spring testingd | 0.11 (-0.23 to 0.45) | 53.3 |
| Low FABQ-W scoresd | | |
| Hip medial (internal) rotation ROM >35° | 0.56 (0.24 to 0.87) | 80.6 |
| Increasing episode frequency | 0.33 (-0.05 to 0.71) | 77.4 |
| Peripheralization with motion testing | 0.45 (0.13 to 0.77) | 77.4 |
| Pain with spring testing | -0.08 (-0.16 to 0.00) | 83.9 |
| Stabilization | | |
| Patient’s age <40 y | 1.00 | 100 |
| Average SLR ROM >91 | 0.36 (-0.07 to 0.79) | 83.9 |
| Aberrant movements presentd | 0.00 (-0.32 to 0.32) | 50 |
| Prone instability test positive | 0.30 (-0.03 to 0.62) | 67.7 |
| Hypermobility with spring testingd | 0.10 (-0.29 to 0.49) | 63.3 |
| 3 or more prior episodes | 0.62 (0.32 to 0.92) | 83.9 |
| Discrepancy in SLR ROM >0° | 0.23 (-0.19 to 0.64) | 77.4 |
| Low FABQ-PA scoresd | | |
| Traction | | |
| Symptoms distal to the buttockc | 0.21 (0.08 to 0.34) | 74.2 |
| Signs of nerve root compression present | 0.41 (0.08 to 0.75) | 74.2 |
| Peripheralize with extension | 0.78 (0.38 to 1.19) | 96.8 |
| Positive crossed straight leg raise | 0.65 (0.02 to 1.28) | 96.8 |
| Peripheralization with no ability to centralize with movement | 0.31 (-0.09 to 0.70) | 80.6 |

a FABQ-W=Fear-Avoidance Beliefs Questionnaire work subscale, ROM=range of motion, FABQ-PA=Fear-Avoidance Beliefs Questionnaire physical activity subscale.
b Kappa value adjusted for low base rate in reliability study (prevalence and bias adjusted kappa [PABAK] values).
c Information was collected only once, therefore analysis of reliability is not applicable.
d n=30; 1 patient was removed, as missing data were present for 1 rater only.
the lower prevalence rate for the stabilization subgroup in the present study is unclear. For the specific exercise subgroup, previous prevalence rates ranged from 21% to 74% \(^{10,18,35,40}\) (present study: 44.8%, 95% CI = 38.6% to 51.0%). It is important to note that a full McKenzie diagnosis and treatment (MDT) \(^{41,42}\) assessment was not performed in the present study; rather, a modified version of patient response to repeated spinal movements was completed, excluding an assessment of lateral movements. Therefore, it is not surprising that when compared with prevalence rates of directional preference using a full MDT assessment \(^{74}\), our prevalence rates were lower (45%). For traction, prevalence rates in the literature were recently reported to be 9% \(^{40}\) (present study: 9.6%, 95% CI = 6.0% to 13.3%).

This study showed that approximately 25% of the patients with LBP met the criteria for more than one subgroup when using the individual subgroup criteria. This finding confirms the need for a hierarchical algorithm to define which subgroup will have priority when overlap occurs. Although some overlap was anticipated, the specific findings of this study provide important information for guiding future research into optimizing the comprehensive algorithm.

The most common subgroups that we found to overlap were the specific exercise and manipulation groups (using the individual subgroup criteria), accounting for 68.4% of patients who met 2 subgroups. Comparing the prevalence of classifications using the individual subgroup criteria with the classification made based on the comprehensive algorithm, it is clear that the algorithm tends to prioritize manipulation over the specific exercise classification. The reasons for this finding are apparent from examining the current algorithm. Although a clear classification of specific exercise is prioritized, as it is assessed prior to the manipulation criteria, the criteria for clear classification of specific exercise require centralization with movement and not the more prevalent finding of a directional preference. Use of a more conservative criterion of centralization likely moves many patients with a directional preference for movement, but not clear centralization, into the manipulation classification when applying the comprehensive algorithm for decision making. It is possible that some of these patients classified into the manipulation subgroup may achieve better outcomes with specific exercise. It also is possible that all patients with an overlap between specific exercise and manipulation should receive manipulation regardless of the presence of centralization. These are important areas for future research. It also is possible that patients meeting these 2 groups (or other combinations of groups) may do better if they receive both treatments initially rather than just the one treatment as per the decision based on the hierarchical comprehensive algorithm. This again is an important issue for future studies.

Another important finding of this study was that approximately 25% of the patients with LBP did not meet the criteria for any of the treatment subgroups when using the individual subgroup criteria. This is essential information for future research designed to refine the comprehensive algorithm and improve clinical outcomes. Currently, in these 25% of patients, use of the comprehensive algorithm would result in choosing a treatment subgroup that is unclear, as the algorithm forces all patients into a subgroup. Although the current study did not examine the clinical outcomes of the physical therapy care, it may be the case that patients with an unclear classification have inferior treatment outcomes. The implications of an unclear classification on clinical outcomes are an important topic for future research.

There are several possibilities for reducing the prevalence of unclear classifications using the comprehensive algorithm. First, the current comprehensive algorithm resulted in 34.0% of patients receiving an unclear classification using the additional criteria in the bottom table. Considering that 25% of the patients did not meet the individual subgroup criteria, this finding suggests that some modifications to the upper part of the algorithm could be made so the 9.0% of patients who did not meet the criteria for a subgroup based on the individual subgroup criteria could receive a clear classification based on the top part of the comprehensive algorithm. Second, future research into existing classifications may result in refined criteria that reduce the number of unclear classifications. Third, the current algorithm does not include all possible physical therapy interventions for LBP; it may be that the addition of another treatment subgroup is appropriate (as evidence permits). Finally, it may be that not all patients are suitable for application of the algorithm and that some patients with unclear classifications should be evaluated with a different decision-making paradigm. Future research could investigate whether clinical outcomes would be improved by limiting application of the comprehensive algorithm to patients with a clear classification and applying a different strategy to patients with an unclear classification based on the algorithm.

It is important to consider the clinical implications of these findings. The finding that many patients meet the criteria for more than one...
subgroup or do not clearly fit a subgroup (based on the individual subgroup criteria) suggests that when using the comprehensive algorithm, a therapist may begin with the treatment subgroup indicated by the algorithm and should monitor the patient’s response to the treatment. Careful monitoring of a patient’s response to treatment may be particularly important for those patients who do not clearly fit a subgroup and instead are classified using the bottom table of the algorithm.

Our study was the first to test the reliability of the comprehensive algorithm using 2 independent assessors. In a previous study, pairs of examiners scored previously obtained assessment findings from 60 patients with LBP using the earlier 3-treatment group classification algorithm. This design examines the reliability of the structure of the algorithm, but does not include error that may be due to the testing procedures of the 2 raters and is likely to overestimate the true interrater reliability. In this previous study, reliability was good (kappa = 0.60, 95% CI = 0.56 to 0.64, percentage of agreement = 76%). We found comparable reliability with kappa values at moderate levels (kappa = 0.52, 95% CI = 0.27 to 0.77, percentage of agreement = 67%). In our study, when considering only patients with a clear classification, the reliability of the classification decision when using the comprehensive algorithm was good (kappa = 0.69). When considering only patients with an unclear classification during the initial assessment, the reliability of the comprehensive algorithm decision was poor (kappa = 0.23). These findings are important because, as mentioned above, the proportion of patients for whom the bottom table was used to determine their classification was 34.0%. Based on the reliability findings, it is unlikely that 2 therapists who are novices to the comprehensive algorithm would choose the same treatment for a patient when using the bottom table of the algorithm. This finding also indicates the need for further research to reduce the prevalence of unclear classifications. If the reliability of classification when the bottom table is used is better in therapists experienced in algorithm use, this would provide valuable information to support intensive training of therapists who are novices to algorithm use to reduce variability in clinical decisions.

Although the overall reliability of the classification decision when using the comprehensive algorithm was moderate, the reliability of the specific criteria used to classify patients was variable. Some criteria had very good reliability, such as “peripheralize with extension,” which had a percentage of agreement of 96.8% (kappa = 0.76). However, other criteria, such as detecting the presence of aberrant movements, demonstrated poor reliability (kappa = 0.00). A previous study that investigated the reliability of detecting aberrant movements also demonstrated low kappa values (0.18; 95% CI = −0.07 to 0.43). Interestingly, in criteria related to assessment of centralization, negative kappa values were found, indicating that the reliability of detecting centralization was less than chance. However, for these categories, percentage of agreement was good (approximately 80%). In this situation, it appears that a low base rate paradox occurred. For example, the occurrence of patients who “centralize with 2 or more movements [positions] in the same direction” was rare, and when it was present, the raters never agreed on its presence. The kappa value for this criteria, adjusted for low prevalence, is 0.61, suggesting good reliability. Previous studies showed good to excellent reliability for the detection of centralization, with kappa values ranging from 0.50 to 1.00, which is in line with the adjusted kappa value. There is controversy regarding whether this adjustment should be made; therefore, it is suggested that both adjusted and unadjusted values should be reported.

A limitation of this study is that a relatively small sample was used for the reliability study, which resulted in somewhat imprecise reliability estimates. Furthermore, the majority of the patients in the reliability study were recruited from a hospital setting. Comparison of patient demographics between the prevalence study and the reliability study suggests that the latter had a greater percentage of women and recruited patients with a longer duration of LBP, which may result in reliability findings that are different from what we would find in patients who are recruited from a private practice or with a shorter duration of symptoms. Additionally, a relatively large percentage of patients who were screened did not meet the eligibility criteria (54%). Therefore, it is important to apply the results of this study only to a similar population. The large number of patients excluded emphasizes that the comprehensive algorithm was designed to apply to non-elderly patients with acute or subacute LBP. Elderly patients with LBP or those with chronic symptoms comprise a significant proportion of individuals with LBP receiving physical therapy, and these patients likely require a different decision-making strategy.

Although limitations are present, this study has good external validity. Patients were recruited from numerous locations (Sydney, Bunbury, and Perth, Australia, and Salt Lake City, Utah) and from a variety of settings (hospital and private practice physical therapy), generating a broad sample. The generalizability of the results was further supported by a
post hoc analysis of the country-specific prevalence rates (using the individual subgroup criteria) that demonstrated similar values between Australia and the United States (manipulation: 38.4% versus 32.0%; stabilization: 12.8% at both sites; specific exercise: 45.6% versus 44.0%; traction: 7.2% versus 12.0%). These prevalence rates are remarkably similar, given some of the differences in patient access to physical therapy care (eg, direct access in Australia, referral in the United States).

Conclusions
The results of this study suggest that approximately one half of patients with acute or subacute LBP seeking care meet the criteria for only one treatment subgroup using the individual subgroup criteria. Approximately 50% of patients either meet the criteria for more than one subgroup or none of the subgroups using the individual subgroup criteria. The decision-making strategy evaluated in this study (eg, the comprehensive algorithm) fit approximately one half of patients seeking physical therapy care for their LBP. This information is important for guiding future research that aims to refine or improve the comprehensive algorithm and physical therapy clinical decision making for patients with LBP. Further research is needed to explore the hierarchical ordering of the algorithm, the possibility of adding additional physical therapy treatments, and the refinement of criteria used in the algorithm. The overall reliability of the comprehensive algorithm classification decision is adequate for clinical use, but the reliability of the lower table is poor and should further be investigated.

Ms Stanton, Dr Fritz, Dr Hancock, Associate Professor Latimer, Professor Maher, and Dr Parent provided concept/idea/research design. All authors provided writing and consultation (including review of the manuscript before submission). Ms Stanton, Dr Fritz, Dr Hancock, and Associate Professor Wand provided data collection and participants. Ms Stanton, Dr Fritz, Dr Hancock, Associate Professor Latimer, and Professor Maher provided data analysis. Ms Stanton, Dr Fritz, Dr Hancock, Associate Professor Latimer, Professor Maher, and Associate Professor Wand provided project management and fund procurement. Dr Fritz, Dr Hancock, Associate Professor Latimer, and Professor Maher provided facilities/equipment. Dr Fritz, Dr Hancock, Associate Professor Wand, and Dr Parent provided institutional liaisons.

The study received ethical approval from the Human Research Ethics Committee of The University of Sydney, from the Human Research Ethics Committee of the University of Notre Dame Australia, and from the Sydney South West Area Health Service Ethics Board. The study qualified for exempt status from the Institutional Review Board of Intermountain Healthcare.

This research was presented at the Low Back Pain Forum; March 15–18, 2011; Melbourne, Victoria, Australia.

This project was supported by the Physiotherapy Research Foundation tagged Physiotherapists’ Registration Board of Western Australia research grant. The authors also acknowledge The University of Sydney International National Research Scholarship for Ms Stanton’s PhD funding, Australia’s National Health and Medical Research Council for funding of Professor Maher’s research fellowship, and the Australian Research Council for funding of Associate Professor Latimer’s Future Fellowship.

This article was submitted August 18, 2010, and was accepted December 3, 2010.

DOI: 10.2522/ptj.20100272

References
1. Mahler CG, Moseley AM, Sherrington C, et al. A description of the trials, reviews, and practice guidelines indexed in the PEDro database. *Phys Ther.* 2008;88:1068–1077.
2. Machado LA, Kamper SJ, Herbert RD, et al. Analgesic effects of treatments for non-specific low back pain: a meta-analysis of placebo-controlled randomized trials. *Rheumatology (Oxford).* 2009;48:520–527.
3. Costa LO, Mahler CG, Latimer J, et al. Motor control exercise for chronic low back pain: a randomized placebo-controlled trial. *Phys Ther.* 2009;89:1275–1286.
4. Hancock MJ, Maher CG, Latimer J, et al. Assessment of diclofenac or spinal manipulative therapy, or both, in addition to recommended first-line treatment for acute low back pain: a randomised controlled trial. *Lancet.* 2007;370:1638–1643.
5. Machado LA, Maher CG, Herbert RD, et al. The effectiveness of the McKenzie method in addition to first-line care for acute low back pain: a randomized controlled trial. *BMJ Clin Evid.* 2010;18.
6. Skillgate E, Vingard E, Alfredsson L. Naprapathic manual therapy or evidence-based care for back and neck pain: a randomised, controlled trial. *Clin J Pain.* 2007;23:431–439.
7. Juni P, Battaglia M, Nuesch E, et al. A randomised controlled trial of spinal manipulative therapy in acute low back pain. *Ann Rheum Dis.* 2009;68:1420–1427.
8. Delitto A. Research in low back pain: time to stop seeking the elusive “magic bullet.” *Phys Ther.* 2005;85:206–208.
9. Kent P, Keating J. Do primary-care clinicians think that nonspecific low back pain is one condition? *Spine (Phila Pa 1976).* 2004;29:1022–1031.
10. Browder DA, Cleland JD, Cleland JA, Fritz JM. Effectiveness of an extension-oriented treatment approach in a subgroup of subjects with low back pain: a randomized clinical trial. *Phys Ther.* 2007;87:1608–1618.
11. Cleland JD, Fritz JM, Flynn TW, et al. A clinical prediction rule to identify patients with low back pain most likely to benefit from spinal manipulation: a validation study. *Ann Intern Med.* 2004;141:920–928.
12. Flynn T, Fritz J, Whitman J, et al. A clinical prediction rule for classifying patients with low back pain who demonstrate short-term improvement with spinal manipulation. *Spine (Phila Pa 1976).* 2002;27:2835–2843.
13. Fritz JM, Lindsay W, Matheson JW, et al. Is there a subgroup of patients with low back pain likely to benefit from mechanical traction: results of a randomized clinical trial and subgrouping analysis. *Spine (Phila Pa 1976).* 2007;32:E793–E800.
14. Hicks GE, Fritz JM, Delitto A, McGill SM. Preliminary development of a clinical prediction rule for determining which patients with low back pain will respond to a stabilization exercise program. *Arch Phys Med Rehabil.* 2005;86:1753–1762.
15. Hoffman S, Maher CG, Latimer J, et al. Independent evaluation of a clinical prediction rule for spinal manipulative therapy: a randomised controlled trial. *Eur J Pain.* 2008;12:936–943.
16. Fritz JM, Brennan GP, Clifford SN, et al. An examination of the reliability of a classification algorithm for subgrouping patients with low back pain. *Spine (Phila Pa 1976).* 2006;31:77–82.
17. Fritz JM, Cleland JA, Cleland JD. Subgrouping patients with low back pain: evolution of a classification approach to physical therapy. [Erratum in: *J Orthop Sports Phys Ther.* 2007;37:769. *J Orthop Sports Phys Ther.* 2007;37:290–302.
Evaluation of a Treatment-Based Classification Algorithm for Low Back Pain

18 Brennan GP, Fritz JM, Hunter SJ, et al. Identifying subgroups of patients with acute/subacute “nonspecific” low back pain: results of a randomized clinical trial. *Spine (Phila Pa 1976).* 2006;31:623–631.

19 Gill NW, Teyhen DS, Lee IE. Improved contraction of the transversus abdominis immediately following spinal manipulation: a case study using real-time ultrasound imaging. *Man Ther.* 2007;12:280–285.

20 May S, Rosedale R. A case of a potential manipulation responder whose back pain resolved with flexion exercises. *J Manipul Physiol Ther.* 2007;30:539–542.

21 Jensen MP, Turner JA, Romano JM. What is the maximum number of levels needed in pain intensity measurement? *Pain.* 1994;58:387–392.

22 Mann NH III, Brown MD, Hertz DB, et al. Initial-impression diagnosis using low-back pain patient pain drawings. *Spine (Phila Pa 1976).* 1993;18:41–55.

23 Fritz JM, Irrgang JJ. A comparison of a modified Oswestry Low Back Pain Disability Questionnaire and the Quebec Back Pain Disability Scale [erratum in: *Phys Ther.* 2008;88:138–139]. *Phys Ther.* 2001;81:770–788.

24 Waddell G, Newton M, Henderson I, et al. A Fear-Avoidance Beliefs Questionnaire (FABQ), and the role of fear-avoidance beliefs in chronic low back pain and disability. *Pain.* 1993;52:157–168.

25 Magee Dj. *Orthopaedic Physical Assessment.* 3rd ed. Philadelphia, PA: Saunders; 1997.

26 Waddell G, Somerville D, Henderson I, Newton M. Objective clinical evaluation of physical impairment in chronic low back pain. *Spine (Phila Pa 1976).* 1992;17:617–628.

27 Nachemson A. Lumbar spine instability: a critical update and symposium summary. *Spine (Phila Pa 1976).* 1985;10:290–291.

28 Cyriax JH. Diagnosis of soft tissue lesions. In: Cyriax JH, ed. *Textbook of Orthopaedic Medicine.* 6th ed. Baltimore, MD: Williams & Wilkins; 1976:389.

29 MacNab I, McCulloch JA. *Backache.* 2nd ed. Baltimore, MD: Williams & Wilkins; 1990:159.

30 Ellison JB, Rose SJ, Sahrman SA. Patterns of hip rotation range of motion: a comparison between healthy subjects and patients with low back pain. *Phys Ther.* 1990;70:537–541.

31 Maher CG, Simmonds M, Adams R. Therapists’ conceptualization and characterization of the clinical concept of spinal stiffness. *Phys Ther.* 1998;78:289–300.

32 Maher CG, Adams R. Reliability of pain and stiffness assessments in clinical manual lumbar spine examination. *Phys Ther.* 1994;74:801–809.

33 Hicks GE, Fritz JM, Delitto A, Mishock J. Interrater reliability of clinical examination measures for identification of lumbar segmental instability. *Arch Phys Med Rehabil.* 2003;84:1856–1864.

34 Beurskens AJ, de Vet HC, Koke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. *Pain.* 1996;65:71–76.

35 Long A, Donelson R, Fung T. Does it matter which exercise: a randomized control trial of exercise for low back pain. *Spine (Phila Pa 1976).* 2004;29:2593–2602.

36 Cunningham M. *Paper 242: More Than Just the Kappa Coefficient: A Program to Fully Characterize Interrater Reliability Between Two Raters.* Washington, DC: SAS Global Forum; 2009.

37 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.

38 Buchbinder R, Goel V, Bombardier C, Hogg-Johnson S. Classification systems of soft tissue disorders of the neck and upper limb: do they satisfy methodological guidelines? *J Clin Epidemiol.* 1996;49:141–149.

39 Riddle DL. Classification and low back pain: a review of the literature and critical analysis of selected systems. *Phys Ther.* 1998;78:708–737.

40 George SZ, Bielosky JE, Donald DA. The centralization phenomenon and fear-avoidance beliefs as prognostic factors for acute low back pain: a preliminary investigation involving patients classified for specific exercise. *J Orthop Sports Phys Ther.* 2005;35:580–588.

41 McKenzie R, May S. The *Lumbar Spine: Mechanical Diagnosis & Therapy.* Vol 1. Raumati Beach, New Zealand: Spinal Publications; 2003.

42 McKenzie R, May S. The *Lumbar Spine: Mechanical Diagnosis & Therapy.* Vol 2. Raumati Beach, New Zealand: Spinal Publications; 2003.

43 Cicchetto DV, Feinstein AR. High agreement but low kappa. II: resolving the paradoxes. *J Clin Epidemiol.* 1990;43:551–558.

44 Feinstein AR, Cicchetto DV. High agreement but low kappa. I: the problems of two paradoxes. *J Clin Epidemiol.* 1990;43:543–549.

45 Aina A, May S, Clare H. The centralization phenomenon of spinal symptoms: a systematic review. *Man Ther.* 2004;9:134–143.

46 Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther.* 2005;85:257–268.