Reliability of a Treatment-Based Classification System for Subgrouping People With Low Back Pain

Despite the high cost associated with the fluctuating clinical course of low back pain (LBP), no treatment strategy surgical or conservative has been shown to be consistently effective in reducing the often persistent symptoms, functional limitations, and disability associated with this condition. The lack of beneficial effects of conservative treatments for LBP may be due to the lack of a pathoanatomy-specific diagnosis, as fewer than 20% of individuals with LBP can be given a specific, structurally based diagnosis.

In the absence of a specific pathoanatomical diagnosis and to better direct treatment, a number of research and clinical groups have suggested that there is a need for a system that classifies individuals with LBP based on key clinical symptoms and multidimensional features of the LBP presentation. The basis for this suggestion is that people with LBP represent a heterogeneous group, consisting of several smaller homogeneous subgroups. Logically, if the subgroups of patients were classified based on criteria relevant to their specific symptoms, these more homogeneous subgroups would have a higher likelihood of responding to matched treatment approaches. Such a classification system could be useful both in prognosis and treatment, rendering the development and testing of classification systems of LBP a top priority.

Numerous classification systems have been described for patients with LBP. Delitto and colleagues described a treatment-based classification (TBC) system and used information gathered from the patient history and physical examination to place a patient into 1 of 4 classification categories that directed patient treatment: manipulation, specific exercise, stabilization, and traction. The...
Using patients with acute or subacute musculoskeletal pain: individual subgroup criteria versus different methods involving the selection of 1 of the 4 treatment options for each patient. Although the reliability of clinicians in performing the individual exam items used in the clinical prediction rules has been shown to be fair to good, agreement among clinicians to classify patients with LBP using the TBC system has only been tested in a preliminary manner.

In a recent study, Stanton et al examined the prevalence of patients meeting the criteria for each TBC treatment subgroup using 2 different methodologies: individual subgroup criteria versus a comprehensive classification algorithm. Using patients with acute or subacute (duration less than 90 days) LBP, the authors found that approximately 50% of the participants met the criteria for only 1 TBC subgroup. Twenty-five percent of the patients met the criteria for more than 1 subgroup, and the other 25% of the patients did not meet the criteria for any subgroup. Given that only 50% of the cases were classified in a mutually exclusive manner, further refinement of the decision-making process is needed to better guide clinicians in treatment selection for the other 50% of patients who did not meet any of the subgroup criteria or met more than 1 subgroup criterion.

To be useful for research or clinical practice, a classification system must demonstrate certain characteristics related to reliability, feasibility, generalizability, and various aspects of validity. Several of these characteristics have been investigated with respect to the TBC system. The classification categories appear to identify meaningful subgroups of patients based on the results of randomized clinical trials comparing the outcomes of patients whose treatment is matched to their classification and those whose treatment is unmatched. Specific criteria have been identified for the various classification categories, with evidence of fair to good reliability among raters for many of these criteria. The overall reliability of classification judgments has also been examined in several studies. Initial studies reported percentage agreements ranging from 55% to 65%, with corresponding kappa values from 0.45 to 0.56. These results have led to the development of a more explicit decision-making algorithm. Fritz et al also examined the interrater reliability of classification decisions made with this algorithm using 7 therapists with varying levels of experience and found an overall agreement between therapists of 76%, with a kappa value of 0.60 (95% confidence interval [CI]: 0.56, 0.64), with no differences in reliability based on experience. Recently, Stanton et al proposed a modified, comprehensive, hierarchical TBC algorithm that would also provide guidance for classifying patients who do not clearly meet the criteria set forth in the original algorithm. In this algorithm, when classifying the approximately 25% of patients who met more than 1 subgroup criterion, raters would need to use additional information outlining “factors favoring” and “factors against” classifications in each treatment category. In the study, the reliability of the 2 novice raters examining 32 patients was moderate, with a kappa value of 0.52 (95% CI: 0.27, 0.77) and a percentage agreement of 67%. Interestingly, following the first examiner’s assessment, 38% of the patients had an unclear classification, whereas 61% had an unclear assessment following the second examiner’s assessment.

Although the interrater reliability of the classification judgments using the TBC system is moderate to good, it is clear that some degree of error persists in the classification decision making associated with the TBC system. Ideally, any individual with LBP should fit primarily 1 classification category. The complex and multidimensional clinical presentation of patients with LBP, however, results in...
The subjects in this study were 12 raters from the University of Vermont (UVM), 11 of whom were licensed physical therapists. The raters completed a questionnaire to assess their education level, credentials, physical therapy experience, and familiarity with the TBC schema (TABLE 1). The physical therapists had an average of 13.3 (range, 4–27) years of practice, and all practiced in an outpatient orthopaedic setting. None had completed a residency program or had an American Physical Therapy Association specialty certification. The 12th rater was a neuroscientist who specialized in posture control. All raters were informed of the experimental protocol and the potential risks of the study, and gave written consent prior to their participation. The protocol and consent form were previously approved by the University of Vermont Institutional Review Board.

Materials and Patient Data Collection

The 12 raters were instructed in the use of a published algorithm24 (FIGURE) to determine patient classification according to the TBC schema, using patient data from a previously published, randomized clinical trial.4 Thus, the raters in the current study did not actually collect the patient data. The previously published, randomized clinical trial was designed to identify some subgroups of patients with LBP; however, the traction subgroup was not included in this trial,4 because patients with signs of nerve root compression were excluded. Thus, the traction category of the TBC algorithm was not included in the current study.

In the previously published trial,4 patients with LBP who qualified for inclusion were between the ages of 18 and 65 years. They were referred to physical therapy for their LBP symptoms, the duration of which had to be less than 90 days. The patients also had to have a modified Oswestry Disability Questionnaire48 score of greater than 25% to be included. Potential patient participants were excluded if they had any history of surgery to the lumbar sacral region, were pregnant, had a visible lateral shift or acute kyphotic deformity, had no reproduction of symptoms with lumbar range of motion or palpation, or had signs of

![FIGURE. Treatment-based classification decision-making algorithm used by the 19 raters. Abbreviations: FABQ, Fear-Avoidance Beliefs Questionnaire; FABQPA, FABQ physical activity subscale; FABQW, FABQ work subscale; LBP, low back pain; ROM, range of motion; SLR, straight leg raise. Adapted with permission from Fritz et al.23](42-09 Henry.indd)
nerve root compression (positive straight leg raise test and/or lower-limb reflex or strength deficits).

Demographic information (age, sex, prior history of LBP, aggravating and relieving factors, and the duration and location [low back only, below the buttock, or below the knee] of current symptoms) was collected, as well as pain rating (11-point numeric pain rating scale) and fear avoidance data using the Fear-Avoidance Beliefs Questionnaire, which includes work and physical activity subscales. The physical examination included range-of-motion measurements with inclinometers for total lumbar spine flexion and extension and straight leg raise. Centralization or peripheralization of low back symptoms with lumbar movement was recorded. Patients were also asked to perform 10 repetitions of trunk extension movements in standing and trunk flexion movements in sitting, as well as to hold a trunk extension position for 30 seconds in prone lying. Centralization or peripheralization of low back symptoms was also recorded for each of these movement tests. The presence of any aberrant movement patterns during trunk flexion/extension movements was noted and the prone instability test was performed as described previously. Lumbar intervertebral mobility (normal, hypomobile, hypermobile) was assessed with the patient in prone lying, by applying a posterior-to-anterior force over each lumbar spinous process. The presence of pain, either local (directly under the therapist’s hand) or distal, with each mobility assessment was also recorded.

The previously published, randomized clinical trial included 123 patients with LBP, from which 24 cases were randomly selected for inclusion in this reliability study. The questionnaire and clinical exam data for each patient with LBP were transcribed onto a standardized, 2-page clinical examination form. These data were then used by the 12 raters in the current study to classify each patient using the TBC decision-making algorithm (FIGURE).

Procedures
During a brief training session (2 hours), the 12 UVM raters were oriented to the goals of the study, the paperwork, the TBC schema, and the decision-making algorithm (FIGURE). Several practice cases (different from those used in this study) were used and discussed to familiarize the raters with the data and the algorithm. Following that training, each of the 12 raters was given 24 clinical examination forms that included the patients’ history and physical exam data and a recording form on which to record the classification choice. Working independently, each rater was instructed to use the information on the clinical examination form to assign 1 of 3 classification categories (manipulation, specific exercise, or stabilization) to each of the 24 cases using the decision-making algorithm (FIGURE). Raters were blind to the judgments of the other raters in this study and to those made in the original clinical trial. Only 1 classification judgment was permitted per subject, and each rater was instructed to submit a classification judgment on each of the 24 patients.

To examine which classification judgments were most difficult to distinguish from one another and to examine the interrater reliability of the TBC schema, the ratings of the 12 UVM raters were combined with those of 7 more experienced, expert (EXP) physical therapy raters from a previous study. Two of the

### TABLE 2

Descriptive Characteristics of the 24 Patients With LBP Using the Treatment-Based Classification Approach

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, y</td>
<td>39.2 ± 11.4</td>
</tr>
<tr>
<td>Duration of symptoms (median [range]), d</td>
<td>20 (1-90)</td>
</tr>
<tr>
<td>Sex (female), %</td>
<td>50%</td>
</tr>
<tr>
<td>Numeric pain rating scale (0-10)</td>
<td>5.6 ± 1.5</td>
</tr>
<tr>
<td>Oswestry Disability Questionnaire score (0%-100%)</td>
<td>40.2% ± 11.5%</td>
</tr>
<tr>
<td>FABQ work subscale (0-42)</td>
<td>11.5 ± 8.7</td>
</tr>
<tr>
<td>FABQ physical activity subscale (0-24)</td>
<td>16.4 ± 5.6</td>
</tr>
<tr>
<td>Symptoms distal to buttock (yes), %</td>
<td>41.7%</td>
</tr>
<tr>
<td>Prior history of LBP (yes), %</td>
<td>70.8%</td>
</tr>
</tbody>
</table>

Abbreviations: FABQ, Fear-Avoidance Beliefs Questionnaire; LBP, low back pain.
*Values are mean ± SD unless otherwise indicated.

### TABLE 3

Percent Agreement Among the Raters and for the Categories

<table>
<thead>
<tr>
<th></th>
<th>UVM Raters (n = 12)</th>
<th>EXP Raters (n = 7)</th>
<th>All Raters (n = 19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent agreement for raters</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kappa (95% CI)</td>
<td>0.62 (0.59, 0.65)</td>
<td>0.47 (0.41, 0.54)</td>
<td>0.57 (0.55, 0.69)</td>
</tr>
<tr>
<td>Overall percentage agreement</td>
<td>80.9%*</td>
<td>68.5%</td>
<td>75.5%</td>
</tr>
<tr>
<td>Kappa (95% CI) by classification category for each rater group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specific exercise</td>
<td>0.64 (0.48, 0.80)</td>
<td>0.60 (0.35, 0.85)</td>
<td>0.61 (0.49, 0.72)</td>
</tr>
<tr>
<td>Manipulation</td>
<td>0.73 (0.49, 0.98)</td>
<td>0.54 (0.19, 0.90)</td>
<td>0.66 (0.47, 0.84)</td>
</tr>
<tr>
<td>Stabilization</td>
<td>0.57 (0.40, 0.74)</td>
<td>0.42 (0.15, 0.69)</td>
<td>0.49 (0.23, 0.62)</td>
</tr>
</tbody>
</table>

Abbreviations: CI, confidence interval; EXP, more experienced; UVM, University of Vermont.
*Significantly higher than the EXP group (P<.01).
7 raters were considered experts in the TBC schema, 3 were experienced physical therapists, and 2 were novice physical therapists. These 7 EXP raters rated the same 24 cases that were rated by the UVM raters. The purpose of combining the 19 raters’ data was to increase the number of raters and to compare their ability to apply the TBC decision-making algorithm to an existing data set.

Data Analysis
Descriptive characteristics were examined in the 19 raters and 24 patients with LBP. To assess interrater reliability of the TBC schema, overall percentage agreement and kappa coefficients with 95% CIs, as well as the distribution of classification judgments, were computed and compared among the 12 UVM raters and the 7 EXP raters. A chi-square test was used to test for a difference in percentage agreement between UVM and EXP raters, as well as to examine differences in classification distributions between UVM and EXP raters. To explore the nature of the disagreements in the classification decisions, we examined the percentage of judgments in agreement and those in disagreement compared to the percentage of the majority of judgments in each classification category. For each patient with LBP included in the analysis, we considered all raters’ judgments and categorized them as in agreement (ie, the same classification judgment compared to the majority of raters) or in disagreement (ie, different classification judgments compared to the majority of raters). For the single case in which there was a tie in the judgments, the case was classified according to the judgment of the 2 expert raters. Finally, we examined classification judgments that were in disagreement pairwise, counting each pair only once, and categorized them into 3 possible categories of disagreement: manipulation-stabilization, manipulation-specific exercise, or stabilization-specific exercise. The frequency of each disagreement category was calculated and a chi-square test was done for each of 3 possible categories of disagreement: manipulation-stabilization, manipulation-specific exercise, or stabilization-specific exercise. The frequency of each disagreement category was calculated and a chi-square test was done for each of 3 possible categories of disagreement: manipulation-stabilization, manipulation-specific exercise, or stabilization-specific exercise. The frequency of each disagreement category was calculated and a chi-square test was done for each of 3 possible categories of disagreement: manipulation-stabilization, manipulation-specific exercise, or stabilization-specific exercise. The frequency of each disagreement category was calculated and a chi-square test was done for each of 3 possible categories of disagreement: manipulation-stabilization, manipulation-specific exercise, or stabilization-specific exercise.

RESULTS

The descriptive characteristics of the 19 raters and the 24 subjects with LBP are provided in TABLES 1 and 2, respectively. Of the 24 patients, the 2 expert raters determined that 5 cases should be classified as specific exercise, 13 cases as manipulation, and 6 as stabilization.

UVM Raters Only
Of the possible 288 classifications (24 patients × 12 raters), there were a total of 277 classification judgments available for analysis, thus 1464 pairwise comparisons, using each pair only once, among the 12 raters’ chosen classification. There was an overall 80.9% agreement in the pairs of classification, with a kappa coefficient of 0.62 (95% CI: 0.59, 0.65) (TABLE 3). When examining rater agreement for a particular classification category, the kappa coefficient was 0.64 (95% CI: 0.48, 0.80) for the specific exercise category, 0.73 (95% CI: 0.49, 0.98) for the manipulation category, and 0.57 (95% CI: 0.40, 0.74) for the stabilization category.

Combined Data Across 19 Raters
Data from the 12 UVM raters were combined with those of 7 EXP raters, resulting in a total of 445 classification judgments made for 24 patients, for a total of 3907 pairwise combinations of rater classification judgments. Overall, 24.0% of the classification judgments were specific exercise, 48.3% were manipulation, and 27.6% were stabilization (TABLE 4). There was no difference in the distribution of classifications among raters from UVM or EXP (chi-square P = .92).

Of the 3907 pairwise combinations of rater classification judgments, 2951 pairs were in agreement, for an overall agreement of 75.5% with a kappa coefficient of 0.57 (95% CI: 0.55, 0.69) (TABLE 3). The overall agreement for the EXP raters was 68.5% with a kappa coefficient of 0.47 (95% CI: 0.41, 0.54) (TABLE 3). The percentage agreement (80.9%) among the UVM raters noted above was greater than that of the EXP raters (P<.01). When examining agreement by a particular classification category for all 19 raters, the kappa coefficient was 0.61 (95% CI: 0.49, 0.72) for the specific exercise category, 0.66 (95% CI: 0.47, 0.84) for the manipulation category, and 0.49 (95% CI: 0.36, 0.62) for the stabilization category.

Additionally, we examined the frequency of agreement versus disagreement judgments compared to the majority judgments for each classification category (TABLE 5). Raters were most likely to agree on the classification of specific exercise. When the majority judgment was specific exercise, 85.9% of the individual raters made this judgment. Raters were only slightly less likely to agree on a classification of manipulation (84.3%),
and were least likely to agree on a classification of stabilization (77.5%). Thus, the stabilization category had the highest percentage of judgments that disagreed (22.5%), followed by the manipulation category (15.7%) and, last, the specific exercise category (14.2%) (TABLE 5).

To explore the nature of the disagreements in the classification decisions, we examined the percentage of disagreements in pairwise raters’ judgments for each of 3 possible classification disagreements (manipulation-stabilization, manipulation-specific exercise, or stabilization-specific exercise) for the 19 raters. The overall percentage of pairwise classification judgments that disagreed was 24.5% (TABLE 6). Of the 24.5% of the discordant pairwise classification judgments, the most common disagreement occurred with 1 rater making a judgment of manipulation and the other rater making a judgment of stabilization (11.0%), followed by the disagreement between the stabilization and specific exercise categories (8.2%). The least common pairwise disagreement occurred between judgments of manipulation and specific exercise (5.3%). It is important to note that in 4 patients the expert raters were not in agreement with the majority of classifications. Three of these 4 disagreements involved a manipulation-stabilization mismatch. The distribution of the pairs that disagreed significantly differed between UVM and EXP (chi-square P < .01). Post hoc comparisons for the individual categories revealed that the UVM raters had similar proportions of manipulation-stabilization and specific exercise-stabilization disagreements, whereas EXP raters had more manipulation-stabilization disagreements and comparatively fewer specific exercise-stabilization disagreements. For both groups, the fewest disagreements occurred between the specific exercise and manipulation categories.

**DISCUSSION**

This study examined the interrater reliability of applying the classification criteria of the TBC system to clinical data that were previously collected, and also explored the pattern of disagreements in classifications to determine which categories were most difficult to distinguish. In some instances, the agreement among the more novice UVM raters was comparable to or better than that among the EXP raters.

The percentage of disagreements in classification judgments, compared to majority ratings, was highest in the stabilization category, involving mismatches with either the specific exercise or manipulation category.

**Interrater Reliability for TBC Classification**

The 12 UVM raters had little to no prior experience with the TBC schema and, with minimal training, were able to demonstrate good interrater reliability in applying the classification criteria of the TBC system. These results, similar to those reported by others,13,23,36 demonstrate the ease of learning this classification system and its potential clinical utility. Interestingly, the UVM rater who was not a licensed physical therapist had the highest percentage agreement with the majority rating, lending further

### TABLE 5

<table>
<thead>
<tr>
<th>Majority Classification Category</th>
<th>Patients per Expert Rater, n</th>
<th>Classification Judgments, n</th>
<th>Specific Exercise</th>
<th>Manipulation</th>
<th>Stabilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specific exercise</td>
<td>5</td>
<td>92</td>
<td>79 (85.9%)</td>
<td>2 (2.2%)</td>
<td>11 (11.9%)</td>
</tr>
<tr>
<td>Manipulation</td>
<td>13</td>
<td>242</td>
<td>12 (5.0%)</td>
<td>204 (84.3%)</td>
<td>26 (10.7%)</td>
</tr>
<tr>
<td>Stabilization</td>
<td>6</td>
<td>111</td>
<td>16 (14.4%)</td>
<td>9 (8.1%)</td>
<td>86 (77.5%)</td>
</tr>
</tbody>
</table>

*Values are n (%).

### TABLE 6

<table>
<thead>
<tr>
<th>UVM Raters (n = 12)</th>
<th>EXP Raters (n = 7)</th>
<th>All Raters (n = 19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall percentage of pairwise judgments that disagreed</td>
<td>19.1%</td>
<td>31.5%</td>
</tr>
<tr>
<td>Percentage of manipulation-specific exercise pairwise judgments that disagreed</td>
<td>3.9%*</td>
<td>7.3%</td>
</tr>
<tr>
<td>Percentage of manipulation-stabilization pairwise judgments that disagreed</td>
<td>7.7%*</td>
<td>16.5%</td>
</tr>
<tr>
<td>Percentage of specific exercise-stabilization pairwise judgments that disagreed</td>
<td>7.5%*</td>
<td>7.7%</td>
</tr>
</tbody>
</table>

*Significantly different proportions from EXP group (chi-square P = .01); post hoc comparisons for the individual categories revealed that the UVM raters had similar proportions of manipulation-stabilization and specific exercise-stabilization disagreements, whereas EXP raters had more manipulation-stabilization disagreements and comparatively fewer specific exercise-stabilization disagreements.
support to the ease of learning the TBC system.

Other studies have used a similar methodology to that of the present study, using written patient cases rather than actual patients. Dankaerts et al\textsuperscript{10} distributed 25 cases (patients’ subjective information and videotaped functional tests) to 13 clinicians, who independently classified each case using particular pieces of the history and physical examination on a mechanism-based classification system for nonspecific LBP.\textsuperscript{32} Kappa coefficients ranged from 0.47 to 0.80, and the percentage agreement among raters was 70% (range, 60%-84%), indicating moderate reliability for this classification approach. The use of paper cases removes additional confounding factors, such as exam performance, that may influence reliability outcomes and thus allows the examination of the classification algorithm itself.

Other studies, in contrast, have had raters examine patients repeatedly.\textsuperscript{11,21,23,32,36,38,39,42} The same patients were examined by different clinicians to determine if they would obtain similar clinical results and assign each patient to the same classification category. Physical therapists with varying levels of experience used the TBC schema and agreed on the categorization of 67% to 79% of mechanical cases.\textsuperscript{33,36} In another classification schema that used pain pattern recognition as the basis for subgrouping of patients with LBP (n = 204),\textsuperscript{42} paired physical therapists performed independent examinations on each patient and assigned the patient to 1 of 5 pain patterns. Agreement on patient classification by independent examiners was 78.9% ($\kappa = 0.61$). Although this methodology may simulate actual clinical decision making more closely than paper cases do, it does not allow direct testing of mutual exclusivity of the subgroups within the algorithm. If the reliability of a classification schema is shown to be poor and there is a high percentage of disagreements in classification judgments, one does not know if this is due to a lack of robustness of the algorithm or to poor reliability of any number of individual test items that are used in the algorithm to make the classification determination.

**Patterns of Classification Disagreements**

In our study, there was an overall percentage of discordant pairwise judgments of 24.5%, indicating that some patients with LBP are difficult to categorize with the classification decision making associated with the TBC system. This prevalence is similar to that reported by Stanton et al\textsuperscript{36} which indicated that 25% of the patients examined met the subgroup criteria for more than 1 subgroup. Our results point to 2 more prevalent patterns of disagreement involving the stabilization category: stabilization-manipulation and stabilization-specific exercise. In contrast, Stanton et al\textsuperscript{36} reported that the most common subgroup combination for patients who met the criteria for 2 subgroups was manipulation and specific exercise. One possible reason for this difference is that in the current study, the sampling of patients in the specific exercise subgroup (according to the expert raters) was the lowest (n = 5) of the 3 subgroups sampled. It is also important to note that there were 4 patients about whom the expert raters were not in agreement with the majority of classifications. This may reflect the greater difficulty of classifying the cases chosen for the current study and a lack of sufficient guidance in using the algorithm\textsuperscript{36} in these more challenging cases.

Raters reported that in about 50% of the patients, assignment of the appropriate category was clear and they were confident in their selection. For the other half of the patients, they reported needing to use the lower portion of the decision-making algorithm, in which they had to weigh factors favoring versus factors against a particular classification category (**FIGURE**). This made the pattern recognition for distinguishing categories more difficult, and raters reported being less confident in their category selection. In these instances, the decision-making process clearly differed among the raters, as each weighted these factors differently, likely based on their own biases or previous clinical experience. Similarly, Stanton et al\textsuperscript{36} reported that the 2 novice raters who participated in the reliability portion of their study had different percentages of unclear assessments (38% of patients for the first rater and 61% of patients for the second rater), which necessitated that the raters use the additional classification criteria in the bottom table of the algorithm used in that study.

The implication of the stabilization-manipulation and stabilization-specific exercise mismatches is that, clinically, patients may not be achieving optimal treatment outcomes because of misclassification. Although achieving 100% sensitivity and specificity in a classification algorithm is unrealistic, caution must be taken when using a classification algorithm that lacks mutual exclusivity too prematurely, as there is a risk of reaching incorrect conclusions regarding the effectiveness of the treatment tested\textsuperscript{36} and the clinical features that determine patient subgrouping. Our results suggest that additional refinement may be necessary for the stabilization category to distinguish the clinical features associated with stabilization treatment success from those of the specific exercise or manipulation categories.

**Study Limitations**

One limitation of the study involves the sampling of cases. The 3 categories examined in this study were not represented equally, and the traction category, the fourth possible TBC category, was not included. Based on the expert raters’ judgment, as well as the majority rating, the sampling was biased toward the manipulation category. Given that percentage of agreement is influenced by prevalence of categories, raters had more opportunities to correctly (or incorrectly) classify cases for the manipulation category, and this is reflected in the highest kappa (0.66) for this category in the combined data. That said, the specific exercise category
had a kappa of 0.61 and had the lowest sample, based on the EXP raters’ judgments as well as the majority rating. Although the sample of the present study was adequate, a larger sample could have provided a greater diversity of cases, thereby increasing the generalizability of the results.

Even though we used raters from 4 different clinical sites, the raters did not examine the patients themselves, and we must acknowledge the possible error in original collection of data from the previous study. The raters based their classification judgments solely on data previously recorded by a physical therapist. Although this removed 1 source of potential error (ie, inconsistency of clinical exam performance), it did not replicate the decision making that occurs in the clinic. Thus, other factors that impact clinical decision making, which might have improved or worsened the raters’ classification success, cannot be taken into account. However, this study design allowed us to examine the reliability and the pattern of disagreements that existed when using the TBC algorithm.

CONCLUSION

The TBC algorithm appears to be easy to learn for novice users. With minimal training, novice raters were able to apply the algorithm in classification judgments with good interrater reliability. However, error persisted in the classification decision making associated with the TBC system, in particular for the stabilization category. Additional refinement through the identification of additional and/or novel patient characteristics is needed to improve the clinical utility of the TBC schema.

KEY POINTS

FINDINGS: Raters were most likely to agree on the classification of specific exercise and least likely to agree on a classification of stabilization. The most common disagreement occurred with 1 rater making a judgment of manipulation and the other rater making a judgment of stabilization. The least common disagreement occurred between judgments of manipulation and specific exercise.

IMPLICATIONS: The issue of mutual exclusivity for the 3 TBC categories studied needs to be further addressed through refinement of the decision-making algorithm. Clinical application of the TBC algorithm for patients considered in the stabilization category should be judicious, as this category had the highest percentage of classification mismatches among raters.

CAUTION: Raters were only provided information about the patient on paper and did not conduct the clinic exam themselves. Thus, other factors impacting clinical decision making that might have improved or worsened the raters’ classification success were not taken into account.

ACKNOWLEDGEMENTS: The authors would like to acknowledge the physical therapy practices that participated in this study: Dee Physical Therapy, South Burlington, VT; Evolution Physical Therapy and Yoga, Burlington, VT; Fletcher Allen Health Care, Burlington, VT; and Timberlane Physical Therapy, South Burlington, VT.

REFERENCES


16. Fritz JM, Delitto A, Vignovic M, Busse RG. Inter-


24. 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.


NOTIFY JOSPT of Changes in Address

Please remember to let JOSPT know about changes in your mailing address. The US Postal Service typically will not forward second-class periodical mail. Journals are destroyed, and the USPS charges JOSPT for sending them to the wrong address. You may change your address online at www.jospt.org. Visit “INFORMATION FOR READERS”, click “Change of Address”, and select and complete the online form. We appreciate your assistance in keeping JOSPT’s mailing list up to date.